# The New Encyclopædia Britannica

Volume 16

MACROPÆDIA

Knowledge in Depth

FOUNDED 1768
15 TH EDITION

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

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



THE UNIVERSITY OF CHICAGO

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

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

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


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



Eleventh Edition
© 1911
By Encyclopædia Britannica, Inc.

Twelfth Edition
© 1922
By Encyclopædia Britannica, Inc.

Thirteenth Edition
© 1926
By Encyclopædia Britannica, Inc.

Fourteenth Edition
© 1929, 1930, 1932, 1933, 1936, 1937, 1938, 1939, 1940, 1941, 1942, 1943, 1944, 1945, 1946, 1947, 1948, 1949, 1950, 1951, 1952, 1953, 1954, 1955, 1956, 1957, 1958, 1959, 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 1970, 1971, 1972, 1973
By Encyclopædia Britannica, Inc.

Fifteenth Edition
© 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991
By Encyclopædia Britannica, Inc.

© 1991
By Encyclopædia Britannica, Inc.

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

No part of this work may be reproduced or utilized
in any form or by any means, electronic or mechanical,
including photocopying, recording, or by any
information storage and retrieval system, without
permission in writing from the publisher.

Printed in U.S.A.



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

# CONTENTS


1 CHICAGO
9 CHILDHOOD DISEASES
21 CHILE
36 CHINA
231 CHINESE LITERATURE
241 CHORDATES
251 CHRISTIANITY
367 CHUNGKING
371 CHURCHILL
377 CIRCULATION AND CIRCULATORY SYSTEMS
418 CIRCUS
425 CITIES
436 CLIMATE AND WEATHER
523 CNIDARIANS
529 COINS AND COINAGE
556 COLLECTIVE BEHAVIOUR
568 COLOGNE
571 COLOMBIA
585 Biological COLORATION
595 COLOUR
605 COLUMBUS
611 COMBINATORICS and Combinatorial Geometry
623 COMMUNICATION
629 COMPUTER SCIENCE
638 COMPUTERS
653 Confucius and CONFUCIANISM
663 CONSERVATION OF NATURAL RESOURCES
687 CONSTANTINE the Great
690 CONSTITUTION AND CONSTITUTIONAL GOVERNMENT
695 CONSTITUTIONAL LAW
704 CONTINENTAL LANDFORMS
760 COPERNICUS
762 The COSMOS
796 CRIME AND PUNISHMENT
817 CRIMINAL LAW
822 Oliver CROMWELL
827 The CRUSADES
840 CRUSTACEANS
860 CRYPTOLOGY
874 The Concept and Components of CULTURE
894 CYPRUS
902 CZECHOSLOVAK LITERATURE
905 CZECHOSLOVAKIA
931 DAMASCUS
935 The Art of DANCE
958 The History of Western DANCE
971 DANTE
977 DARWIN
982 DEATH

# Chicago

Until the 1830s a minor trading post at a swampy river mouth near the southwestern tip of Lake Michigan, Chicago used its strategic location as the interior land and water hub of the expanding United States to become the centre of one of the world's richest industrial and commercial complexes. It is the third most populous city and metropolitan area in the United States. Chicago's achievements are distinctly characteristic of the country as a whole, and its problems are the problems of the modern United States; in a sense it may be—as a series of observers has called it—the typical U.S. city.

The relations between this youthful city and its rural environment are also noteworthy. Throughout its history, Chicago and the surrounding counties of what became its metropolitan area, now containing about two-thirds of the population of Illinois, have existed as almost a separate entity—politically, socially, and spiritually—from largely rural "Downstate" Illinois. The attitudes and lives of the early settlers in and around the burgeoning city, mainly from the Northeastern states or from Europe, were in contrast to those of Downstaters, many of whom came from Appalachian or Southern states. While Chicago was, for example, a major supplier of goods and manpower to the Union during the Civil War, in southern Illinois there was an unsuccessful but strong movement toward secession and alliance with the Confederacy. This alienation continues to plague the political and social life of both the city and the state.

This article is divided into the following sections:


Physical and human geography 1
- The character of Chicago 1
- The landscape 3
  - The city site
  - Climate
  - The city layout
- The people 4
- The economy 5
  - Industry
  - Commerce
  - Finance
  - Transportation
- Administrative and social conditions 5
  - Government
  - The social milieu
  - Housing
  - Education
  - Health
- Cultural life 6
  - The arts
  - Museums
  - Recreation
  - Journalism and broadcasting

History 7
- Settlement and early activity 7
- Growth and development 7
  - Explosive economic growth
  - The rebuilt city and its people
  - Symbols of civic consolidation
  - The 20th century

Bibliography 9


## Physical and human geography

### THE CHARACTER OF CHICAGO

A by-product of Chicago's growth on the raw frontier of U.S. industry was its reputation as a city in which "anything goes," a city whose name became an international byword for underworld violence during and after the Prohibition era of the 1920s and early 1930s. This sort of mayhem has long been overshadowed in Chicago as elsewhere in the United States by the random violence of daily urban life. Municipal corruption, another commodity on which Chicago was long thought to have cornered the market, is likewise not in fact a local monopoly, though Chicagoans perhaps have a higher tolerance for human frailty among politicians—politics in Chicago being to an extent an expensive form of public entertainment—than do the citizens of other municipalities. As theologian Martin Marty of the University of Chicago observed, after revelations that politicians of both parties had profited enormously from ownership of racetrack stocks, "Someplace else it might be shocking. . . . Children grow up here knowing things are rigged and fixed."

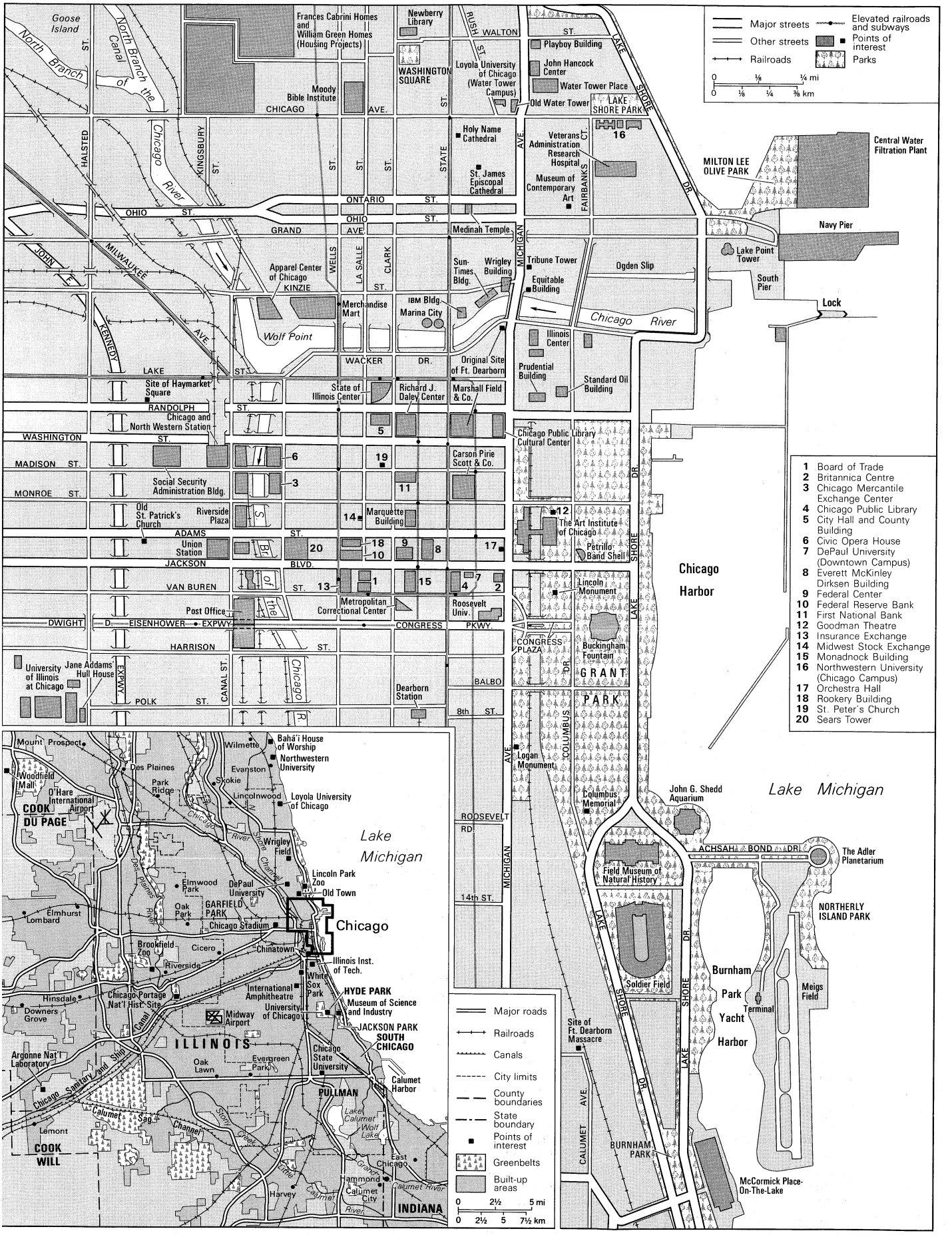
However much Chicago's political and social life may have deserved the brickbats of its numerous critics, there is little disagreement that the city's physical presence is stunning. Chicago arose from the ashes of its Great Fire in 1871 to develop the skyscraper as well as many of the other major innovations of modern architecture. In the decades immediately following World War II, however, exigencies of the marketplace often conquered civic pride in maintaining the great landmarks of Chicago's past. There were exceptions—notably, the Auditorium Building and the Newberry Library—but these were preserved through limited private initiative. More recently, public awareness and effective legislation have fostered increased conservation efforts. This factor and the desire for more land on which to build new structures have aided in the southward and westward expansion of Chicago's downtown into formerly blighted areas, so that the city's striking skyline, containing some of the world's tallest buildings, rises along a continually widening strip.

Urban vistas

Behind this impressive facade lies a sprawling industrial city, its monotony accentuated by the flat Midwestern landscape and by a repetitive gridiron pattern of streets broken only by the radial avenues that cover old Indian trails to the northwest and southwest and the great freeways and railroad lines that for many years have made the city a major hub of commerce. The whole mass reaches out over the former prairie, spilling over city limits into an irregular and continuously expanding belt of suburbs and industrial satellites.

The magnificent downtown lakeside strip nevertheless remains the focus of attention in the mind of resident, commuter, and visitor alike. A person strolling north on Michigan Avenue (in the downtown area) passes the green acres of Grant Park, with the neoclassical building of the Art Institute of Chicago and the well-hidden tracks of the Illinois Central Gulf Railroad; the Chicago Public Library Cultural Center, with arched rooms and hallways decorated with fine mosaic work of a past era; and one of the world's greatest skyscraper complexes. From the bridge that spans the Chicago River the stroller is confronted with what many people regard as one of the most beautiful and open urban spaces in the world, stretching along both sides of what once was the river's estuary. North of the river along Michigan Avenue is "the Magnificent Mile"—Chicago's answer to New York City's Fifth Avenue in commercial elegance—which includes the Old Water Tower, whose medieval stone turrets survived the conflagration of 1871 to become an eccentric monument to civic nostalgia.

Outside these areas of downtown Chicago, the stroller finds a complex city, a kaleidoscope of neighbourhoods mirroring the ethnic and racial diversity of U.S. life.




Central Chicago and (inset) its metropolitan area.

Chicago remains essentially the "blue-collar" city characterized by the poet Carl Sandburg as the "city of big shoulders," heavily populated by the descendants of labourers from the streets and soils of 19th-century Europe and of former slaves from the Deep South. The latest influx—that of Spanish-speaking residents and of immigrants from Southeast Asia and eastern Europe—has added further to the complexity.

Civic disasters of the 19th and 20th centuries

Its widely scattered ethnic neighbourhoods and its suburbs have retained memories of a long series of disasters, running from the time of the Great Chicago Fire itself. In the 19th century these included the police assault on strikers that left several wounded or killed, the apparently retaliatory bomb throwing (attributed to anarchists) that killed seven policemen, and the ensuing reaction against German-American leaders, all associated with the Haymarket Square Riot. Many other bitter, often fatal, labour disputes occurred in the steel, railroad, packinghouse, and other industries. These were followed by the catastrophic Iroquois Theater Fire; the sinking of the cruise ship "Eastland" in the Chicago River, drowning more than 800 persons; the reputation of gangsterism and intermittent mayhem evoked by mention of Al Capone, John Dillinger, and the St. Valentine's Day Massacre in the early 20th century; the televised violence between police and protesters during the 1968 Democratic convention; and a continuing succession of major and minor municipal corruptions. At the same time, "the windy city"—a meteorologically correct nickname nevertheless derived from the inflated claims of the early municipal "boosters"—has laid claim to a distinguished list of citizens who have significantly enriched the intellectual, artistic, and social life of the United States.

It has been said that Chicago's intelligentsia is hindered by a "second city" mentality and an accompanying tendency toward self-disparagement. That may indeed be true, but such a designation is also less likely to inhibit risk-taking experimentation, as evidenced by Chicago's artistic achievements. The city has once again regained its reputation as a major theatre centre, and its contemporary styles of architecture are imaginative and sometimes controversial. This willingness to try something new—epitomized by a huge abstract Picasso sculpture, a gift to the city from the artist himself—continues to entertain Chicagoans and the city's millions of tourists and conventioneers. (R.Do./Ed.)

## THE LANDSCAPE

City and regional populations and areas

**The city site.** The city of Chicago proper occupies 228.1 square miles (590.8 square kilometres). The Chicago Standard Metropolitan Statistical Area (SMSA) consists of Cook County and five surrounding Illinois counties, and the Chicago–Gary–Kenosha Standard Consolidated Statistical Area (SCSA) is made up of nine counties—two of them in northwestern Indiana and one in southeastern Wisconsin.

Chicago's site is generally level, rising from Lake Michigan, 579 feet (176 metres) above sea level, to slightly more than 600 feet in outlying portions of the city. Most of Chicago is built on a plain, the remnant of postglacial Lake Chicago, formed when the retreating continental glacier blocked normal northeastward drainage through the St. Lawrence Valley, about 10,000 to 12,000 years ago. Outlying portions of the metropolitan area, formed from material deposited by the glaciers, rise to more than 700 feet.

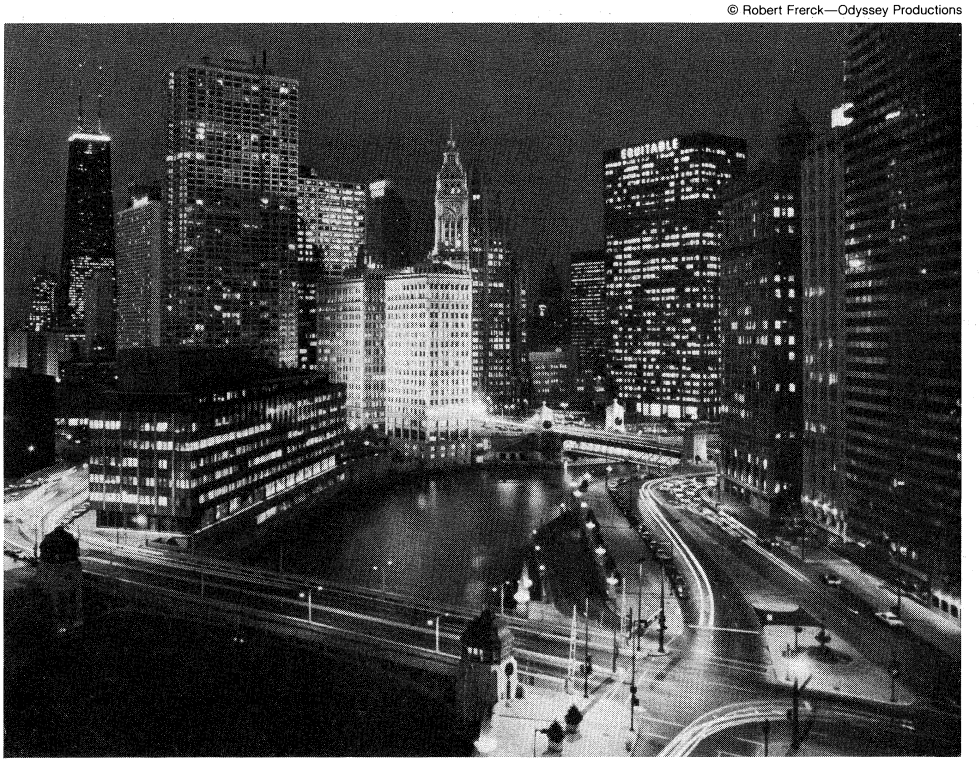
Flow of the Chicago River

The narrow Chicago River extends one mile (1.6 kilometres) inland from Lake Michigan, where it splits, dividing the city into North, West, and South sides. Its original flow was into Lake Michigan, but completion of the Chicago Sanitary and Ship Canal in 1900 reversed it, since the bottom of the canal is below the surface of Lake Michigan. Near the southeastern corner of the city, the flow of the Calumet River was reversed by the Calumet Sag Channel, completed in 1922 and enlarged from 1955 to 1972 as a modern barge route. The two waterways join southwest of the city and receive treated sewage effluent from three plants of the Metropolitan Sanitary District of Greater Chicago, as well as wastes from industrial plants and outlying areas. During high runoff the rivers occasionally revert to their original lakeward drainage, and some flooding occurs in low-lying areas.

**Climate.** The climate is subject to rapid changes of weather as successions of air masses pass generally from west to east. Lake Michigan tends to mitigate extremes, with lower temperatures in summer and higher in winter generally occurring close to the lake.

January temperatures average about 25° F (−4° C) and July 75° F (24° C). Annual precipitation averages about

© Robert Frerck—Odyssey Productions



The Chicago River, looking east. The John Hancock Center is at the far left, and the Wrigley Building, with its clock tower, is in the centre.

33 inches (838 millimetres), and heavy snows occasionally disrupt local transportation.

**The city layout.** Chicago meets its suburbs in a ragged pattern of boundaries on three sides, while on the east the lakefront curves from southeast to northwest. The area centring on the forks of the river was platted on a gridiron pattern in 1830 following specifications of the Northwest Ordinance of 1787. This plan was followed to some degree in the rest of the city, though it was broken often by radial avenues (some following old Indian trails leading to the river mouth) and other features, such as the Burnham Plan of Chicago (1909), rail lines and yards, industrial sites, and parks.

Downtown Chicago has been known as the "Loop" since 1897, when several elevated lines were joined into an overhead loop of tracks encircling an area that covers some 35 blocks and receiving feeder lines from north, west, and south. The building boom that began in the mid-1950s extended the highly concentrated business district westward from the Loop and, in the 1970s, into the Near West Side beyond the river's south branch. Many new skyscrapers have radically altered the city's skyline. North Michigan Avenue, initially developed following completion of the Michigan Avenue Bridge in 1920, and adjacent Near North sites have experienced much high-rise commercial and residential building since the 1960s, the most notable being the 100-story John Hancock Center and the 74-story Water Tower Place. Both are multipurpose skyscrapers containing shopping facilities, restaurants, offices, and apartments; Water Tower Place also includes a hotel. Other major downtown office buildings completed during the 1970s and early 1980s included the 110-story, 1,454-foot Sears Tower—the world's tallest building—just west of the Loop; the 80-story Standard Oil Building east of the Loop; and the Illinois Center complex of office and apartment buildings and a hotel. In the early 1980s an average of 2,800,000 square feet (260,000 square metres) of new office space was being constructed in downtown Chicago each year, in addition to a considerable volume of office-building construction in suburban areas.

Grant Park, in downtown Chicago, Lincoln Park on the North Side, and Jackson and Burnham parks on the South Side, stretch for miles along the lakefront. The city has an impressive park system inland as well.

Principal industrial areas lie along the two branches of the Chicago River and in the Calumet region to the southeast, as well as along railroad lines and in satellite cities, such as Waukegan, Aurora, Joliet, and Chicago Heights in Illinois and the Gary–Hammond–East Chicago complex in Indiana, up to 40 miles from downtown Chicago. In South Chicago, along and near the Calumet River and along the lakefront in adjacent Indiana, are many oil refineries and iron and steel, chemical, and fabricating plants. The first steel plant in the area was established at the entrance of the Calumet River in 1880, followed by additional large plants nearby and at Indiana Harbor in the early 1900s. Gary was established as a major steel-producing centre in 1906. In the 1960s two large steel plants were established at Burns Harbor, east of Gary in the Indiana Dunes area. Meanwhile, the principal terminal of the Great Lakes–St. Lawrence Seaway overseas shipping route was developed after 1956 by the Chicago Regional Port District in Lake Calumet, six miles down the Calumet River from Lake Michigan, supplemented by the opening in the late 1970s of the Iroquois Landing terminal at the entrance of the Calumet River.

Patterns of industrial and suburban growth

## THE PEOPLE

The proportion of foreign-born population of Chicago and the metropolitan area has fluctuated with trends in immigration. For example, among the principal countries of origin in the census of 1970 were Poland, Germany, Italy, the Soviet Union, Sweden, and Ireland. The census of 1980 showed a large increase in Spanish-speaking people of Latin-American origin and in people of Asian origin.

Although there were a few blacks in Chicago from the earliest period of the city's growth, immigration was accelerated during and after World War I. The black population increased from 233,000 in 1930 to almost 1,200,000 in 1980, when it represented two-fifths of the city's population.

North of the Loop, the Lake Shore Drive mansions have been replaced rapidly by equally luxurious high-rise apartment buildings that extend the lakefront Gold Coast almost to the city's border. Inland from this strip lies a narrow band of two- and three-story apartment buildings and older homes occupied largely by single persons and families of professional people.

The area adjacent to the University of Chicago on the South Side forms one of the city's major intellectual communities, though it is virtually surrounded by one of the most blighted sections of the city. Other universities contributing to local social patterns are DePaul and Loyola on the North Side and the University of Illinois at Chicago, the construction of which in the 1960s uprooted much of the old Italian community southwest of downtown Chicago.

Many ethnic and racial groups continue to form more or less homogeneous communities in various parts of the city. The Irish, long in control of the city's politics, are widespread, and a predominantly Irish region on the South Side spawned four mayors of Chicago—all Democrats—between 1933 and 1976. Chicago's Polish community, the largest in the country, remains heavily concentrated on the Near Northwest Side. Swedish and German neighbourhoods reach through the North and Northwest sides, and Bohemian and Slovak Czechs have spread into the southwestern suburb of Cicero, while Greeks have settled just west of the Loop.

Ethnic communities, city and suburban

Heavily Jewish populations are characteristic of the Far North and the adjoining suburbs of Lincolnwood and Skokie. The Chinese community is concentrated on the South Side, and Japanese and Korean communities mainly on the North Side, while the nation's second largest urban enclave of American Indians shares its poverty with Appalachian whites in the Mid-North area.

Many of these ethnic communities have lost most of their distinctive character as the newer generations have become homogenized into U.S. life. Foreign tongues remain in evidence, however, together with storefront restaurants and traditional shops that add old-world flavour to the Chicago neighbourhoods.

The original main axis of black settlement was along mass-transit lines, especially through the South Side, where access to industrial employment was favourable. During and after World War II the South Side "black belt" expanded within the city and into adjacent suburbs, while additional areas of the West Side, once heavily Jewish, and parts of the Near North were occupied by blacks. In both the city and suburbs, most black areas previously had been occupied by first or second generations of foreign origin who either left already deteriorating neighbourhoods or fled a growing influx of blacks. Middle-class and affluent black neighbourhoods developed in areas of social and economic stability.

Black settlement

Residential suburbs first grew up along the principal rail lines, but since World War II much suburban development has taken place, commonly at lower densities, in the areas between the earlier radial axes. A suburban real-estate boom in 1869 created new communities that were able to absorb many of the people burned out by the Great Fire. Among these was Riverside, west of the city, laid out in irregular streets and with exemplary planning, which has enabled the community to retain its character. To the north of the city, such lakeside communities as Evanston, Winnetka, and Lake Forest began a steady growth that in the 20th century made the North Shore the most prestigious of suburban areas. To the west, Oak Park had similar growth.

Among developments that encouraged residential deconcentration after World War II were new express highways; expansion of industries outside the city; huge regional shopping centres, the sales of which rival those downtown; O'Hare International Airport (later annexed to the city) and its surrounding complex of hotels, motels, shopping centres, office buildings, and industrial districts; nuclear-research facilities, including the Argonne National Laboratory and the Fermi National Accelerator Laboratory (Fer-

milab) near Batavia; and many communities and "new towns." The first of these, Park Forest, was begun in 1947 about 30 miles south of the city's centre.

## THE ECONOMY

**Industry.** Chicago and its metropolitan area have remained the most important focus of economic activity in interior North America. Its economic base, with a balance between industry and commerce, is highly diversified. Nevertheless, the city has suffered, along with many other metropolitan areas of the Northeast and Middle West, from the shift in population and economic activity to the "Sun Belt" of the South and West and from foreign industrial competition.

Manufacturing provides about one-fourth of the region's employment; leading categories are steel, metal products, food products and confections, metal furniture, chemicals, soap, paint, machine tools, communications equipment and electronic goods, railroad equipment, surgical appliances, and scientific instruments.

The blue-collar and white-collar economy

Chicago's steel supply and its strategic situation as the major transportation node of the continent has enabled it to assume leadership in the manufacture of a wide variety of machinery and fabricated metal products, ranging from diesel-electric locomotives to printing presses, material-handling equipment, and earth-moving and agricultural machinery.

A wide variety of chemicals and allied products serve both industrial and consumer markets. Closely associated are several large petroleum refineries, principally in the Calumet area and northwestern Indiana.

Chicago's printing establishments include several of the world's largest. Many nationally distributed magazines and mail-order catalogs, as well as a substantial proportion of the country's telephone directories, are produced in these plants. Enormous quantities of paper, much of it from Canada, reach Chicago by water. The city ranks second to New York City in the white-collar aspects of publishing, though it tends to specialize in such areas as educational materials, encyclopaedias, and professional and trade publications. It is also the home office of several major advertising and public-relations firms.

Situated between the agricultural Midwest and the urban-industrial Northeast, Chicago remains a leader in food processing, although by the early 1970s the Union Stock Yards had terminated all meat-processing activities.

**Commerce.** Commercial activities of nationwide importance include trading in commodities futures on the Chicago Board of Trade and the Chicago Mercantile Exchange and associated brokerage offices and related establishments.

Convention facilities

Chicago has more trade shows, conventions, and corporate meetings than any other U.S. city. The Merchandise Mart, with most of its 4,000,000 square feet of floor area devoted to wholesaling activities, was for many years the world's largest commercial building. Other facilities, such as the Apparel Center of Chicago, serve specialized wholesaling industries. Convention and trade-show facilities include concentrations of hotels and motels, notably in the downtown area and in the vicinity of O'Hare Airport, together with the convention hall at McCormick Place-on-the-Lake, opened in 1971 to replace a smaller facility destroyed by fire.

**Finance.** Chicago is the site of a Federal Reserve Bank, established in 1914. Most large banks are in the Loop area, and, because Illinois prohibits branch banking, the outlying neighbourhoods and suburbs are served by limited-service facilities, smaller banks, currency exchanges, and savings-and-loan associations.

The city is also the site of the Midwest Stock Exchange and offices of most major brokerage houses. Many insurance companies are in the city or suburbs, including the nation's two largest automobile insurers.

**Transportation.** In addition to Chicago's standing as a major inland port and railroad hub of the nation, O'Hare International Airport is the world's busiest. By the late 1960s the older, smaller Midway Airport had been pressed into service again to relieve congestion. Metropolitan Chicago's dominance as the most important railroad freight centre in North America has continued, although railroad mergers, the rise of intermodal transportation, and the deregulation of many aspects of the transportation industry have presented new challenges to the region's supremacy. Chicago lost many intercity passenger trains before and after the advent of the quasi-public Amtrak system in 1971. Its several commuter lines serving suburbs to the north and west are widely regarded, however, as the finest in the nation in terms of comfort, punctuality, and overall service. The Regional Transportation Authority (RTA), established in 1973, is responsible for most suburban railroad and bus service, as well as for the rapid-transit and bus services of the Chicago Transit Authority (CTA) within the city and in some nearby suburbs. Both the CTA and RTA have become increasingly debt ridden in spite of having one of the nation's highest fare structures.

The expressway system built after World War II has become congested, as has Lake Shore Drive, which reaches nearly from the northern to southern city limits and provides scenic views of both the lake and the city skyline.

## ADMINISTRATIVE AND SOCIAL CONDITIONS

**Government.** The spiritual chasm between Chicago and the rest of Illinois is perhaps widest in the political and social spheres and deepest in the struggles between city hall and the state government. Chicago Democrats and Downstate Republicans have found few issues over the years that could be debated on a basis other than that of partisan politics. Until the one-person, one-vote reapportionment of the legislature in the 1960s, the whip was usually in the hands of the sparsely populated Downstate counties. In addition, the growth in population and wealth of the suburbs after World War II has created a second front on which Chicago, like most large U.S. cities, is forced to wage a defensive campaign against a growing drain on its human and financial resources.

Overlap of administrative institutions

Routine operation and long-term planning in the Chicago SMSA are complicated by the continuous proliferation of overlapping administrative and taxing units of government. In addition to the almost 300 incorporated municipalities and the unincorporated areas under administration of the counties, the SMSA has more than 500 special districts—elementary and high school and community college, park, forest preserve, drainage, sanitary, and the like—established to circumvent state-imposed limitations on borrowing. Such authorities as public housing, ports, transit, and highways operate without taxing power. State and federal funds and, when appropriate, user charges supplement local outlays.

The Northeastern Illinois Planning Commission, created by the state in 1957, coordinates planning, especially among suburban governments. All projects involving federal aid must conform to the commission's comprehensive plan, adopted in 1968 and periodically reviewed, for the six counties of the SMSA.

Chicago's government long has been handicapped by an unwieldy structure not adapted to efficient administration of a modern urban region. Power is concentrated in a mayor who presides over the City Council of aldermen representing the city's 50 wards. The mayor, with City Council approval, also appoints members of the Board of Education, Park District, Housing Authority, and other special-purpose boards and commissions.

Patronage and the political machine

This formal scattering of power long has nurtured an informal but highly structured and disciplined political machine in Chicago that was brought to its peak of efficiency during the administration of Mayor Richard J. Daley, widely regarded as "the last of the big-city bosses" well before his reelection to a sixth four-year term in 1975. As chairman of the Cook County Democratic Party, he wielded great power beyond the city limits and was recognized as the predominant voice in the statewide party and a major power in the national Democratic Party. A deeply entrenched patronage system in all areas of government was moved into its highest gears for elections, and accusations and denials of voting irregularities were a constant feature of Chicago's political life. Opponents often charged that city building inspectors and other officials employed statutory sanctions to enforce party loyalty or punish dis-

affection and that the City Council, in spite of a scattering of liberal independent and Republican aldermen, served as little more than a rubber stamp for mayoral programs.

The foundations of the "Organization," as it is called by its adherents, were laid during the brief term (1931–33) of Anton Cermak, a Bohemian immigrant who quickly mastered the politics of Chicago's ethnic ghettos, opposed the Prohibition that was unpopular with immigrant workers, and carefully balanced Democratic slates and platforms among the many ethnic, labour, and business interests. Innovative programs for municipal conservation and rebuilding renewal that were begun during the reformist administration of Martin Kennelly (1947–55) were moved ahead rapidly only after Daley's accession in 1955. For more than 20 years an unofficial alliance of labour unions, civic and business leaders (often suburbanites), and party faithful from the precinct level upward, with heavy voting support from the blue-collar and ethnic communities, maintained the Organization's hold on Chicago's political life. Daley was the catalyst of the Organization's unity, and his death in 1976 marked the beginning of the end of its total dominance of city politics. In 1979 the city elected its first woman mayor, Jane M. Byrne, and by the early 1980s there were many "independent" (*i.e.*, non-Organization) Democratic aldermen in the City Council. In 1983 Chicago elected its first black mayor, Harold Washington.

**The social milieu.** The tremendous growth and spread of its black population and the concomitant flight of middle- and upper-class whites and of commerce to the suburbs probably was the most dominant feature of Chicago's social picture after World War II. The city's few black aldermen tended to align themselves with the political machine and bring it the votes of the black community. The growing Spanish-speaking community, without political representation, intensified the situation. The impact of these factors has been felt most heavily in housing, although pressing needs in education and in health and welfare services increasingly force the city to look to state and federal sources for relief. These requests frequently intensify hostilities between city and state governments.

Social and political interactions

Black opposition began to grow and become organized in the 1960s. One of the earliest groups to form was Operation Breadbasket—an economic activity of the Southern Christian Leadership Conference founded by Martin Luther King, Jr.—which mounted campaigns, often in conjunction with businesses and citizen groups controlled by whites, to achieve greater economic and political power for the black community. Its successor, Operation PUSH, and other organizations continued to mobilize blacks who were increasingly frustrated with Mayor Daley and the Organization. By the early 1980s massive voter-registration drives had given the black community a powerful constituency that was able to elect Washington and about one-third of the aldermen in the City Council, many of whom were independent Democrats.

**Housing.** A fundamental clash of values became apparent in the 1960s when Daley, in rebutting a charge that Chicago was the most segregated city in the country, declared that the city had no ghettos. Numerous restrictive real-estate practices, however, long had been in effect in Chicago, intensified by a six-day race riot in 1919 that killed at least 33 persons. Scores of incidents occurred in following years. The inevitable outward explosion of the black community after World War II was abetted by unscrupulous real-estate operators but opposed by "block organizations" and other militant white-citizen groups, especially on the South and West sides.

By 1980 the Chicago Housing Authority had built about 45,000 units of low-rent housing, mainly as massive high-rise apartment projects that, in the view of many persons, tended to intensify the crime, isolation, and other evidences of life in the slums they were intended to replace and to epitomize the worst aspects of racial and socioeconomic segregation. Private institutions, such as the Illinois Institute of Technology on the South Side, had better fortune with privately financed middle-income housing projects. A combination of redevelopment, conservation, and social programs in the area centred on the University of Chicago became a prototype for treatment of changing urban communities.

In 1971 the city's participation in the federally funded Model Cities program was jeopardized by the administration's reluctance to locate low-rent public housing in predominantly white neighbourhoods. At the same time, a number of the suburbs, many of which had fair-housing laws, resisted such housing and were largely unreceptive to pleas for housing aid from the city.

**Education.** The racial patterns of Chicago's public schools reflect neighbourhood residential patterns, with few attempts at integration and a relatively low standing in comparison with nationwide achievement standards. Both the public schools and the Roman Catholic parochial schools, which make up the nation's largest private-school system, have teetered on the brink of financial calamity and provoked intense city–state political feuds.

In higher education the University of Chicago long has been among the nation's most prestigious institutions. Both the Illinois Institute of Technology and Northwestern University, the latter with campuses in Evanston and Chicago, have national reputations, while Loyola and DePaul universities are major Roman Catholic institutions. Roosevelt University, in downtown Chicago, founded in 1945, offers a diverse curriculum especially geared toward an urban student body. The University of Illinois at Chicago complements the main campus in Champaign–Urbana.

Higher education

**Health.** Chicago is among the major medical- and dental-training centres in the nation, and its hospitals and research facilities are of high quality. A high proportion of trained personnel leave the area, however, creating an overall shortage in both the city and the state. As in many cities, service to the poor remains deficient, heavily encumbered by partisan political controversy. Publicly supported Cook County Hospital, one of the nation's largest, has often found itself embroiled in political and financial crises that tend to affect its services, while neighbourhood clinics in black and Spanish-speaking areas, staffed mainly by young doctors and medical students, have been in frequent conflict with the politically run Board of Health.

## CULTURAL LIFE

**The arts.** Its reputation as a boisterous and crassly commercial city notwithstanding, Chicago has fostered a robust artistic life throughout most of its history. It was a major theatre centre during the late 19th and early 20th centuries, and, before the discovery of Hollywood in the early 20th century, it was the cradle of the infant U.S. motion-picture industry. During the 1950s the Second City troupe began a series of theatrical innovations that were to provide many new directions and talents to the entertainment world. The nationally recognized Goodman School of Drama, long affiliated with the Art Institute of Chicago, initiated a resident professional company in 1970 and became one of the eight colleges of DePaul University in 1978. There are numerous avant-garde companies, especially in the youth-oriented sections of the city, and several city and suburban "dinner theatres" have attained artistic acclaim.

Theatrical and musical life

The status of the Chicago Symphony Orchestra as one of the world's major musical ensembles was reinforced with critical enthusiasm after the appointment of Georg (later Sir Georg) Solti as conductor in 1969. Chicago opera revived when the Lyric Opera was founded in 1954 to provide Chicago with brief but regular seasons of opera of a high calibre. Chicago's place in literature was at its highest in the early decades of the 20th century, especially with the publication of *Poetry* magazine. Although many Chicago writers have gained renown, the city's peculiar literary genius has shown itself most prominently in the field of journalism—from bucolic Midwestern homily to stinging social and political commentary.

Although Chicago and environs have incubated the most outstanding examples of modern domestic and commercial architecture, the city's record in preserving its landmarks has been a depressing one. The razing of Louis Sullivan's Stock Exchange Building in 1971 epitomized to conservationists and historians a callousness toward the city's aesthetic heritage that already had replaced nu-

Architecture

merous architectural landmarks with more profitable but uninspired buildings. Considerable interest has developed, however, in preserving Chicago's older buildings of architectural value, and the results are evident in the rehabilitation of many structures. Others have been adapted to new and innovative uses, including Fulton House, a former warehouse converted to a luxury apartment building, and the buildings in Printing House Row.

Architecturally Chicago remains among the finest of the world's large cities, but its plan and skyline have been threatened by an increasing number of conventionalized structures indifferently adapted from the buildings of Ludwig Mies van der Rohe, an innovative genius who had renewed Chicago's architectural history in the years following World War II.

**Museums.** Its many and diversified collections of painting, sculpture, prints, photographs, and handicrafts rank the Art Institute of Chicago among the major museums of the world. In addition, its school makes it an important training centre for the fine arts. The Museum of Contemporary Art provides Chicagoans with a complementary point of view through its exhibitions of the leading edge of artistic endeavour. There are also several commercial galleries that sponsor exhibitions of works by artists with local as well as national and international reputations.

The exhibitions of the Museum of Science and Industry, housed in a huge remnant of the world's fair of 1893, are rivalled in the United States only by those of the Smithsonian Institution in Washington, D.C. The public displays and research activities of the Field Museum of Natural History place it among the leading scientific institutions of the world. Nearby are the John G. Shedd Aquarium and the Adler Planetarium.

**Recreation.** The city's extensive park system is supplemented by forest preserves located along the original city limits and in suburban areas of Cook County. Sandy beaches, in intermittent patches along Lake Michigan, provide summertime recreation.

Professional sports spark civic enthusiasm for the White Sox and Cubs in baseball, the Bears in football, the Black Hawks in hockey, the Sting in soccer, and the Bulls in basketball. For Chicagoans and visitors alike, the many entertainments available in the Near North Side area of nightclubs and cabarets are a continuous attraction.

**Journalism and broadcasting.** Chicago has two metropolitan daily newspapers: the *Chicago Tribune* and the *Chicago Sun-Times.* The *Chicago Defender* is oriented primarily to the city's black population. There are many suburban newspapers and weekly and monthly magazines, as well as several dailies and weeklies in foreign languages. *Chicago Commerce,* a monthly, and *Crain's Chicago Business* carry economic and financial news. The magazine *Chicago* features general articles, stories of local interest, and entertainment notes. Chicago has numerous television and radio broadcasting stations; the public television station, WTTW, was one of the nation's pioneers in educational programming.

## History

### SETTLEMENT AND EARLY ACTIVITY

In 1673 the French explorers Louis Jolliet and Jacques Marquette followed an Indian portage to the mudflats over which a Y-shaped river flowed. It emptied into Lake Michigan, while its arms reached nearly to the drainage basin of the Mississippi River system, thus virtually linking two great North American waterways. The meaning of the Indian name for the region remains disputed—among the possibilities are skunk, wild onion, or powerful.

Trappers, traders, and adventurers used the area for portage and barter throughout the 18th century. The first known non-Indian settler was Jean Baptiste Point Sable (or Pointe du Sable), son of a wealthy French merchant who had moved to Haiti and married a black woman there. Sable settled in the Great Lakes area in the 1770s. In 1795 the United States obtained a six-mile-square area about the river mouth.

Ft. Dearborn, built in 1803, was destroyed in 1812 and all but one of its military and civilian population were killed in an Indian raid. The fort was rebuilt in 1816 and was occupied until the 1830s. Outside its walls a cluster of traders' shacks and log cabins were built, but the settlement attracted little interest even after Illinois, with most of its population in the central and southern regions, became a state in 1818.

The opening of the Erie Canal in 1825, joining the Atlantic states and the Great Lakes, shifted the main axis of westward movement northward from the Ohio River route. Soon afterward, Chicago became the principal western terminus. The county of Cook located its seat at the small community, and the regional federal land office opened there. Numerous retail stores opened to outfit newcomers to the West, and the volume of animal pelts and products for Eastern markets increased. In 1837, the year Chicago became incorporated as a city, its population was about 4,200.

Chicago's geographical potentiality as a water gateway was fulfilled by completion in 1848 of the Illinois and Michigan Canal, linking the Great Lakes and Mississippi systems. A pair of railroad lines from the East tied into Chicago in 1852, and by 1856 it had become the nation's chief rail centre. A belt line connected the radiating trunk lines by 1856, and commuter service to outlying neighbourhoods and suburbs began.

Emergence as a rail centre

### GROWTH AND DEVELOPMENT

**Explosive economic growth.** Industry followed the rails. By the late 1850s lake vessels carried iron ore from the Upper Michigan ranges to the blast furnaces of Chicago. Chicago became the nation's major lumber-distributing centre by the 1880s. The railroads brought farm produce from west and south, and Chicago's Board of Trade became the nerve centre of the commodities market. The railroads also hauled cattle, hogs, and sheep to Chicago for slaughtering and packing. The consolidated Union Stock Yards, largely bankrolled by nine railroads and the owners of several other Chicago stockyards, opened on Christmas Day, 1865.

Chicago emerged as the major city of the Midwest. Its 1880 census reported more than 500,000 inhabitants, a 17-fold increase over 1850; by 1870 it had exceeded St. Louis, Mo., in population. It was the site of the 1860 Republican National Convention at which Illinoisan Abraham Lincoln won the presidential nomination. Both Americans and northern European immigrants, drawn by Chicago's factories and carried by the rail network that was anchored in Chicago, continued to pour into the city.

The Great Chicago Fire

Four square miles of Chicago, including the business district, were destroyed by fire on October 8–10, 1871. Starting in the southwest, fed by wooden buildings and pavements and favoured by a long dry spell, flames spread northeastward, leaping the Chicago River and dying out only when they reached Lake Michigan. About 250 lives were lost, some 90,000 people were made homeless, and almost $200,000,000 in property was destroyed.

**The rebuilt city and its people.** Much of the city's physical infrastructure remained, however, including its water-supply and sewage systems and transportation facilities. Chicago rebuilt rapidly in a similar pattern, although with buildings that were more modern and in conformance with new fire regulations.

The central business district, bounded by the Chicago River to the north and west and by the railroad along the lakeshore to the east, held the major department stores, the larger banks, the Board of Trade, the regional headquarters of rising national corporations, and the centres of commerce, law, and government. The district was the birthplace of the steel-frame skyscraper. Completion of the Home Insurance Building in 1885 led during the next nine years to the construction of 21 buildings ranging from 12 to 16 stories throughout the downtown area. Commuter railroads, horse, cable, and electric street railways, and elevated rapid-transit lines served the Loop.

The Lake Michigan shore became the centre for the homes and civic pursuits of Chicago's economic and social elite. Lake Shore Drive north of the Loop emerged as the mainline for society—the Gold Coast, it was soon nicknamed. Although blighted by the Illinois Central Railroad

yards, the waterfront land to the east of the Loop was nevertheless landscaped and named Grant Park.

Heavy industry, warehouses, and rail yards crowded the banks of the Chicago River. Industrial pockets also existed at Chicago's outskirts. At the far south, where the Calumet River met Lake Michigan, the steel mills drew a polyglot community of blue-collar workers and their families. The Union Stock Yards dominated another South Side area, the Back-of-the-Yards district, which was made infamous in Upton Sinclair's scathing novel of industrial oppression, *The Jungle* (1906). Public health and sanitary conditions were an outrage: until 1900 Lake Michigan both supplied fresh water to Chicago and received its untreated sewage, a condition probably responsible for the city's frequent epidemics.

The second wave of foreign immigrants

Many of the working families arrived in the second great wave of European immigration: Russian Jews, Italians, Poles, Serbs, Croatians, Bohemians, and other groups from southern and eastern Europe. The 1890 and 1900 censuses showed that more than three-fourths of Chicago's population was made up of the foreign-born and their children.

The working districts were fertile ground for social action. The labour movement left the mark of its early attempts at industrial organizing: the Haymarket Riot of 1886, in which workers and lawmen alike died; and an 1894 strike against the Pullman Palace Car Company, led by pioneer organizer Eugene V. Debs and others. Social work was another influence: Jane Addams and her followers at Hull House, a West Side settlement, tried to improve the wretched conditions of housing and health there.

In 1889 Chicago annexed numerous inner suburbs, doubling its area and its population (to almost 1,100,000) and surpassing Philadelphia as America's second most populous city. By 1900 it was a centre of nearly all parts of the U.S. economy as well as of social insurgency and reform, immigration, education, and even culture. Chicago also had developed a brawling spirit evident not only along the dingy streets of the immigrant ghettos but also in corporate boardrooms and in the most elegant brothel in the nation, which entertained royalty from abroad and millionaires from the newly sprawling suburbs.

This Chicago was particularly striking to writers and visitors. "I have struck a city—a real city—and they call it Chicago," wrote Rudyard Kipling. "The other places don't count." And, he continued, "Having seen it, I urgently desire never to see it again. It is inhabited by savages."

**Symbols of civic consolidation.** A major expression of the city's character was the Plan of Chicago (1909), by Daniel H. Burnham and Edward H. Bennett, which took the general outlines of turn-of-the-century Chicago, added the notions of style possessed by the city's industrial and mercantile elite, and presented a vision of the future. The plan was inspired by the 1893 World's Columbian Exposition—for which Chicago outbid New York City, Philadelphia, and Washington, D.C.—celebrating the 400th anniversary of the discovery of America. Built on the Midway Plaisance adjacent to the University of Chicago (endowed in 1891 by John D. Rockefeller), the exposition's buildings have been called a stylistic union of Classical Greece, Imperial Rome, Renaissance Italy, and Bourbon Paris.

The World's Columbian Exposition of 1893

Nonetheless, the exposition stimulated activity in city planning not only in Chicago but also throughout the world. The "City Beautiful" movement dominated civic thought for several decades, influencing even some federal buildings in Washington, D.C. The Classicism of the exposition was in marked contrast, however, to the modern Chicago School of architecture, and the two trends proceeded concurrently during the following decades. Chicago became a world centre of architectural innovation in the late 19th and early 20th centuries, with many notable buildings by Dankmar Adler, Louis Sullivan, Frank Lloyd Wright, and Henry H. Richardson.

Features of the Burnham Plan

The Burnham Plan, as it came to be called, proposed many subsequently developed features: park areas along Lake Michigan that included beaches, boulevards, and yacht basins; a belt of forest preserves rimming the city for recreation; the widening of arterial streets; a civic centre; and a double-decked boulevard in the central area along the Chicago River. Until 1939 the quasi-official Chicago Plan Commission promoted individual features of the plan, which, like Burnham's admonition, "Make no little plans," came to have a profound effect on Chicago.

**The 20th century.** Chicago's population growth was less spectacular in the 20th century, though industrial expansion associated with World Wars I and II and the postwar prosperity continued to attract newcomers. Most pronounced was the influx of blacks from the South seeking industrial employment. A building boom in the city and suburbs terminated abruptly following the stock-market collapse of 1929, and during the next decade the population increased only slightly, to about 3,400,000 in 1940. Possibly contributing to this slowed growth were the worldwide notoriety of Chicago (only in part deserved) as the playground of underworld figures during the Prohibition era, the failure of several Chicago banks during the Depression of the 1930s, and the allegedly powerful grip of criminal syndicates on many aspects of economic and political life. In contrast, however, the suburban population increased rapidly during this period.

After World War II construction was slow to resume until Daley's election in 1955. Massive rebuilding programs became a hallmark of his terms in office, including an almost total alteration of the skyline of the Loop and adjacent areas. As in most cities the downtown area, although it continued as the centre for offices, suffered from a decline in other functions, including retailing, entertainment, and wholesale distribution, while rapid expansion of those activities took place on the periphery and in suburban areas. The 1970 census revealed that, for the first time, the city itself had less than half of the metropolitan population.

BIBLIOGRAPHY. General works include IRVING CUTLER, *Chicago: Metropolis of the Mid-Continent,* 3rd ed. (1982), and (ed.), *The Chicago Metropolitan Area: Selected Geographic Readings* (1970); FEDERAL WRITERS' PROJECT, *Illinois,* new rev. ed. (1974); and HARRY HANSEN, *The Chicago* (1942), in the "Rivers of America" series. *Chicago* magazine (monthly) provides articles of local interest and reviews of cultural and entertainment events in the region. Historical development of the city and region are detailed in BESSIE L. PIERCE, *A History of Chicago,* 3 vol. (1937–57, reissued 1975); HOMER HOYT, *One Hundred Years of Land Values in Chicago* (1933, reissued 1982); and HAROLD M. MAYER and RICHARD C. WADE, *Chicago: Growth of a Metropolis* (1969). Aspects of Chicago history are covered in HERMAN KOGAN and ROBERT A. CROMIE, *The Great Fire: Chicago, 1871* (1971); and LOIS WILLE, *Forever Open, Clear and Free: The Historic Struggle for Chicago's Lakefront* (1972). Ethnic aspects are covered in OTIS D. DUNCAN and BEVERLY DUNCAN, *The Negro Population of Chicago: A Study of Residential Succession* (1957); CHICAGO DEPARTMENT OF DEVELOPMENT AND PLANNING, *Historic City: The Settlement of Chicago* (1976); and PETER D'ALROY JONES and MELVIN G. HOLLI (eds.), *Ethnic Chicago* (1981). Materials on Chicago politics and government include LLOYD LEWIS and HENRY JUSTIN SMITH, *Chicago: The History of Its Reputation* (1929, reprinted 1973); PETER H. ROSSI and ROBERT A. DENTLER, *The Politics of Urban Renewal: The Chicago Findings* (1961, reprinted 1981); MARTIN MEYERSON and EDWARD C. BANFIELD, *Politics, Planning, and the Public Interest: The Case of Public Housing in Chicago* (1955); MIKE ROYKO, *Boss* (1971), an attack on the Democratic Party Organization of Mayor Daley; SAMUEL K. GOVE and LOUIS H. MASOTTI (eds.), *After Daley: Chicago Politics in Transition* (1982), an anthology of scholarly papers.

Basic works on planning in Chicago are DANIEL H. BURNHAM and EDWARD H. BENNETT, *Plan of Chicago* (1909, reprinted 1970); CHICAGO DEPARTMENT OF DEVELOPMENT AND PLANNING, *Comprehensive Plan of Chicago* (1966), and *Chicago 21: A Plan for the Central Area Communities* (1973); and NORTHEASTERN ILLINOIS PLANNING COMMISSION, *The Comprehensive General Plan for the Development of the Northeastern Illinois Counties Area* (1968, reissued 1975). The city's architectural heritage is covered in CARL W. CONDIT, *The Chicago School of Architecture* (1964, reissued 1975), on the years 1875–1925, *Chicago, 1910–29: Building, Planning and Urban Technology* (1973), and *Chicago, 1930–70* (1974); IRA J. BACH, *Chicago on Foot: Walking Tours of Chicago's Architecture,* 3rd ed. (1977), and, with ROY FORREY (eds.), *Chicago's Famous Buildings: A Photographic Guide to the City's Architectural Landmarks and Other Notable Buildings,* 3rd ed., rev. and enl. (1980). A book combining history, ethnic interest, and architecture is GLEN E. HOLT and DOMINIC A. PACYGA, *Chicago: A Historical Guide to the Neighbourhoods: The Loop and the South Side* (1979).

(H.M.M.)

# Childhood Diseases and Disorders

The term childhood diseases denotes those diseases that characteristically occur during an age span that begins with the fetus and extends through adolescence. This is a period typified by change, both in the child himself and in his immediate environment. Changes in the child related to growth and development are so striking that it is almost as if the child were a series of distinct yet related individuals as he passes through infancy, childhood, and adolescence. Changes in the environment occur as the surroundings and contacts of a totally dependent infant become those of a progressively more independent child and adolescent. Health and disease during the period from conception to adolescence must be understood against this backdrop of changes.

Although, for the most part, the diseases of childhood are similar to those of the adult, there are several important differences. For example, certain specific disorders, such as precocious puberty, are unique to children; others, such as acute nephritis—inflammation of the kidney—are common in children and infrequent in adults. At the same time, some diseases that are common in adults are infrequent in children. These include essential hypertension (high blood pressure of unknown cause) and gout. Finally, a major segment of pediatric care concerns the treatment and prevention of congenital anomalies, both functional and structural.

Apart from variations in disease due to differences between children and adults, certain other features of diseases in children need to be emphasized. Infectious disorders are prevalent and remain a leading cause of death, although individual illnesses are often mild and of minor consequence. Most instances of the common communicable diseases, such as measles, chicken pox, and mumps, are encountered in childhood. Disorders of nutrition, still of great concern, especially but not exclusively in developing countries, are of extreme importance to the growing and developing child. The unique nutritional requirements of children make them unusually susceptible to deficiency states: vitamin-D deficiency causes rickets, a common disorder of children in developing countries, and only rarely causes any disease in adults. The major environmental hazards that endanger the health of young children are either unavoidable, as in air pollution, or accidental, as in poisoning and in traffic injuries. Older children, especially adolescents, are exposed, as are adults, to environmental hazards that they deliberately seek, such as cigarette smoking and the use of alcohol and other drugs.

This article reviews the scope of diseases that affect children, with particular emphasis on the ways in which the unique attributes of the growing child and special aspects of his environment serve to modify the course, effects, and treatment of particular diseases.

The article is divided into the following sections:


Diagnosis and general considerations of treatment and prevention 9
Disease-affecting differences between children and adults 10
  Anatomical differences 10
  Physiological differences 11
Disorders present at birth 11
  Diseases transmitted through the placenta or due to placental dysfunction 11
  Injuries incurred during birth 11
  Prematurity and low birth weight 12
  Metabolic disturbances 12
  Infections 12
  Respiratory disorders 13
  Cardiovascular disorders 13
  Blood disorders 13
  Gastrointestinal disorders 13
  Kidney and urinary-tract disorders 14
  Nervous-system disorders 14
  Endocrine disorders 14
  Musculoskeletal disorders 14
  Skin disorders 14
  Chromosomal disorders 14
Disorders of later infancy and childhood 15
  Sudden infant death syndrome (SIDS) 15
  Failure to thrive 15
  Malnutrition 15
  Classic infectious diseases of childhood 15
  Respiratory disorders 16
  Cardiovascular disorders 16
  Blood disorders 16
  Gastrointestinal and liver disorders 17
  Kidney and urinary-tract disorders 17
  Nervous-system disorders 18
  Endocrine disorders 18
  Skin disorders 19
  Connective-tissue disorders 19
  Accidents 19
  Child abuse and neglect 19
  Psychological disorders 19
Disorders associated with adolescence 20
Bibliography 20


## DIAGNOSIS AND GENERAL CONSIDERATIONS OF TREATMENT AND PREVENTION

Diagnosis of the diseases of childhood involves special considerations and techniques; for example, in evaluating genetic disorders, not only the patient but his entire family may need to be examined. Inapparent environmental causes of diseases, such as poisonings, must be considered and investigated thoroughly, by methods that at times resemble those of a detective. Diseases of the fetus may derive directly from disorders of the mother or may be caused by drugs administered to her. Diagnostic techniques have been developed that permit sophisticated examination of the fetus despite its apparent inaccessibility. The withdrawal of a small amount of the amniotic fluid that surrounds the fetus permits examination of fetal cells as well as the fluid itself. Chromosomal and biochemical studies at various stages of development may help to anticipate problems in the postnatal period; they may indicate the need for immediate treatment of the fetus by such techniques as blood transfusion; or they may lead to the decision to terminate pregnancy because serious, untreatable disease has been recognized. Other specialized techniques permit examination of the fetus by X ray and ultrasound, and by electrocardiography and electroencephalography (methods for observing and recording the electrical activity of the heart and the brain, respectively). Fetal blood can be obtained for analysis, and certain techniques permit direct viewing of the fetus.

In examination of the infant, inaccessibility is no special problem, but his small size and limited ability to communicate require special techniques and skills. Of even more importance, however, is the fact that adult norms cannot be applied to younger age groups. Pediatric diagnosis requires knowledge of each stage of development, with regard not only to body size but also to body proportions, sexual development, the development and function of organs, biochemical composition of the body fluids, and the activity of enzymes. The development of psychological and intellectual function is equally complex and requires special understanding. Since the various periods of growth

tages of rowth nd devel- pment

and development differ so markedly from one another, they are divided for convenience into the following stages: intrauterine (the period before birth), neonatal (first four weeks), infant (first year), preschool (one to five years), early school (six to 10 years for girls, six to 12 for boys), prepubescent (10 to 12 for girls, 12 to 14 for boys), and adolescent (12 to 18 for girls, 14 to 20 for boys). Only if appropriate norms are established for each stage of development can the child's condition be adequately evaluated and the results of diagnostic tests properly interpreted. Thus, it is of no concern if a 12-month-old infant is unable to walk alone, although some infants are able to do so at nine months of age. The crucial question is at what age one becomes concerned if a developmental milestone has not been reached. Five-year-old boys average 44 pounds (20 kilograms) in weight but may vary from 33 to 53 pounds (15 to 24 kilograms). The hemoglobin level that is of no concern in the three-month-old infant may reflect a serious state of anemia in the older child. The blood levels of certain enzymes and minerals differ markedly in the rapidly growing child from those in the late adolescent, whose growth is almost complete. Failure of a 15-year-old girl to have achieved menarche (the beginning of menstruation) may be indicative of no abnormality in sexual development but requires careful evaluation.

Treatment of childhood disease requires similar considerations with regard to various stages of growth and development. Variation in drug dosage, for example, is based not only on body size but also on the distribution of the drug within the body, its rate of metabolism, and its rate of excretion, all of which change during various stages of development. The inability of infants and small children to swallow pills and capsules necessitates the use of other forms and alternate routes of administration. Drug toxicity of importance at one stage of development may be of no concern at another; for example, the commonly used antibiotic tetracycline is best avoided in treatment of the child younger than age 10 because it is deposited in teeth, in which enamel is also being deposited, and stains them. When permanent teeth are fully formed, the deposition of tetracycline no longer occurs. The delayed consequences of certain forms of treatment, especially with radioactive isotopes—substances that give off radiation in the process of breaking down into other substances—might be of no consequence in the case of an elderly person with a life expectancy of 10 or 20 years but might deter a physician from the use of such treatments for the infant with his whole life in front of him. Finally, the nutritional requirements of the growing child must be considered when treatment of disease requires modification of the diet or administration of drugs that may affect the absorption or metabolism of essential nutrients.

rognosis

The outlook for recovery from diseases in children is often better than it is for adults, since the child's additional capacity of growth may counteract the adverse influences of disease. The bone fracture that results in permanent deformity in the adult, for example, may heal with complete structural normality in the child, as continued growth results in remodeling and reshaping of the bone. Ultimately, the infant who has one kidney removed because of infection or tumour most likely will have entirely normal renal (kidney) function because the remaining kidney will increase its size and functional capacity with growth. In contrast, removal of one kidney in the adult usually results in a residual functional capacity equal to 70 to 75 percent of that of two normal kidneys.

Thus, being in a period of rapid growth and development may favourably affect the child's recovery in the course of a disease. The converse may also be true, however. The rapidly growing and maturing central nervous system, for example, is particularly susceptible to injury during the first two or three years of life; also, adolescents may react unfavourably to psychological stresses that are tolerated readily by more mature individuals.

revention f hildhood isease

In the general consideration of childhood diseases, a final aspect that merits emphasis is the role of prevention. The major factors responsible for the decline in infant and childhood mortality rates over the past decades have been the development and application of preventive measures. By the late 20th century, in most countries the death rate for infants under one year of age had decreased until it was scarcely more than a 10th of the rate in the 1930s. Socioeconomic factors—such as better maternal nutrition and obstetrical care and improved housing, water supplies, and sewage disposal—have been of prime importance in this decline, together with better hygiene at home, safer infant feeding techniques, and widespread immunization against common infectious diseases. In comparison to the favourable effect of these and other preventive measures, an increased capacity to treat diseases, even with such powerful tools as the antibiotic drugs, has had relatively little impact. In the developed countries, where the most common causes of childhood morbidity and mortality are accidents, prevention depends upon a willingness to design and modify communities and homes to make them safer for children. Just as important as the development of public health measures is their practical application; underutilization of established procedures and techniques for prevention of disease is a major health problem.

## DISEASE-AFFECTING DIFFERENCES BETWEEN CHILDREN AND ADULTS

Disturbances in growth may be among the most striking consequences of disease in children. An obvious example of this effect is total growth failure, which is seen in almost every serious disease of infants and children. Local retardation or disturbance in growth patterns may be equally striking. Osteomyelitis, an infection of bone, may, for example, result in retardation or cessation of growth at that site, with subsequent severe asymmetry between the affected limb and its normal counterpart. Enlargement of the heart as a result of cardiac disease may cause gross distortion of the chest, as the growing ribs adapt to the abnormal shape of the heart.

Many differences in the manifestations of disease in children and adults can be ascribed to differences in anatomical structure and in biochemical, immunological, and physiological function. Less well understood are the consequences of differences in psychological function. In general, the younger the child, the more striking these differences are.

**Anatomical differences.** Not only is the child's body smaller than that of the adult, but it has different proportions; for example, the sitting height of the newborn infant represents about 70 percent of total body length. With rapid growth of the extremities, sitting height decreases to about 57 percent of the body length at three years of age and, finally, as growth proceeds more slowly, to the adult proportion of about 50 percent. Growth and development are not necessarily smooth, continuous processes. Weight and height increase rapidly in infancy and at puberty; for example, the head completes half its total growth in the first year of life, and by the age of two years the child has reached half his adult height. In addition to differences in proportion and size, there are marked differences in body composition between children and adults. As examples, in newborn infants muscle mass constitutes approximately 25 percent of total body weight, compared to 43 percent in adults. Total body water, which accounts for 90 percent of early fetal weight, represents 75 percent of body weight at birth, drops to about 60 percent by one year of age, and then declines gradually to reach the adult figure of 55 percent. The higher proportion of body water, due almost entirely to a relatively greater volume of fluid outside the cells, affects the response of the infant, particularly to disturbances in water balance.

Differences in body composition

There are many examples of differences in anatomical structure that affect manifestations of disease. In assessing the health of the infant with cardiac or pulmonary (lung) disease, the thinner chest wall, the relatively more horizontal position of the heart, and the more rapid cardiac and respiratory rates must be taken into account. The thin abdominal wall of the infant permits palpation—examination by touching with the fingers—of the kidneys, whereas in older subjects the kidneys usually can be felt only if they are abnormally large. In the infant, with the bones of the skull still not fused together, obstruction of the flow of cerebrospinal fluid may result in striking enlargement

of the head, a condition referred to as hydrocephalus. In the older child, when the skull sutures have fused, such enlargement is not possible, and the manifestations of spinal-fluid obstruction are similar to those of the adult, including severe headache and visual difficulties as a result of increased intracranial pressure. The primary manifestation of mumps is a painful swelling of the parotid and other salivary glands. In adolescents, involvement of the testes or ovaries occurs only rarely, a phenomenon related in some way to the immaturity of these organs. In the adult, particularly in the male, severe sex-gland involvement is common.

**Physiological differences.** Physiological differences between children and adults that cause differences in the manifestations of disease include all the various functional, endocrine, and metabolic features of the growing and maturing organism. A major characteristic in this regard is the limited ability of the infant to maintain homeostasis (a stable internal environment) during illness because of his greater metabolic and nutritive requirements. Moreover, most of the first year of life is characterized by immaturity of renal function, the capacity of the kidneys to respond to the stresses of disease being less than later in life. The baby with severe diarrhea, for example, cannot conserve water well enough and may become dehydrated. With any degree of stress, metabolic abnormalities are likely to be more severe in the infant than in the older child.

The liver of the newborn child also demonstrates certain features of immaturity. Of particular importance is its limited capacity to excrete bilirubin, a product of the breaking down of hemoglobin (the oxygen-carrying pigment of red blood cells). In certain conditions in which there is a rapid rate of destruction of red blood cells, the inability of the liver to excrete the added load of bilirubin may result in a large increase in the concentration of this substance in the blood; the bilirubin concentration, if high enough, can cause severe brain damage known as kernicterus. Since immaturity of the brain also contributes to the infant's increased susceptibility to this disorder, kernicterus is rarely encountered outside of the neonatal period, even in subjects with severe liver disease.

The ability of the young infant to metabolize and to excrete certain drugs is limited by the immaturity of the liver and of the kidney, and drug dosage must be adjusted accordingly.

Immunologic responsiveness

The immunologic system of the body is responsible for the defense against disease. This highly complex system involves the production of antibodies (proteins that can recognize and attack specific infectious agents); the action of granulocytes and macrophages, cells that destroy infecting organisms by ingesting them (a process called phagocytosis); and the function of a variety of cellular mechanisms involving the complement system (complement is an enzyme-like substance in the blood). Antibody production in the infant is qualitatively and quantitatively different from that in the older child and adult. Although the differences in antibody response cannot be related specifically to differences in the capacity of the infant to withstand infection, they certainly must play some role. On the other hand, many of the clinical features of infectious disease occurring during the first two or three years of life appear to be related to the fact that these are infections occurring for the first time.

Another difference in immunologic response between children and adults is in the functioning of the reticuloendothelial system. This system, which is composed of the macrophages found in the lymph nodes, spleen, and other lymphatic tissues, is relatively more active in childhood. Since macrophages ingest infectious organisms, children with coryza or sore throats commonly have swollen lymph "glands" visible and palpable in the neck. Similarly, their tonsils and adenoids, which are lymphatic tissues, swell rapidly in response to mild infections.

## DISORDERS PRESENT AT BIRTH

**Diseases transmitted through the placenta or due to placental dysfunction.** Infectious diseases of the fetus are caused by many different types of organisms, including viruses, bacteria, spirochetes, and protozoa (*e.g.*, toxoplasmosis). Most of these infections are the result of infection of the mother, the infectious agents being transmitted through the placenta (the temporary organ by means of which the fetus receives nourishment and discharges waste) by way of fetal circulation. Bacterial infection is most often associated with premature rupture of the membranes and infection of the amniotic fluid.

Maternal rubella (German measles) occurring during the first eight weeks of pregnancy is associated with congenital malformation of the fetus in more than 50 percent of cases, the figure decreasing to about 20 percent by the 16th week and dropping sharply thereafter. Infection of the fetus with a virus of the cytomegalovirus type involves many organs, has a high fatality rate, and may result in severe brain damage in fetuses who survive. Infection by the intracellular parasite *Toxoplasma gondii* produces a disease called toxoplasmosis, which may cause death or may result in microcephalus (abnormal smallness of the head), hydrocephalus (excessive accumulation of fluid in the brain cavities), or mental retardation. Congenital syphilis may have a variety of effects in the infant, including involvement of the skin, liver, spleen, lymph nodes, and kidneys. Malformations of the bones and teeth appear later, and severe involvement of the central nervous system may become apparent after many years.

Just as infectious agents may cross the placenta, so also most drugs administered to the mother may pass through the placenta and have important effects on the fetus. A most dramatic and devastating example of this effect occurred in Europe during the early 1960s, when the birth of thousands of infants with absent or short limbs resulted from the maternal ingestion of the apparently harmless drug thalidomide. Anesthetics, analgesics (pain relievers), sedatives, antihypertensive drugs, and antibiotics all may have adverse effects on the fetus. Congenital goitres (enlargement of the thyroid) have been produced by administration of antithyroid drugs to the mother. It is now clear that adverse effects on the fetus must be considered whenever drug therapy of the mother is contemplated.

The abuse of narcotics or alcohol by the mother can also lead to dire fetal consequences. Infants born to mothers addicted to heroin, morphine, or other opiates commonly share their mothers' addiction and suffer withdrawal symptoms within 72 hours of birth. Many infants of alcoholic mothers are afflicted with a combination of malformations, known as the fetal alcohol syndrome, which include mental retardation, growth retardation, and microcephaly.

The entire nutrient supply of the fetus derives from the mother. Although maternal deficiency states may, therefore, be reflected by parallel deficiencies in the fetus, in general the needs of the fetus will be met ahead of those of the mother, and an adequate amount of a given nutrient may be supplied to the fetus, despite maternal deficiency. Mild to moderate deficiencies of iron or calcium in the mother, for example, are not usually associated with deficiencies in the fetus. On the other hand, protein and caloric malnutrition may be associated with decreased fetal size.

Placental insufficiency

Deficiencies in the fetus may also arise from placental dysfunction (malfunctioning). The consequences of abnormalities of the placenta depend upon the time of onset and the severity of placental inadequacy. Serious placental insufficiency early in pregnancy may result in the death of the fetus. It is also likely that placental insufficiency can be a factor in decreasing fetal growth. Toward the end of pregnancy, placental dysfunction is associated with premature delivery or evidence of varying degrees of fetal distress, ranging from yellow staining of the skin to fetal wasting and to signs of severe lack of oxygen.

**Injuries incurred during birth.** The physical trauma of delivery may result in a number of injuries to the infant. Of little consequence is the diffuse soft-tissue swelling of the scalp referred to as caput succedaneum. Difficult delivery may result in more extensive bruising, abrasions, and edema—particularly after breech delivery; however, serious harm is rare. Bleeding under the periosteum (the covering membrane) of the skull produces a large swelling in 1 to 2 percent of babies, and in some it is associated with a small fracture of the underlying skull; fortunately, spontaneous healing occurs speedily. Injuries to the spinal

cord are rare, but injuries of peripheral nerves as a result of traction on the head are not uncommon. Such injuries include Erb's paralysis, with weakness of the arm and shoulder because of damage to the fifth and sixth cervical nerves. Injury to the phrenic nerve, with paralysis of the diaphragm—the muscular partition between the chest and the abdomen—and facial-nerve injury resulting in facial palsy also are encountered. In the vast majority of such instances of peripheral-nerve injury, recovery is complete.

An extremely important form of birth injury is that associated with lack of oxygen (anoxia). Fetal anoxia may occur from inadequate oxygenation of the mother, low maternal blood pressure, or abnormalities in the uterus, placenta, or umbilical cord that result in inadequate blood flow to the fetus. After birth, anoxia may result from blood loss, shock, or inadequate respiration. Clinical manifestations include decreased activity, slowing of the heart, and blueness of the skin (cyanosis). Severe anoxia may cause death of the newborn, although recovery is more common. The major significance of anoxia is that it may result in brain damage if prolonged more than a few minutes.

**Prematurity and low birth weight.** The usual length of the gestation period is 40 weeks. Infants born prior to 37 weeks of gestation are considered to have been born early and are referred to as preterm or premature. Infants who at birth weigh 2,500 grams (about 5.5 pounds) or less are considered to be of low birth weight and either are prematurely born or have had less than the expected rate of growth within the uterus.

Definition of prematurity

Infants whose weight is low at birth account for as many as 10 to 15 percent of births among low socioeconomic groups and as few as 4 to 5 percent of births among those of higher socioeconomic status. Clinical examination of the baby helps to differentiate between the preterm baby and the small baby born at term, but such determination of gestational age (the age from conception to birth) is not precise. The correct classification of the baby is important because maturity, in terms of gestational age, is a major factor determining the ease with which the baby will adapt to life outside the uterus. In the infant born too early, many organ systems will not be fully developed. The preterm infant who is large and only slightly immature does as well as the full-term infant, but the very small preterm infant, below 1,000 grams (about 35 ounces) in weight, has a high fatality rate and is prone to many complications.

The complications encountered in coping with extrauterine existence involve primarily the respiratory and gastrointestinal systems. In addition to anatomical immaturity of the lungs, a handicapping feature of the premature infant may be a lack of a substance called a surfactant, which plays an important role in permitting the air spaces, or alveoli, of the lungs to remain open. Surfactant appears in some fetuses at 24 weeks' gestation but is absent in others until about 30 weeks. Because of these respiratory handicaps—particularly the lack of surfactant—many premature infants suffer from respiratory distress syndrome, a condition described below under *Respiratory disorders.*

Inability to suck adequately and limitations in the capacity to digest foodstuffs and absorb them through the intestinal tract provide other serious handicaps for the premature infant. To circumvent these problems, infants may be fed (by stomach tube) specially prepared formulas tolerated by even the smallest of babies.

The relatively large surface area of the small infant and his inability to maintain body temperature may require his being kept in an incubator. In addition to temperature control, the incubator makes it possible to provide extra oxygen to the infant who has respiratory difficulties, although this must be done with care because excessive oxygen may lead to damage to the eyes, a condition known as retrolental fibroplasia.

Survival of premature infants

As indicated above, the prematurely born infant is considerably less likely to survive than are full-term infants. Premature infants, accounting for less than 8 percent of all live births, account for two-thirds of infant deaths. Even after the first year of life, the mortality rate among infants with low birth weights is greater than among infants with birth weights above 2,500 grams. The cause of this increased rate is not completely known, although a higher prevalence of congenital anomalies accounts for some of the difference. Moreover, retarded intellectual development and other abnormalities of the nervous system are more common in such infants, particularly those with birth weights of less than 1,500 grams (3.3 pounds). The majority of infants with low birth weights remain small throughout the childhood years, which may reflect a continued pattern of slow growth, first evidenced in the uterus.

**Metabolic disturbances.** Infants of diabetic mothers represent a unique group with special metabolic problems. Intrauterine death is common and unexplained. The placenta is often abnormal. The infants at birth generally are large and have large organs, a condition referred to as macrosomia. Respiratory distress and low levels of sugar in the blood (hypoglycemia) are common complications.

Neonatal hypoglycemia is a relatively common disorder, particularly among infants whose birth weight is low. Fifteen percent of hypoglycemic infants have associated abnormalities of the central nervous system. In most instances hypoglycemia is transient and responds readily to treatment.

Jaundice in the newborn is ordinarily related to an imbalance between the rate of destruction of red blood cells and the metabolism of hemoglobin to bilirubin and the rate of excretion of bilirubin in the bile; there is a resultant temporary elevation of bilirubin level in the blood. Jaundice may, however, be due to septicemia, to several different diseases of the liver, or to obstruction of the ducts through which bile flows into the intestinal tract. Abnormally high bilirubin levels have also been found in association with breast feeding; it is an extremely rare condition resulting from the presence of an unusual substance in the milk.

The significance of jaundice depends on the underlying cause and the amount of excess bilirubin in the blood. In extreme cases, bilirubin can be deposited in brain cells, resulting, as mentioned above, in severe nerve-cell damage, called kernicterus. This condition, which may lead to deafness and cerebral palsy, is encountered most often in infants with erythroblastosis fetalis, a blood disorder discussed below. Brain damage from an excess amount of bilirubin can usually be prevented by means of exchange transfusions (in which most of the infant's blood is replaced with blood from donors), which in the most severe cases may need to be repeated many times.

Kernicterus

Tetany of the newborn, a condition that appears within a few days after birth, is characterized by increased neuromuscular irritability, with muscular twitching, tremors, and convulsions. In most cases, the blood concentration of calcium is low, and that of inorganic phosphate is high. In some infants the disorder appears to be due to a low concentration of magnesium in the blood. The infant's condition is usually dramatically improved by the intravenous administration of calcium. The disorder is transient, so that treatment with oral calcium supplements can be discontinued after one or two weeks.

In contrast to the metabolic disturbances described above, which are generally transient conditions of the newborn, are the long-term disorders known as the inborn errors of metabolism. These result from the absence of a functional enzyme in a particular metabolic pathway. Because of this "enzyme block," there is a deficiency in the products of the affected pathway and an excessive build-up of harmful chemicals that cannot be processed normally. Inborn errors of metabolism are genetically determined, and most are very rare. Many lead to severe illness and brain damage unless effective and early treatment can be started. A well-known example is phenylketonuria, which can be detected by a simple blood-screening test (the Guthrie test) during the first week of life. Once identified, the affected infant is given a special diet that prevents brain damage and allows normal growth. The diet has to be continued until at least the age of 10 years, and some clinicians recommend that it be followed for life.

Inborn errors of metabolism

**Infections.** The newborn infant is subject to the ordinary infections and, in addition, to infection with commonly encountered organisms such as *Escherichia coli, Staphylococcus aureus,* and group B hemolytic streptococci, which are not usual causes of serious infection in

older age groups. Infection may be acquired in the uterus, during delivery, or later, in the nursery. Commonly encountered serious infections are pneumonia, meningitis (inflammation of the coverings of the brain and spinal cord), and septicemia (infection of the bloodstream). Often the infant shows few signs of the disease other than poor feeding, lethargy, pallor, or slight fever. Since the newborn infant's resistance to infection is poor, early diagnosis and treatment are particularly important. Often, treatment is given when infection is merely suspected.

Common infections

Congenital defects of each part of the immunologic system have been discovered. The most striking feature of these diseases is the inability of the patient to combat infection. Thus, untreated patients with some forms of agammaglobulinemia (lack of antibodies in the blood) may die from overwhelming infection in infancy or early childhood.

**Respiratory disorders.** Numerous abnormalities of respiratory function are common in the newborn infant. One of the most severe is respiratory distress syndrome (RDS; also called hyaline membrane disease). RDS occurs in 0.5 to 1 percent of all deliveries, and, as previously mentioned, is especially common in premature infants. In addition, it is encountered commonly in infants of diabetic mothers and after cesarean section (delivery through the wall of the mother's abdomen). RDS also occurs, albeit infrequently, in full-term infants without any apparent predisposing cause. Soon after birth, affected infants begin to take rapid, shallow breaths and can be shown by appropriate tests to be exchanging air (*i.e.,* absorbing oxygen and exhausting carbon dioxide) only poorly. Without expert treatment, they may die within a few hours or may have a protracted course over a period of several days, with later demise or gradual improvement and recovery. Treatment is directed at relieving the symptoms and includes correction of an associated acidosis, administration of oxygen, and assisting the infant to breathe, if necessary with a mechanical ventilation machine. With modern care, death has become less common.

Respiratory distress syndrome

Pneumonia is in infants a serious problem. The onset is either within hours after birth, in infants whose infection is contracted from the mother, or after 48 hours of life, when the infection is acquired after birth. Infants show signs of difficulty in breathing, and often there is an associated infection of the blood (septicemia). Treatment consists of the administration of carefully selected antibiotics in appropriate dosages and respiratory support.

Pneumonia

An infant may inhale meconium (a semisolid discharge from the infant's bowels) during the course of delivery, leading to obstruction of the upper airway. Clearing the airway with suction, the administration of oxygen, and general respiratory support are usually effective in promoting recovery within two to three days.

Leakage of air into the pleural space (between the membrane lining the chest and that enveloping the lungs and other thoracic organs), with consequent partial or complete collapse of the lung (pneumothorax), bleeding into the lung, and failure of expansion of the lung (atelectasis), also causes respiratory failure in the newborn infant. Prompt treatment is often necessary to ensure survival.

**Cardiovascular disorders.** Cardiovascular disturbances in the newborn are related primarily to congenital malformations that affect about seven out of every 1,000 infants. They vary from those that are incompatible with life to those that cause no illness and require no treatment. Sometimes the cause is known because of an association with a chromosomal disorder (*e.g.,* Down's syndrome and Turner's syndrome; see below); in a few the cause is maternal rubella infection. The lesions arise early in fetal development, and the result is usually either an obstruction of normal blood flow or an abnormal communication between different parts of the heart or the circulation. When the structural abnormality causes severe disturbance, heart failure results. The baby in heart failure may present such symptoms as a blue complexion (cyanosis), breathlessness, or feeding difficulties. Most congenital heart defects are associated with heart murmurs that can be heard with a stethoscope. The most common congenital lesion is a ventricular septal defect, which is a hole between the two lower chambers of the heart (the left and right ventricles). Many of these close spontaneously without treatment. Diagnosis of an infant with suspected congenital heart disease has been made surer and easier with the development of echocardiography. Most of the disorders that cause illness can be corrected by surgery, which—unless the defect is immediately life-threatening—is usually deferred until the child is older. For a fuller discussion of specific congenital cardiovascular defects, see CIRCULATION AND CIRCULATORY SYSTEMS: *Cardiovascular system diseases and disorders; Congenital heart disease.*

A specific cardiovascular problem common in the preterm infant is patent ductus arteriosus, which is the persistence of an essential feature of fetal circulation. The ductus arteriosus is a fetal blood vessel that connects the descending aorta and the pulmonary artery. It shunts blood from the lungs (which are nonfunctional in the fetus), channeling it toward the placenta (where oxygenation takes place). Normally, the ductus closes shortly after birth. When it remains patent (open) after birth, it functions as a shunt in the opposite direction, diverting blood from the aorta to the lungs. Thus, too much blood is delivered to the lungs, and the subsequent pulmonary congestion causes breathing difficulties. Drugs can be given to encourage the ductus to close. If drug treatment proves ineffective, the ductus may be closed surgically.

**Blood disorders.** The diseases affecting the blood of newborn infants include diseases of the red blood cells (particularly the anemias, which involve an inadequate level of functioning hemoglobin in the blood) and of the clotting factors (*e.g.,* hemophilia). These diseases and others that affect the blood of the newborn are discussed below, in the sections dealing with disorders associated with later infancy and childhood, and are covered in BLOOD: *Blood diseases.*

Erythroblastosis fetalis is a disease in which the red blood cells of the fetus are destroyed because of an incompatibility between the infant's blood and that of the mother. The severest form results from incompatibility between an Rh-negative mother and an Rh-positive fetus. If the mother has been sensitized (by previous exposure) to Rh-positive red blood cells, she will have circulating antibodies against the Rh factor. These antibodies can cross the placenta and destroy the red blood cells of her Rh-positive fetus. Unless the mother has been sensitized by blood transfusions, her first Rh-positive fetus is normally not affected. This is because her exposure to the fetal red blood cells is minimal until the delivery of the baby, when there is substantial transfer of fetal red blood cells to the maternal circulation. This exposure can sensitize the mother, and any future Rh-positive fetuses will be at risk. It is now standard procedure to administer anti-Rh serum promptly to an Rh-negative mother who has given birth to an Rh-positive child. The serum destroys any fetal red blood cells in her circulation before she becomes sensitized, thereby protecting future Rh-positive fetuses from erythroblastosis fetalis.

**Gastrointestinal disorders.** Vomiting, a common symptom among newborn infants, may be due to intestinal obstruction or to overfeeding or may occur without apparent cause. Continuous contraction of the muscle governing the opening between the stomach and the intestine may cause vomiting. This condition, called pyloric stenosis, may occur at any time in early infancy and usually requires surgical treatment.

The first bowel action and passage of meconium by the baby usually occurs within 12 hours. Delay may indicate an obstruction of the bowel. Important causes of obstruction are congenital narrowing (stenosis) or occlusion (atresia) of the intestine. These can occur at any site—from the duodenum (the first section of the small intestine) to the rectum and anus. Some babies are born with a small dimple or pit rather than a patent anus. Duodenal stenosis is particularly common in babies with Down's syndrome. Congenital obstructions of the intestines cause vomiting and constipation in early life; most can be corrected surgically.

Congenital bowel obstruction

Meconium ileus, intestinal obstruction by hard lumps of meconium, occurs almost exclusively in infants with cystic fibrosis, an inherited disease that is described below.

Recovery, except in some instances of perforation of the intestine, is the rule.

**Kidney and urinary-tract disorders.** The kidneys of the newborn infant are entirely capable of maintaining homeostasis, or balance, of fluids and electrolytes in normal circumstances, adapting readily, for example, to the various formulas utilized in infant feeding, despite the wide range of solute content and the consequent large variation in the excretory load imposed. (Electrolytes, in this context, are substances that become ionized in solution; that is, are given a positive or negative electrical charge.) In situations of stress, however, abnormalities in the regulation of salt and water balance and of acid–base metabolism are common. Limitations in the excretory capacity of the newborn infant's kidneys require adjustment of drug dosage and fluid therapy.

The most common disorders of the kidneys and urinary tract encountered in the neonatal period are congenital anomalies. Some, such as absence of one kidney, do not matter, since one healthy kidney will suffice; but other infants are born with no kidneys or with malformed (dysplastic) kidneys that function poorly. Polycystic disease of the kidneys is an example of a serious congenital abnormality. In this disorder, the kidneys contain numerous large cysts that severely impair renal function.

Congenital obstructions of the urinary tract—either of the ureter above the bladder or of the urethra below it—predispose the infant to urinary infection and to kidney damage. Failure of a newborn to pass urine within 12 hours of birth leads to a search for possible obstruction.

Nephritis (inflammation of the kidneys) is rare in the newborn. In one well-known type—congenital nephrosis—large amounts of protein are lost in the urine, with consequent development of severe, generalized edema. The outlook for recovery in congenital nephrosis and in other forms of nephritis in infants is extremely poor.

Infections of the kidneys and urinary tracts are difficult to recognize clinically in young infants. If, however, they are diagnosed early and treated promptly, such infections respond well to treatment, unless there is an associated congenital obstructive lesion.

**Nervous-system disorders.** Congenital malformations of the nervous system rank among the most common severe congenital abnormalities. A variety of brain malformations may occur, some incompatible with life (*e.g.*, anencephaly—the absence of the cerebral hemispheres), others resulting in permanent disability. Common brain malformations include microcephaly, an abnormally small head due to limited brain growth, and hydrocephalus, in which there is an increase in the volume of cerebrospinal fluid associated with increased pressure. The obvious evidence of the latter condition is the large size of the head. Some infants with hydrocephalus die before birth. After birth, the condition may arrest spontaneously. The major treatment is relief of pressure by diversion of the spinal fluid or by surgical correction of any obstruction. The prevention of progressive damage is the goal of therapy.

Spina bifida

Spina bifida is a congenital disorder in which the vertebral column fails to close over a portion of the spinal cord, usually in the lumbar region, leaving that section of cord unprotected. Part of the unprotected cord—nervous tissue, meninges (the cord's membranous covering), or both—may protrude through the defect in the vertebral column. Protrusion of the meninges, with or without neural elements, is frequently accompanied by hydrocephalus. The spinal-cord abnormality usually results in defective nerve function below the level of the lesion; thus weakness or paralysis of the legs and urinary incontinence are common.

Acquired conditions, including those secondary to insufficient oxygen and bleeding, have been mentioned above in the section on birth injuries. Meningitis (inflammation of the coverings of the brain and spinal cord) may occur in the newborn. Unfortunately, the diagnosis is often delayed because of the lack of characteristic symptoms and findings in infants.

Convulsions (seizures) are common in the newborn. These may result from damage to the brain during delivery or from infections and metabolic problems.

**Endocrine disorders.** Although rare, congenital defects of the endocrine (hormone-producing) glands can have severe consequences. Congenital hypothyroidism (subnormal secretion by the thyroid glands, also called cretinism) is an especially important endocrine disease of infancy in that failure to identify and treat it early may result in severe mental retardation. It is due either to an absence of the thyroid or to a metabolic disturbance in the function of the gland. Early diagnosis and proper therapy with thyroid drugs result in dramatic improvement, with rapid disappearance of all signs and symptoms of disease. In countries with well-developed health services, a drop of the newborn's blood is used in a chemical screening test for cretinism. Effective treatment thus can be initiated shortly after birth.

Congenital adrenal hyperplasia is a group of conditions in which there is a defect in the production of normal adrenocortical-steroid hormones (secretions of the cortex, or outer substance, of the adrenal glands). Excessive stimulation of the cortex of the adrenals by a pituitary hormone (adrenocorticotropic hormone, or ACTH) results in abnormal enlargement of the glands and overproduction of androgenic (masculinizing) adrenal hormones. As a result, there may be abnormal development of the genitalia of females in utero and evidence of excessive androgenic effect in either sex during infancy, with accelerated growth, premature appearance of pubic hair, and enlargement of the phallus.

**Musculoskeletal disorders.** Common congenital musculoskeletal defects include abnormalities of the feet and the hips. Classic clubfoot, or talipes equinovarus, is a congenital twisting of the foot in which the heel bends upward and the front part of the foot is turned inward and bent toward the heel. Correction usually involves the use of splints and plaster casts to force the foot into the correct position; severe cases may necessitate surgery. In talipes calcaneovalgus, the front part of the foot is bent upward and turned outward. This form of clubfoot generally results from mechanical pressure in the uterus having held the foot in an unusual posture. Passive stretching exercises usually can correct this condition, but stubborn cases may require the use of splints or casts.

Congenitally dislocated hips are associated with lax joints and are most common in girls born at term by breech delivery. The condition is usually detectable by careful clinical examination and, if diagnosed early, responds to simple treatment. If undetected until a two-year-old is noticed to walk with a limp, major surgery may be needed.

**Skin disorders.** The infant's skin has a thin epidermis and immature glands and is particularly susceptible to blistering and infection. Diaper, or napkin, rashes, which affect the areas of skin in contact with a wet diaper, are very common and can become severe when additional infection occurs.

There are many common birthmarks. Most result from either developmental anomalies of the blood vessels, called hemangiomas, or from an excess of pigment in the skin, called nevi. A common worry to parents is the "strawberry" hemangioma, which is red, raised, and unsightly. Although it may increase in size in the early weeks, it gradually fades away by the age of seven years. A Mongolian blue spot, usually on the buttocks or back, looks like a faint bruise and is a common pigmented birthmark in infants of black or Oriental stock. It fades and is rarely visible after the age of seven.

**Chromosomal disorders.** A normal person has 46 chromosomes, but sometimes developmental faults occur that result in the fetus' having extra chromosomes. Most of these abnormal fetuses result in miscarriages or stillbirth, but those with Down's syndrome (mongolism) commonly do survive. Down's syndrome occurs approximately once in every 600 births. The affected child carries an extra chromosome number 21 and has a characteristic appearance that includes a round skull; flat face; oblique eyes; small, drooping mouth; and a short, broad neck and hands. The main problem of Down's syndrome victims is moderate to severe mental retardation. As adults, most are incapable of leading independent lives. They also suffer from an excess of respiratory infections in early life and

have an increased incidence of serious congenital abnormalities. In developed countries, however, most of them grow up to be reasonably healthy adults, though their life expectancy is shorter than that of a normal person.

Disorders of the sex chromosomes are also common. These disturb the development of the gonads more than they influence the external genitalia; therefore, many of the conditions are not diagnosed until after puberty, when the child (or parents) becomes concerned about the lack of development of sexual characteristics. Normal girls have two X sex chromosomes. Those with Turner's syndrome have a single X chromosome. The syndrome may be detected early in life because the girls are short and have other visible characteristic features. The diagnosis can only be confirmed, however, by careful analysis of the chromosomes in the blood cells. These girls remain short, and secondary sexual characteristics do not appear unless additional hormones are given. Even then, an affected girl remains infertile because her uterus, vagina, and gonads are very small.

Turner's syndrome

Normal boys have one X chromosome and one Y sex chromosome. Those with Klinefelter's syndrome have an extra X chromosome. Although the condition occurs as often as Down's syndrome, it is not usually detected until the testicles fail to enlarge at puberty. The boys are healthy, but infertility is usual.

It is well documented that more males than females are affected by nonspecific mental handicaps and that in some families the males are regularly affected. Several different forms of mental handicap linked with the X chromosome in the male have been identified. In some of the affected males, a "fragile site" can be identified on the X chromosome with appropriate laboratory techniques. Such males are said to have the fragile-X syndrome.

### DISORDERS OF LATER INFANCY AND CHILDHOOD

**Sudden infant death syndrome (SIDS).** In developed countries, SIDS (also called crib death or cot death) accounts for 20 percent of deaths between the ages of one month and one year. SIDS is a categorization rather than an explanation, for the label is given when no reason for death can be found from the infant's medical history or even after autopsy.

Most crib deaths occur in the first five months of life and strike at home during the night. They are more common in the winter and in poor social circumstances. A preceding minor respiratory infection is common. This has prompted some investigators to suggest that the underlying defect is the presence of a virus in the bloodstream, leading to instability of cardiac and respiratory mechanisms. Many other hypotheses have been proposed to explain such deaths, however, and it is likely that several different causes may be involved.

**Failure to thrive.** Failure to thrive is the term used to describe the condition in which a young child fails to gain weight satisfactorily. Common reasons for such poor weight gain are parental neglect or lack of food. On the other hand, a large number of important gastrointestinal disorders may be responsible, including those associated with vomiting, such as food intolerance or obstruction of the upper bowel by pyloric stenosis; disorders of digestion and absorption, including celiac disease and cystic fibrosis; and bowel infections. Alternatively, the body, because of other serious disorders (*e.g.,* chronic infection or heart or kidney disorder), may fail to use the food that is given and absorbed appropriately.

**Malnutrition.** Malnutrition refers to any disorder brought on by improper diet. In developed countries, the most common form of malnutrition is obesity, the excess accumulation of fat brought on by a diet containing too many calories. Obesity is a major contributor to ill health throughout life. In nonindustrialized nations, by contrast, most malnutrition stems from the lack of food or of particular nutrients. Such deficiency diseases remain an enormous problem. In addition, specific nutritional disturbances are encountered regularly in all populations.

Malnutrition due to inadequate intake of food results in muscle wasting, stunted growth, pallor, increased susceptibility to infection, and fatigue. A special form of malnutrition, in which the intake of calories is adequate but that of protein is not, is referred to as kwashiorkor; it is prevalent in areas of Africa, Asia, and Latin America. Kwashiorkor primarily affects children from six months to five years of age, the onset usually coinciding with the child's being weaned from breast milk (which provides adequate protein) to a diet consisting largely of starchy carbohydrates. The affected children are small, have excess fluid in their tissues, and often have enlarged livers. They have unusual pigmentation of the skin and sparse, reddish hair. Permanent aftereffects of kwashiorkor, especially on the intellectual functions, are matters of great concern.

Kwashiorkor

Vitamin deficiencies can result in a variety of diseases. Rickets is a disorder secondary to deficiency of vitamin D. The major consequence is bone disease, with defective growth of the epiphyseal cartilage. (This cartilage, present in several bones, especially near the ends of the long bones of the arms and legs, ossifies as a person matures.) Scurvy occurs as a consequence of a deficiency of vitamin C. Clinical manifestations include bone disease, irritability, and bleeding under the skin and mucous membranes. Pellagra is due to a deficiency of niacin and is manifested clinically by diarrhea, dermatitis, and dementia. Riboflavin deficiency results in lesions of the skin and corners of the mouth, with a peculiar smoothing of the tongue. Beriberi is a consequence of thiamine deficiency. The major clinical features often relate to cardiac impairment. Defects in the functioning of the nervous system also are common. Deficiency of vitamin A results in ocular abnormalities, growth retardation, anemia, and dermatitis.

**Classic infectious diseases of childhood.** All of the various types of infectious disease, which can involve virtually every organ and every part of the body, are encountered in children. There are, however, certain infectious illnesses that have become almost synonymous with the term childhood disease. These diseases occur chiefly among children, and one bout usually provides lifelong immunity against further attacks. Such classic infectious diseases of childhood include the exanthematous viral infections (*i.e.,* measles, chicken pox, German measles, and other viral infections that produce skin eruptions) and mumps. The incidence of these diseases, which were once endemic among childhood populations throughout much of the world, now varies markedly. Smallpox, the most serious of the exanthematous viral diseases, has been eradicated worldwide through immunization programs. Other classic childhood disorders—including measles, German measles, and mumps—have been all but eradicated via immunization in countries with high standards of medical care. They remain endemic, however, in areas with poorer health-care systems.

Measles (rubeola) is a viral disease transmitted by the respiratory route, with an incubation period of 10 to 14 days. The initial symptoms include a runny nose, conjunctivitis (inflammation of the membrane lining the eyelids and covering part of the front of the eye), cough, and a characteristic eruption on the mucous membranes of the mouth (Koplik spots). The characteristic rash then appears on the skin, usually beginning over the neck and the face, and spreads to the rest of the body. Recovery is the rule, although serious neurologic complications and secondary bacterial infections of the lungs may occur. Measles ranks as an important cause of death among undernourished children in poor countries, but it rarely causes death or permanent disability in developed countries.

Measles

German measles, or rubella, is a milder disease, also viral, with an incubation period of 14 to 21 days. As discussed earlier, its major significance is the likelihood of its causing severe malformations of the fetus if contracted by the mother during the first three months of pregnancy.

Chicken pox (varicella) is a highly contagious viral disease with an incubation period of 13 to 17 days. At the start there is mild to moderate fever, followed by a generalized eruption of papules, small, solid elevations that appear in crops, initially small and red, becoming vesicular (*i.e.,* becoming small blisters). After several days, no new lesions develop, and the vesicles gradually crust over and heal. Severe itching usually accompanies the rash.

Mumps is a viral disease of the parotid and other salivary

glands, which has an incubation period of 14 to 24 days. The predominant feature of the disease is painful swelling of the parotid glands, which are below and in front of the ears. The pancreas and gonads (sex glands) may also be involved, although rarely in children.

**Respiratory disorders.** The common cold, or acute nasopharyngitis, the most common respiratory disease in children, is caused by a large number of viruses and may be complicated by superimposed bacterial infection. There is no specific treatment.

Tonsillitis (acute infection of the tonsils) is more properly considered a part of the acute-pharyngitis (throat-inflammation) syndrome. Enlargement of the tonsils as a result of recurrent infection often leads to the decision to remove the tonsils, a course many physicians now believe is rarely indicated. Enlarged tonsils do not cause irritability, poor appetite, or poor growth.

Enlargement of the adenoids (lymphoid tissue in the nasal part of the pharynx) as a result of recurrent infection can result in mouth breathing and a so-called adenoidal facial appearance, the most conspicuous feature of which is the constantly open mouth. By blocking the eustachian tube, it can contribute to infections of the middle ear (otitis media) and to hearing loss. In children with chronic middle-ear disease and a specific type of hearing loss, removal of adenoids may be indicated.

Croup

Croup is an inflammatory disease of the larynx (voice box) or epiglottis (the plate of cartilage that shuts off the entrance into the larynx during the process of swallowing), most often caused by viral infection; it is encountered in infants and small children. Inflammation and swelling of the vocal cords lead to respiratory obstruction, particularly in the inspiratory phase, and a croupy cough, which sounds like the bark of a seal.

Allergic rhinitis (inflammation of the nasal passages) is the most common allergic disorder of childhood. Seasonal allergic rhinitis, or hay fever, due to sensitization to house dust, pollen, or molds, is characterized by attacks of sneezing, nasal itching, and a watery nasal discharge during the season when the specific allergens are prevalent. Similar symptoms are present in perennial allergic rhinitis but without seasonal pattern. In addition to inhalants, sensitization to specific foods may underlie the disorder. Treatment consists of avoidance of the substances causing the reaction, desensitization, and use of decongestant drugs and antihistamines (drugs that, by inactivating the histamine given off by injured cells, suppress many of the symptoms of an allergic attack).

Asthma

Asthma is a common allergic disorder of children that affects the bronchi and bronchioles (the large and small air passages in the lungs). Spasm, edema, and abnormal secretion of mucus result in obstruction of the lower respiratory tract and characteristic wheezing and laboured breathing. Inhalant allergens, particularly dust, molds, and pollens, and foods may play important causal roles. Psychologic stress may be a precipitating factor, but viral or bacterial infection of the respiratory tract is a more common triggering factor. A variety of effective treatments is available, together with preventive measures that reduce the chances of recurrent attacks. The outlook generally is good, with only a small percentage of children continuing to have severe asthma into adult life.

In discussing childhood respiratory diseases, tuberculosis and cystic fibrosis should be included. Both of these disorders predominantly affect the lungs, although many other organs may also be involved. Tuberculosis continues to be a major world health problem. As countries improve public-health standards and increase their socioeconomic level, the illness and mortality from this disease decrease steadily. Tuberculosis appears mostly in a primary form consisting of a small localized lesion of the lung that either heals completely or remains quiescent for many years. Only infrequently among children does the disease extend to involve other parts of the lung or other parts of the body, such as bones, kidneys, or the central nervous system. Miliary tuberculosis, a generalized form of infection, and tuberculous meningitis are the most severe forms of the disease and have an extremely high mortality, although recovery may occur with proper treatment. These forms most commonly occur in young children. As with other diseases, tuberculosis is better prevented than treated. A form of immunization (BCG—bacille Calmette-Guérin—vaccine) is utilized in areas of the world in which the disease is endemic. In other areas, control depends on prevention of contacts and early identification and treatment, if necessary, of infected individuals. A variety of antibiotic agents is effective in treatment, particularly the drugs isoniazid and rifampin.

Cystic fibrosis is a hereditary disorder of the exocrine glands (*i.e.,* those glands that release secretions through ducts). It affects many organ systems, but the lungs suffer most severely. Estimates of incidence vary from one in 3,700 to one in 1,000 live births. It is rare among blacks and Orientals and is transmitted as a recessive trait. The underlying metabolic defect is unknown, but the disease appears to start with the secretion of unusually thick and sticky mucus. In fetuses, intestinal obstruction may result from the production of viscid meconium. Pulmonary involvement may be apparent in the newborn or may develop during childhood, with repeated bouts of atelectasis (collapse of the lungs) and ultimate bronchiectasis (chronic dilation and degeneration of bronchi and bronchioles). Pancreatic insufficiency leads to a malabsorption syndrome, with fatty, bulky stools and malnutrition. The liver may be involved. Abnormality of the sweat glands is evidenced by a high salt content of the sweat, which, in hot weather, may lead to salt depletion and collapse. Treatment is directed toward the many organs involved, particularly with regard to aggressive therapy for respiratory tract infections. Regulation of diet and administration of pancreatic enzymes contribute to the maintenance of adequate nutrition. The ultimate outlook is grave, although therapy has been successful in markedly prolonging life. Many affected persons survive into adult life.

Sinusitis, otitis, bronchitis (inflammation of the sinuses, the ears, and the bronchi, respectively), and pneumonia occur commonly in children and do not differ in essential detail from the same diseases in adults. Other conditions that affect children and adults alike are described in RESPIRATION AND RESPIRATORY SYSTEMS: *Respiratory system diseases.*

**Cardiovascular disorders.** Congenital heart defects, treated earlier in this article, rank among the most common sources of cardiovascular difficulties in children. Among acquired heart diseases in children, rheumatic fever is the most important cause worldwide, although it has become far less common and less severe in developed countries. Rheumatic fever strikes mainly between the ages of five and 15, occurring as an abnormal reaction to a beta-hemolytic streptococcal throat infection of a few weeks previous. Heart involvement may not be apparent early, but 60 percent of the victims develop rheumatic heart disease in later life; mitral stenosis (narrowing of the mitral valve) is a particularly common complication.

Most disorders of cardiac rate and rhythm in childhood are benign. An exception is paroxysmal atrial tachycardia, a disorder characterized by a steady, rapid heart rate, which in infants may exceed 300 beats per minute. If the disorder persists, it may lead to heart failure. Treatment with digitalis usually restores normal rhythm.

Pericarditis and myocarditis, inflammation of the sac enclosing the heart and of the heart muscle, are caused by a variety of infectious agents; they may result from systemic diseases. The most common cause is acute rheumatic fever. Symptoms include pain, fever, and evidence of heart failure. Treatment and prospects of recovery depend on the underlying cause.

Bacterial endocarditis (bacterial infection of the heart lining) occurs most frequently in children with preexisting heart disease. The most common organism is the alpha streptococcus, which accounts for 80 percent of cases. Common symptoms are fever, a sense of ill health, and fatigue. The outlook depends on the sensitivity of the infecting organism to antibiotic drugs, the age of the affected child, and the type of underlying heart disease.

**Blood disorders.** Virtually all of the recognized blood diseases of adults are encountered in children. Of particular importance are the conditions in which abnormal types

of hemoglobin are formed. The abnormal hemoglobin present in sickle-cell anemia, also called sickle-cell disease and sicklemia, must be inherited from both parents to cause the disease, the effects of which include hemolytic anemia (anemia involving destruction of red blood cells and release of their hemoglobin) and recurrent crises with episodes of painful swelling of the hands and feet, abdominal pain, and increase of the anemia. Persons who have inherited the defect from one parent and are said to have the sickle-cell trait constitute approximately 10 percent of the U.S. black population. There are a number of other abnormal hemoglobins. Thalassemia, or Cooley's anemia, is a condition in which there is severe, progressive hemolytic anemia, beginning at about six months of age. Like sickle-cell anemia, thalassemia is a recessive hereditary disorder and thus must be inherited from both parents. It occurs in a broad equatorial belt extending from the Mediterranean countries through India to the Far East. Its underlying defect is the deficient production of adult hemoglobin (hemoglobin A). Repeated transfusion of blood and, in certain instances, removal of the spleen are the only available treatments.

Hereditary spherocytosis and hereditary elliptocytosis cause hemolytic anemia because of abnormalities in the structure of the red blood cell. A number of abnormalities in red-blood-cell enzymes also can lead to increased red-cell destruction.

The most common form of anemia in infants and children is caused by iron deficiency. Fetal stores of iron usually prevent development of anemia during the first six months of life, but it is common thereafter, when the diet may not be adequate to meet the high requirements for iron. Apart from pallor, children usually are well, although they may show irritability and lack of appetite. Treatment consists of the administration of iron and modification of the diet to include sufficient iron to prevent recurrence.

Leukemia

Leukemia is a neoplastic (cancerous) disorder of the leukocyte precursors (*i.e.,* young forms) of the white blood cells in the blood-forming tissues. The incidence in childhood is about four cases per 100,000 population. It is the most common malignant disease of children, with a peak onset between two and four years of age. Most cases are of the lymphoblastic type. (Lymphoblasts are precursors of lymphocytes.) Clinical manifestations include anemia, thrombocytopenia (deficit of blood platelets), and infiltration of various organs of the body with leukemic cells. A number of drugs are available for the treatment of leukemia. Remission (disappearance of symptoms) can be induced in about 90 percent of children with acute lymphoblastic leukemia, and half of these survive more than five years.

Thrombocytopenia is a disorder characterized by a tendency toward bleeding because of a decrease in circulating platelets. (The platelets help to stop bleeding in two ways: they contain a clotting factor, and they serve to block rents in blood-vessel walls.) The causes of most cases remain unknown. Treatment consists of replacement of blood when there is a major hemorrhage, transfusion of platelets for emergency management, and, in selected cases, administration of adrenocortical steroids and removal of the spleen. Spontaneous recovery occurs in 80 to 90 percent of cases within three months from onset.

Congenital disorders of the coagulation process usually become manifest during infancy or early childhood. The most common of these is hemophilia, a disease caused by deficiency in a specific coagulation factor. The disease is manifested only in males who have inherited the trait from their mother and occurs in about one of every 10,000 male births. Treatment consists of intravenous injection of the deficient factor, along with measures to control bleeding locally and transfusion of blood when necessary.

**Gastrointestinal and liver disorders.** Abdominal pain, one of the most common symptoms of childhood, can be indicative of many gastrointestinal disorders but usually occurs without evidence of disease. Recurrent abdominal pain without detectable disease may be a psychosomatic disorder. See also DIGESTION AND DIGESTIVE SYSTEMS: *Digestive system diseases.*

Acute appendicitis occurs in all age groups, although it is rare in extremely young infants. The clinical manifestations (abdominal pain, vomiting, fever) in older children are similar to those in adults. In infants and younger children, the systemic manifestations are more severe, and rupture of the appendix is more frequent.

Intussusception

Intussusception is a condition encountered in the first and second years of life in which one section of intestine doubles (invaginates) into the section next distant from the stomach. Gastrointestinal bleeding and symptoms of obstruction result. Sometimes the intussusception is eliminated by administration of a barium enema. Surgical correction is more usually required, however.

Young children often put things in their mouths, and sometimes they accidentally swallow them. Foreign bodies lodged in the esophagus must be removed. Objects small enough to pass through the esophagus into the stomach usually will pass through the entire intestinal tract, and no treatment is necessary.

Food intolerance is an important cause of vomiting, diarrhea, and failure to thrive in early life. Sometimes the intolerance to a specific food item is temporary, so all that is required is avoidance of that substance for a few months. Other intolerances are more serious and may require lifelong avoidance. Celiac disease is caused by a peculiar sensitivity to the gluten fraction of wheat, rye, or other cereals; therefore, symptoms develop after the introduction of cereals into the diet. Affected children characteristically pass large, bulky, greasy stools; have poor appetite; and are generally miserable. Although they are thin, there may be marked gaseous abdominal distension. The diagnosis is usually made by a biopsy of the affected segment of the intestine (the jejunum). Treatment, consisting of exclusion of gluten from the diet, has a dramatic effect within a few weeks. The patient needs to remain on the special diet throughout childhood and probably should continue it throughout life.

Intolerance to particular sugars may be associated with gluten sensitivity or may occur on its own. The disaccharide sugars—for example, sucrose (table sugar) or lactose (milk sugar)—are the most common offenders. Inability of the intestine to handle these sugars in the normal way leads to diarrhea, malabsorption, and failure to thrive. Fortunately, sugar intolerance is generally a transient event.

Types of viral hepatitis

Viral hepatitis (inflammation of the liver due to infection with a virus) has its highest incidence but lowest mortality rates among children of school age. Two main forms of the disease, hepatitis A and hepatitis B, occur in children. These two forms were distinguished initially by their clinical characteristics and are now recognized to be caused by two different viruses. Hepatitis A (infectious hepatitis) is highly contagious and can be passed from person to person directly, as well as being acquired from contaminated water or food. It is particularly common in communities with poor sanitation. Hepatitis A has an incubation period of 14 to 40 days and usually has an abrupt onset. Fever, headache, and feelings of ill health are followed by loss of appetite, nausea, and vomiting. Jaundice ensues, and the liver becomes enlarged and tender. Improvement is usually noted in a few weeks. In children, complete recovery is common.

Hepatitis B (serum hepatitis) is acquired as a result of receiving blood or blood products from someone who is a carrier of the disease. Thus, it may be acquired from a blood transfusion or from being inoculated with a contaminated needle used by another person (*e.g.,* among drug addicts using the same needle). It has a much longer incubation period (60 to 160 days) and more gradual onset than does hepatitis A. The clinical symptoms are much the same as in hepatitis A, and the outlook for recovery is good in children.

**Kidney and urinary-tract disorders.** Infection of the urinary tract is common and occurs predominantly in females. The most frequent infection is cystitis, a superficial infection of the lining of the bladder, but pyelonephritis, infection of the kidney, is not uncommon. *Escherichia coli* is the organism responsible in 80 percent of the cases. Symptoms of cystitis include urgency, frequency, painful urination, and suprapubic pain (pain just above the frontal pelvic bones). Pyelonephritis may be without

Cystitis and pyelonephritis

symptoms or may cause fever, back pain, and shaking chills. The patient, who should drink plenty of fluids and void frequently, usually receives antibiotics, which clear the infection rapidly. Ultrasound or X-ray investigation for underlying congenital abnormalities is especially important in young children. Infection recurs in up to 50 percent of cases.

The presence of bacteria in the urine without manifestation of symptoms (asymptomatic bacteriuria) is found in about 1 percent of schoolgirls. This condition is associated with an increased frequency of minor voiding disturbances, of urinary-tract abnormalities, and of symptomatic urinary infections in later life.

Various forms of glomerulonephritis (kidney disease in which there is inflammation of the glomeruli—the knots of minute blood vessels in the capsules of the nephrons, the functioning units of the kidneys) affect children. The type most commonly encountered in children worldwide—though infrequently seen in developed countries—is acute post-streptococcal nephritis. This disorder occurs as a late complication of infection with certain strains of group A beta streptococci. The onset is heralded with blood in the urine, excess fluid in the tissues, or headache due to high blood pressure. Spontaneous recovery ordinarily occurs. A rare patient with unusually severe disease may suffer irreversible kidney damage.

The nephrotic syndrome is a group of symptoms that occurs as a consequence of any kidney disease; characteristically, there is excretion of great amounts of protein in the urine, and generalized edema occurs in the absence of evidence of glomerulonephritis or systemic disease. Most of these children respond to treatment with adrenocortical steroids and ultimately recover. As previously mentioned, congenital nephrosis is an especially severe form that may be apparent at the time of birth. There is no effective treatment, and infants do not usually survive beyond the first year of life.

All forms of glomerulonephritis described in adults are seen also in children. If sufficient information is available, most instances of hereditary nephritis can be shown to have their onset in childhood.

Disorders of specific tubular functions (*i.e.,* functioning of the nephrons) are rare. Nephrogenic diabetes insipidus is a disease of male infants in which there is failure of the kidneys to respond to antidiuretic hormone, with consequent inability to concentrate urine. The symptoms are polyuria (copious urine), polydipsia (excessive thirst), and chronic dehydration. The Fanconi syndrome is a group of diseases in which there are multiple abnormalities in renal-tubular function. In one of these, cystinosis, there is progressive impairment in renal function.

Dialysis and kidney transplant

Children with renal failure and uremia (nitrogenous wastes in the blood) can be treated with dialysis and renal transplant. The major role of dialysis in children is to support patients until a transplant can be performed. Transplant of a kidney from a living, related donor yields the best results, but many children have had successful transplants of kidneys from cadavers (see TRANSPLANTS, ORGAN AND TISSUE).

**Nervous-system disorders.** Congenital anomalies of the nervous system are common and have been discussed earlier in this article.

Mental retardation is a major problem, affecting about 0.5 percent of young children. For the majority the cause is unknown. Fortunately, most retarded children have only a mild handicap, with an intelligence quotient (IQ) between 50 and 75, and are, therefore, educable and trainable to a reasonable degree of independence as adults. Up to 15 percent of affected children have a more severe handicap, however, and cannot be expected to be independent as adults.

Cerebral palsy refers to a condition in which there is a nonprogressive lesion of the brain causing impairment of movement and posture. Unfortunately, its causes are sometimes responsible for brain or nerve damage in other areas and for mental handicap; these additional problems tend to influence most strongly the quality of the child's life. For children with cerebral palsy, the brain damage is permanent, but it does not increase. Much rarer are the degenerative diseases of the nervous system, most of which are of unknown cause and are untreatable.

Brain tumours are the second most common malignancy of childhood (after leukemia); they are, nevertheless, very rare. The most common brain tumours are situated at the base of the brain and are associated with raised intracranial pressure, causing head enlargement or pain and vomiting. Scanning by computerized axial tomography, which provides a cross-section image of the brain, helps with the diagnosis. A combination of radiotherapy and surgery may be successful in treating such tumours.

Convulsive disorders

Convulsive disorders in children are common. As many as 5 percent of children have a seizure at least once during their lifetime. So-called febrile seizures occur in association with high fever. They are most common between six months and four years of age, and there is a high familial incidence. Spontaneous recovery is usual. Epilepsy, or recurrent seizures, has a prevalence of about 0.5 percent. There are many known causes, but in most cases none is found. Treatment with anticonvulsant drugs is successful in suppressing seizures in most cases.

Eighty percent of all cases of meningitis occur in the first five years of life, and the majority of these strike during the first two years. In most cases, meningitis results from a bacterial or viral infection of the cerebrospinal fluid. Bacterial meningitis is a serious acute illness, in that the bacterial infection may damage the brain permanently. By contrast, most of the common viral causes—including the mumps virus—rarely produce serious illness. Poliomyelitis, however, is a serious form of viral meningitis because of the risk of permanent paralysis. Fortunately, poliomyelitis is rarely seen in countries that provide mass immunization against the disease. The diagnosis of meningitis is made by analyzing a sample of cerebrospinal fluid withdrawn by needle from the spinal canal (*i.e.,* by lumbar puncture). Effective antibiotic therapy is available for the bacterial forms.

**Endocrine disorders.** In addition to the congenital disorders discussed earlier, a variety of endocrine diseases can occur during childhood. These include precocious puberty, hyperthyroidism, pituitary or adrenal insufficiency, and diabetes mellitus.

Precocious puberty includes a large group of conditions in which there is premature onset of sexual development. Although precocious puberty can result from disease of the brain, adrenals, or gonads, in most instances no underlying disease can be detected.

Overactivity of thyroid function, or hyperthyroidism, is uncommon. Affected children exhibit nervousness, weight loss, irritability, and hyperactivity. Usually there is enlargement of the thyroid gland. A variety of drugs that suppress thyroid function is available. In some instances, surgical removal of most of the thyroid gland is indicated.

General pituitary or adrenal insufficiency results in deficiencies of many hormones and produces a complex disturbance of many body functions, usually requiring urgent treatment. Therapy consists of administration of those hormones that are not being produced in sufficient quantity. A deficiency of the pituitary secretion growth hormone may exist without other deficiencies, in which case it causes merely extreme shortness, the child being otherwise well. Once the condition has been identified by serial measurements of the rate of growth and by measurement of growth hormone in the blood, injections of growth hormone can restore the child to normal height.

Diabetes mellitus in childhood is nearly always of the type 1 variety; *i.e.,* resulting from a deficiency of the pancreatic hormone insulin. Because there is a familial tendency for the condition, and because more children have been treated and have grown up to have their own children, there has been an increased incidence of diabetes worldwide. The most striking clinical features, elevated levels of glucose in the blood and increased excretion of glucose in the urine, are due to the patient's inability to metabolize glucose normally. Abnormalities in fat and protein metabolism are also present. Control of the abnormal handling of glucose by the daily administration of insulin and some restrictions of diet can keep most children asymptomatic and enable them to lead normal,

healthy lives. Even the best control, however, might not prevent vascular and neurologic complications that occur 20 or more years later. The outcome, therefore, is unsure, and the majority of persons with onset of diabetes mellitus in childhood appear to develop significant complications in adult life and to have a reduced life expectancy.

**Skin disorders.** Skin disorders are common during childhood. Birthmarks and diaper rashes (both considered above), eczema, and local infections are often seen.

Eczema is characterized by reddening of the skin, papules, oozing, and crusting with intense itching. In infants the lesions often appear first on the cheeks and then develop on other areas, while older children are most affected on the insides of the elbows and the knees. Treatment includes attention to any underlying allergic causes and local application of a variety of different medications, especially adrenocortical-steroid ointments.

Impetigo contagiosa is a superficial infection of the skin with *Staphylococcus aureus* or hemolytic streptococci. Vesicular or pustular lesions exude moisture and become crusted. Untreated, the lesions tend to become widespread and may involve any area of the skin or the scalp. Treatment consists of keeping the affected areas clean and local or systemic administration of antibiotics.

Fungal infections of the skin are also common. Thrush, a disease characterized by small, white spots in the mouth or a diffuse rash on the body, affects infants infected by the fungus *Candida albicans.* In the older child, tinea capitis (ringworm of the scalp), tinea corporis (ringworm of the body), and tinea pedis (athlete's foot) are all common superficial fungal infections.

Warts, also called verrucae, are the most common viral skin infection and are probably more common in childhood than at any other time. The average life of a wart is three to four months, so treatment is usually reserved for long-lasting warts. On the sole of the foot a verruca that becomes rather flattened is called a plantar wart.

Various parasites may cause skin infestations. The common head louse (*Pediculus humanus capitis*) causes irritation of the scalp and lays tiny, whitish eggs (nits) on the hair. Head lice are easily eradicated by the application of special lotions to the scalp of the child (and the rest of the family). Scabies is an infection caused by the itch mite (*Sarcoptes scabici*), which lays its eggs in burrows beneath the skin. After a few weeks of infestation, the child becomes sensitized to the parasite and develops an itchy rash, particularly on the hands and armpits. The infestation is transferred by bodily contact, so that other family members are commonly infected, and all should be treated with creams or lotions that eradicate the mite.

**Connective-tissue disorders.** Henoch-Schönlein purpura (anaphylactoid purpura) is the most common connective-tissue disorder in children. It is characterized by a purpuric rash, painful swollen joints, and abdominal pain with vomiting. In a minority of patients, the kidneys become involved and nephritis develops; this is the only complication that may cause permanent damage. In general there is complete recovery.

Juvenile rheumatoid arthritis, or Still's disease, is rare. In very young children it is characterized by general illness, fever, and rashes, with comparatively mild joint involvement. In older children, the adult pattern of the illness is seen, with predominant joint involvement and little or no general illness. More than half of affected children make a complete recovery; the rest have recurrences requiring treatment.

**Accidents.** In developed countries, accidents cause more loss of life and disability among children (except infants) than any disease. Road-traffic mishaps account for nearly half of the accidental deaths—usually the child involved being a pedestrian or cyclist. Accidents in the home, by way of burns and falls, account for another quarter. Boys are more at risk than girls, particularly if they are from a large family living in a poor, inner-city area. Children are more likely to suffer serious burns and scalds than adults because of the fact that their skin is thin and more liable to full-thickness damage.

Poisoning

Accidental poisoning is very common, particularly among two- to four-year-olds, who are inquisitive and use their mouths to feel and taste new objects. Accidental ingestion of household fluids and medicines is common. Fortunately, it is usual for the child to swallow only a tiny amount, and severe illness from such poisoning is rare. Medicinal drugs are much more likely to cause illness than are household and garden products, berries, or toadstools.

Lead poisoning has become less common worldwide, though there is increasing worry about prolonged exposure to low levels of lead and its possible relationship to abnormal childhood behaviour and intelligence. Low-level lead poisoning generally results from unavoidable exposure to atmospheric lead pollutants. This is a problem in some heavily industrialized areas and in those regions where leaded gasoline is still burned in automobiles.

**Child abuse and neglect.** The spectrum of child abuse is wide. It includes not only children who have suffered physical abuse with fractures and bruises ("the battered child") but also those who have experienced emotional abuse, sexual abuse, deliberate poisoning, and the infliction of fictitious illness on them by their parents (Munchausen syndrome by proxy). Children under the age of two are most liable to suffer direct physical abuse at the hands of their parents. Such abuse is more common in families who are poor and are living under stress and in which the parents themselves suffered cruelty as children. Frequently, the child shows other evidence of poor nutrition or neglect. Most developed countries have a well-established system for dealing with suspected cases of abuse, involving child-protection agencies, social workers, and, if necessary, the police.

Sexual abuse

Sexual abuse, in which dependent, developmentally immature children are involved in sexual activities that they do not fully comprehend and to which they cannot give informed consent, has become increasingly recognized. Girls are involved mainly, and their fathers are the usual offenders. Sexual abuse frequently does not come to light until the older girl develops a psychosomatic illness, runs away from home, or is truant from school.

**Psychological disorders.** All disorders have both a physical and psychological component. For many disorders—such as asthma, eczema, and ulcerative colitis—the importance of physical and emotional factors varies at different times during the course of the disease. Moreover, the individual's concept of illness and his worries about it inevitably contribute to the severity and duration of a particular illness. A three-year-old may have no concept of disease or of time. Consequently, he may not worry about the cause or duration of a disease but instead be much more upset by immediate discomforts associated with the illness. A young child may view admission to the hospital as particularly frightening and unpleasant. Unless the parents can be with him, he may see their absence as a complete loss and cannot appreciate that he may be back with them and well two days later. Thus, great efforts are made to avoid hospitalizing children. When a youngster is admitted to a hospital, the parents are encouraged to be with him as much as possible and, when conditions permit, to sleep beside the young child in the hospital.

In other ways, the fact that the young child has no concept of illness is an advantage, for as soon as an acute illness is over, the child resumes normal health with startling rapidity—he does not feel in need of convalescence in the way that an adult does after a frightening experience. Children have great and speedy powers of recovery.

Stress precipitates symptoms in people of all ages. Headaches, leg aches, stomachaches, and vomiting are common symptoms of stress in children. The sort of stress that causes such symptoms may be problems at home—such as parental discord, inconsistent parental behaviour, rivalry with siblings, or unrealistic expectations by parents—or problems at school or with friends. The loss of a parent or a move to a new home can be an acute stress.

Minor behaviour disturbance involving enuresis (urinating), soiling (defecating), or sleep disorders are common. Most children who exhibit such behaviours should not be considered psychologically abnormal. Similarly, habit spasms (tics) involving repetitive involuntary movements, usually of the head and neck, are extremely common, affecting up to 10 percent of 10-year-olds.

Autism

Severe behaviour disorders are much less common, and true childhood psychosis is most uncommon. Autism is a well-known condition but extremely rare. The onset typically is in infancy, when the baby seems not to enjoy being cuddled. In some cases, the disorder is only recognized when the child is slow to talk. Autistic children have limited communication and avoid normal social contact, preferring things to people. Minor degrees of autistic behaviour are common in children with other major problems such as blindness, deafness, or mental handicap.

Hyperactivity is a behaviour disorder characterized by perpetual overactivity. Hyperactive children refuse to concentrate on one task for long, are always on the go, and need relatively little sleep. They are very easily distracted, and, because of the lack of concentration, school problems arise. The incidence of hyperactivity varies enormously from country to country, and it is likely that local fashions and beliefs greatly affect the criteria for diagnosis. Most young children are very active and exhaust their parents, and few concentrate on their schoolwork as much as their parents wish. Thus, parents often see a child as overactive and readily suggest hyperactivity as the problem, though strict measurement of psychological criteria rarely demonstrates its presence. Enthusiasts embark on behaviour-modification therapy and sometimes drug therapy.

## DISORDERS ASSOCIATED WITH ADOLESCENCE

Adolescence begins with the onset of sexual maturation and continues through the transition state from childhood to young adulthood. The beginning is biologically defined by the onset of puberty, usually during the 10th to 13th year. The end is less definable and, depending upon environmental factors, may be as early as 16 years or as late as 20. In addition to rapid anatomical and physiological changes occurring during adolescence, the period is one of rapid psychosocial and psychosexual change, with tremendous turmoil generated over feelings of inadequacy, increase in sexual and aggressive drives, internal disorganization, and the attempt to attain self-control.

During adolescence, body weight almost doubles, and an additional 25 percent in height is gained. Secondary sexual characteristics appear, menstruation begins in girls, spermatogenesis (sperm formation) starts in boys, and fertility is established in both sexes. Adolescents establish a sense of identity and achieve a degree of independence that ultimately prepares them to take their place in adult society. It is expectable, therefore, that the major disorders of adolescence concern problems of growth, sexual development, and psychological disturbances.

Disturbances of growth

Disturbances of growth chiefly concern short stature in boys and tall stature in girls, both conditions being a potential source of psychological handicap. Although organic and genetic causes of short stature in boys must all be considered, most relatively short but otherwise healthy children are simply late maturers. Graphic plots of height gain with age reveal steady, normal progression but a delayed pubertal growth spurt, concordant with the delay in sexual maturation. With further sexual maturation, acceleration in growth will occur, and adult height within normal limits will be achieved. Similarly, many excessively tall adolescent girls are early maturers; with early sexual and skeletal maturation, their linear growth stops at an adult height well within normal limits.

Sexual development

The sequence of sexual development in girls is extremely variable. Widening of the bony pelvis, growth of the nipples and breasts, changes in external and internal genitalia, and the menarche occur sequentially as pituitary gonadotropin (sex-gland-stimulating hormone) causes ovarian release of estrogen (female sex hormone). Axillary (armpit) and pubic hair and some of the changes of the external genitalia develop under the stimulus of androgens of adrenal origin. Since these arise from a different source of pituitary stimulation, there is considerable variation among girls in the relationship of their appearance and, for example, development of the breasts. Recognition of this helps to allay anxieties over "abnormal" sexual development.

Menstruation in adolescence is characterized by many functional disturbances, including oligomenorrhea (scant menstruation), amenorrhea (absent menstruation), menorrhagia (excessive bleeding), and dysmenorrhea (painful menstruation). Amenorrhea requires a thorough evaluation for possible organic abnormality, such as underfunctioning sex glands, absence of the uterus, or obstruction to the menstrual flow. In most instances, skipped menstrual periods during the first year or so after the menarche reflect the common irregularity of menstruation during early adolescence. Later in adolescence, transitory amenorrhea may be associated with stress, such as onset of the school year or moving to a new home. Treatment is not usually required.

Sexual development in boys usually follows a more predictable sequence, although there is great variation in the time of onset of puberty and the time of achievement of full sexual maturation. Stimulation of the testes by pituitary gonadotropins results in the release of the hormone testosterone, which causes growth of the internal and external genitalia, development of pubic, axillary, and facial hair, changes in the larynx that result in deepening of the voice, and increased statural growth and muscular development. In about half of all boys, noticeable swelling of mammary-gland tissue occurs midway through adolescence. When the enlargement of the breasts is great enough to engender concern, it is called gynecomastia. In most instances, the enlargement disappears spontaneously.

Acne vulgaris (common acne) is a prevalent skin condition that has its onset during adolescence. At puberty, androgenic stimulation of the skin's sebaceous (oil) glands (which empty into the canals of the hair follicles) causes increased production of the fatty substance sebum. In susceptible individuals, there is oversecretion of sebum. Sebum and cellular debris then form a plug in the follicle canal, and the growth of bacteria in the plug produces unsightly pustules. Prolonged treatment is often needed.

Psychological disturbances

The psychological disturbances of adolescence are universal and protean, ranging from minor emotional upsets to schizophrenia—from mild feelings of inadequacy to suicide. The sexual and aggressive impulses of the preadolescent period are complicated by the advent of physical and sexual maturity. Both an inability to control urges and desires and an excessive degree of self-control are characteristic. Some adolescents remain too dependent upon their parents; others attempt to achieve independence too quickly. As many as 10 percent of adolescents may have psychological disturbances that seriously interfere with their functioning and the development of social relations.

One well-known major psychological disorder that generally begins in adolescence is anorexia nervosa. The onset is usually at puberty. The victims, overwhelmingly girls, at first appear healthy but then refuse to eat, and they lose weight. As they lose weight they begin to look ill, and expert help is required in order to encourage them to eat again and regain health. Anorexia nervosa is rare.

Unfortunately, certain other major behavioral disturbances of adolescence have become increasingly prevalent in the late 20th century. Suicide has become much more common; suicidal gestures are particularly common in girls. In many industrialized countries, suicide ranks as the second or third most common cause of death during adolescence (after accidents and, in some countries, malignancy). Delinquency, vandalism, and dropping out of school have become increasingly important and are often associated with addiction to drugs or alcohol.

**BIBLIOGRAPHY.** BENJAMIN SPOCK and MICHAEL B. ROTHENBERGER, *Baby and Child Care,* new rev. ed. (1985), provides comprehensive coverage of current medical practices and parental concerns. Comprehensive pediatric texts include ABRAHAM M. RUDOLPH, JULIEN I.E. HOFFMAN, and SUSAN AXELROD (eds.), *Pediatrics,* 17th ed. (1982); VICTOR C. VAUGHAN III, *et al.* (eds.), *Nelson's Textbook of Pediatrics,* 12th ed. (1983); HARRY C. SHIRKEY (ed.), *Pediatric Therapy,* 6th ed. (1980); STEPHEN H. SHELDON and HOWARD B. LEVY, *Pediatric Differential Diagnosis,* 2nd ed. (1985); and MORRIS GREEN, *Pediatric Diagnosis: Interpretation of Symptoms and Signs in Different Age Periods,* 3rd ed. (1980). Specific aspects of child care are explored in TOMAS SILBER (ed.), *Ethical Issues in the Treatment of Children and Adolescents* (1983); MORRIS GREEN (ed.), *The Psychological Aspects of the Family: The New Pediatrics* (1985); and JAMES H. HUMPHREY (ed.), *Stress in Childhood* (1984).

(S.R.M./C.M.E./H.L.B.)

# Chile

The Republic of Chile (República de Chile), situated on the Pacific coast of South America, is an extremely long and narrow country that extends about 2,700 miles (4,300 kilometres) from its boundary with Peru in the tropical zone, at latitude 17°30′ S, to the tip of South America at Cape Horn, latitude 56° S, a point only about 400 miles north of Antarctica. This unusually shaped country has an average mainland width of slightly more than 100 miles, with a maximum of 217 miles at the latitude of Antofagasta and a minimum of 9.6 miles near Puerto Natales. It is bounded in the north by Peru and Bolivia, on its long eastern border by Argentina, and to the west by the Pacific Ocean. Continental Chile and its offshore islands comprise 284,520 square miles (736,903 square kilometres), but if the Antarctic territory claimed by Chile were included, it would total 774,763 square miles. Chile exercises sovereignty over Easter Island, the Juan Fernández Archipelago, and the volcanic islets of Sala y Gómez, San Félix, and San Ambrosio, all in the South Pacific. The country also claims a 200-mile offshore limit.

Chile's relief is chiefly mountainous, with the Andes range dominating the landscape. Because of the country's extreme length it has a wide variety of climates, from the coastal desert beginning in the tropical north to the cold subantarctic southern tip. It is also a land of extreme natural events: volcanic eruptions, violent earthquakes, and tsunamis originating along major faults of the ocean floor periodically beset the country. Fierce winter storms and flash floods alternate with severe summer droughts.

Much of northern Chile is desert; the central part of the country is a temperate region where the bulk of the population lives and where the larger cities, including the capital, Santiago, are located; south-central Chile, with a lake and forest region, is temperate, humid, and suitable for grain cultivation; and the southernmost third of the country, cut by deep fjords, is an inhospitable region: cold, wet, windy, and limited in resources. The economy of Chile is based on primary economic activities: agricultural production; copper, iron, and nitrate mining; and the exploitation of sea resources.

Chile exhibits many of the traits that typically characterize Latin-American countries. It was colonized by Spain, and the culture that evolved was largely Spanish; the influence of the original Indian inhabitants is negligible. The people became largely mestizo, a blend of Spanish and Indian bloodlines. The society developed with a small elite controlling most of the land, the wealth, and the political life. Chile, however, did not depend as heavily on agriculture and mining as did many Latin-American countries, but rather developed an economy based on manufacturing as well. Thus, Chile has become one of the more urbanized Latin-American societies, with a burgeoning middle class. Chile has also had a history of retaining representative democratic government. Except for a military junta that held power from September 1973 to March 1990, the country has managed to stay relatively free of the coups and constitutional suspensions common to many of its neighbours.

This article is divided into the following sections:


Physical and human geography 21
- The land 21
  - Relief
  - Drainage
  - Soils
  - Climate
  - Plant and animal life
  - Settlement patterns
- The people 27
- The economy 27
  - Resources
  - Agriculture
  - Industry
  - Trade and finance
  - Transportation
- Administration and social conditions 29
  - Government
  - Education
  - Health and welfare
- Cultural life 30

History 30
- Precolonial period 30
- Colonial period 31
- Struggle for independence 31
- Chile from 1818 to 1920 32
  - The conservative hegemony, 1830–61
  - The widening of liberal influence, 1861–91
  - Political development, 1891–1920
- Chile after 1920 33
  - Political uncertainty, 1920–38
  - The Radical presidencies, 1938–52
  - Political stagnation, 1952–64
  - A period of change, 1964–73
  - The military dictatorship, from 1973

Bibliography 35


## Physical and human geography

### THE LAND

**Relief.** The major landforms of Chile are arranged as three parallel north–south units: the Andes mountains to the east; the intermediate depression, or longitudinal valley, in the centre; and the coastal ranges to the west. These landforms extend lengthwise through the five latitudinal geographic regions into which the country is customarily subdivided. From north to south, with approximate boundaries, these are Norte Grande (extending to 27° S); the north-central region, Norte Chico (27° to 33° S); the central region, Zona Central (33° to 38° S); the south-central region, La Frontera and the Lake District (38° to 42° S); and the extreme southern region, Sur (42° S to Cape Horn).

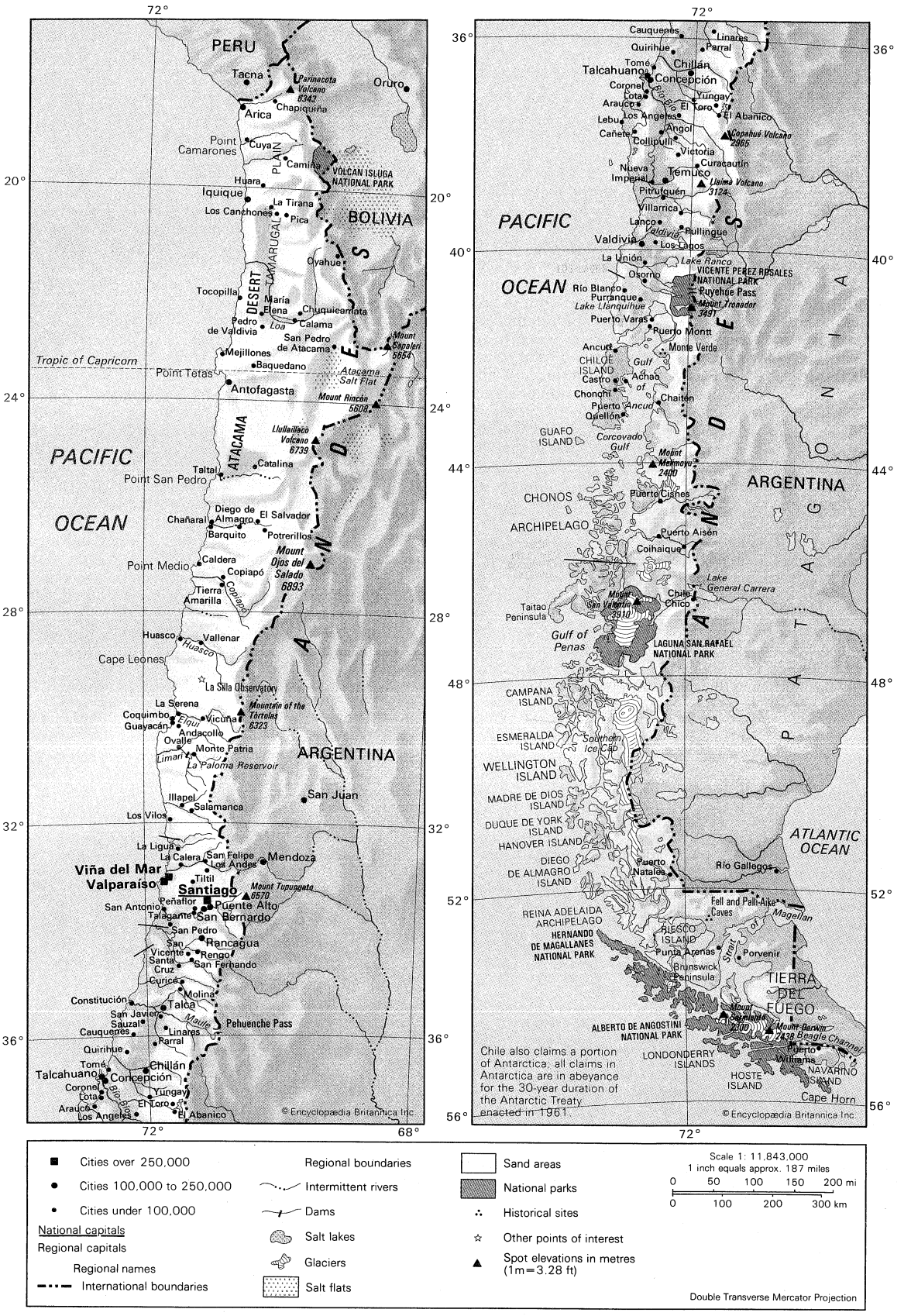
Regions of Chile

*The Chilean Andes.* Extending almost the length of the country, the Chilean Andes, which form most of the border with Argentina, include the highest segment of the Andes mountain chain, which acts as both a physical and a human divide. The Chilean Andean system consists of lofty, often snow-capped mountains, deeply incised valleys, and steep slopes.

Geologic history of the Chilean Andes

The formation of the western Andes ranges began during the Jurassic Period, some 190,000,000 years ago. Marine and terrestrial sediments that had accumulated in the Andean geosyncline were folded and lifted as the Pacific Plate was overridden by the South American Plate. In the early Tertiary Period (beginning 65,000,000 years ago) active volcanism and the injection of effusive rocks laid down the paleovolcanic materials (rhyolites and dacites) that contain the rich copper, iron, silver, molybdenum, and manganese ores of Chile. Also of Tertiary origin are the coal deposits of central Chile.

Later in the Tertiary Period the uplift of the Andes continued, accompanied by further outbursts of volcanism. This active tectonism led to the separation of the Andes from the older coastal ranges and the formation of the intermediate depression. At the beginning of the Quaternary Period the Andes had reached a higher elevation




than at present. During the global cooling that occurred some 2,500,000 years ago, the higher summits were covered by ice masses whose glacier tongues descended into the intermediate depression. Rich sediments were washed down the glacial valleys and deposited into the longitudinal depression. The numerous lakes in the Lake District of south-central Chile are remnants of the ice melting that began some 17,000 years ago. Since the advent of the Holocene Epoch (beginning 10,000 years ago) the Chilean Andes have not changed significantly, but they still experience uplift and episodic volcanic eruptions.

The Andes of northern Chile to latitude 27° S are wide

**MAP INDEX**


**Political subdivisions**

Aisén del General Carlos Ibáñez del Campo .....46 30 s 73 30 w
Antofagasta .....23 30 s 69 00 w
Araucanía .....38 30 s 72 00 w
Atacama .....27 30 s 70 00 w
Bío-Bío .....37 00 s 72 30 w
Coquimbo .....30 45 s 71 00 w
Libertador General Bernardo O'Higgins .....34 30 s 71 00 w
Los Lagos .....41 45 s 73 00 w
Magallanes y Antártica Chilena .....53 00 s 72 00 w
Maule .....35 30 s 71 30 w
Región Metropolitana de Santiago .....33 30 s 70 30 w
Tarapacá .....20 00 s 69 20 w
Valparaíso .....33 00 s 71 15 w

**Cities and towns**

Achao .....42 28 s 73 30 w
Ancud .....41 52 s 73 50 w
Andacollo .....30 14 s 71 06 w
Angol .....37 48 s 72 43 w
Antofagasta .....23 39 s 70 24 w
Arauco .....37 15 s 73 19 w
Arica .....18 29 s 70 20 w
Baquedano .....23 20 s 69 51 w
Barquito .....26 21 s 70 39 w
Calama .....22 28 s 68 56 w
Caldera .....27 04 s 70 50 w
Camiña .....19 18 s 69 26 w
Cañete .....37 48 s 73 24 w
Castro .....42 29 s 73 46 w
Catalina .....25 13 s 69 43 w
Cauquenes .....35 58 s 72 21 w
Chaitén .....42 55 s 72 43 w
Chañaral .....26 21 s 70 37 w
Chapiquiña .....18 23 s 69 33 w
Chile Chico .....46 33 s 71 44 w
Chillán .....36 36 s 72 07 w
Chonchi .....42 38 s 73 47 w
Chuquicamata .....22 19 s 68 56 w
Coihaique .....45 34 s 72 04 w
Collipulli .....37 57 s 72 26 w
Concepción .....36 50 s 73 03 w
Constitución .....35 20 s 72 25 w
Copiapó .....27 22 s 70 20 w
Coquimbo .....29 58 s 71 21 w
Coronel .....37 01 s 73 08 w
Curacautín .....38 26 s 71 53 w
Curicó .....34 59 s 71 14 w
Cuya .....19 07 s 70 08 w
Diego de Almagro .....26 21 s 70 03 w
El Abanico .....37 20 s 71 31 w
El Salvador .....26 14 s 69 37 w
El Toro .....37 17 s 71 28 w
Guayacán .....29 59 s 71 22 w
Huara .....19 59 s 69 47 w
Huasco .....28 28 s 71 14 w
Illapel .....31 38 s 71 10 w
Iquique .....20 13 s 70 10 w
La Calera .....32 47 s 71 12 w
La Ligua .....32 27 s 71 14 w
La Serena .....29 54 s 71 16 w
La Tirana .....20 21 s 69 40 w
La Unión .....40 17 s 73 05 w
Lanco .....39 27 s 72 47 w
Lebu .....37 37 s 73 39 w
Linares .....35 51 s 71 36 w
Los Andes .....32 50 s 70 37 w
Los Angeles .....37 28 s 72 21 w
Los Canchones .....20 27 s 69 37 w
Los Lagos .....39 51 s 72 50 w
Los Vilos .....31 55 s 71 31 w
Lota .....37 05 s 73 10 w
María Elena .....22 21 s 69 40 w
Mejillones .....23 06 s 70 27 w
Molina .....35 07 s 71 17 w
Monte Patria .....30 42 s 70 58 w
Nueva Imperial .....38 44 s 72 57 w
Osorno .....40 34 s 73 09 w
Ovalle .....30 36 s 71 12 w
Oyahue .....21 14 s 68 16 w
Parral .....36 09 s 71 50 w
Pedro de Valdivia .....22 36 s 69 40 w
Peñaflor .....33 37 s 70 55 w
Pica .....20 30 s 69 21 w
Pitrufquén .....38 59 s 72 39 w
Porvenir .....53 18 s 70 22 w
Potrerillos .....26 26 s 69 29 w
Puente Alto .....33 37 s 70 35 w
Puerto Aisén .....45 24 s 72 42 w
Puerto Cisnes .....44 45 s 72 42 w
Puerto Montt .....41 28 s 72 57 w
Puerto Natales .....51 44 s 72 31 w
Puerto Quellón .....43 07 s 73 37 w
Puerto Varas .....41 19 s 72 59 w
Puerto Williams .....54 56 s 67 37 w
Pullingue .....39 33 s 72 11 w
Punta Arenas .....53 09 s 70 55 w
Purranque .....40 55 s 73 10 w
Quirihue .....36 17 s 72 32 w
Rancagua .....34 10 s 70 45 w
Rengo .....34 25 s 70 52 w
Río Blanco .....40 50 s 73 32 w
Salamanca .....31 47 s 70 58 w
San Antonio .....33 35 s 71 38 w
San Bernardo .....33 36 s 70 43 w
San Felipe .....32 45 s 70 44 w
San Fernando .....34 35 s 71 00 w
San Javier .....35 36 s 71 45 w
San Pedro .....33 54 s 71 28 w
San Pedro de Atacama .....22 55 s 68 13 w
San Vicente .....34 26 s 71 05 w
Santa Cruz .....34 38 s 71 22 w
Santiago .....33 27 s 70 40 w
Sauzal .....35 45 s 72 07 w
Talagante .....33 40 s 70 56 w
Talca .....35 26 s 71 40 w
Talcahuano .....36 43 s 73 07 w
Taltal .....25 24 s 70 29 w
Temuco .....38 44 s 72 36 w
Tierra Amarilla .....27 29 s 70 17 w
Tiltil .....33 05 s 70 56 w
Tocopilla .....22 05 s 70 12 w
Tomé .....36 37 s 72 57 w
Valdivia .....39 48 s 73 14 w
Vallenar .....28 35 s 70 46 w
Valparaíso .....33 02 s 71 38 w
Victoria .....38 13 s 72 20 w
Vicuña .....30 02 s 70 44 w
Villarrica .....39 16 s 72 13 w
Viña del Mar .....33 02 s 71 34 w
Yungay .....37 07 s 72 01 w

**Physical features and points of interest**

Alberto de Angostini National Park .....54 40 s 70 00 w
Ancud, Gulf of .....42 05 s 73 00 w
Andes, *mountains* .....36 00 s 71 00 w
Atacama Desert .....24 30 s 69 15 w
Atacama Salt Flat .....23 30 s 68 15 w
Atlantic Ocean .....51 00 s 68 00 w
Bío-Bío, *river* .....36 49 s 73 10 w
Brunswick Peninsula .....53 30 s 71 25 w
Camarones Point .....19 13 s 70 17 w
Campana Island .....48 20 s 75 15 w
Chiloé Island .....42 30 s 73 55 w
Chonos Archipelago .....45 00 s 74 00 w
Copahué Volcano .....37 51 s 71 10 w
Copiapó, *river* .....27 19 s 70 56 w
Corcovado Gulf .....43 30 s 73 30 w
Darwin, Mount .....54 44 s 69 27 w
Diego de Almagro Island .....51 25 s 75 10 w
Duque de York Island .....50 40 s 75 20 w
Elqui, *river* .....29 54 s 71 17 w
Esmeralda Island .....48 57 s 75 25 w
Fell and Palli-Aike Caves, *historical site* .....52 35 s 71 25 w
General Carrera, Lake .....46 30 s 72 00 w
Guafo Island .....43 36 s 74 43 w
Hanover Island .....51 00 s 74 40 w
Hernando de Magallanes National Park .....53 45 s 73 00 w
Horn, Cape .....55 59 s 67 16 w
Hoste Island .....55 15 s 69 00 w
Huasco, *river* .....28 27 s 71 13 w
La Paloma Reservoir .....30 43 s 71 06 w
La Silla Observatory .....29 08 s 70 45 w
Laguna San Rafael National Park .....47 10 s 73 30 w
Leones, Cape .....28 59 s 71 32 w
Limarí, *river* .....30 44 s 71 43 w
Llaima Volcano .....38 43 s 71 43 w
Llanquihue, Lake .....41 08 s 72 48 w
Llullaillaco Volcano .....24 43 s 68 33 w
Loa, *river* .....21 26 s 70 04 w
Londonderry Islands .....55 00 s 70 40 w
Madre de Dios Island .....50 15 s 75 05 w
Magellan, Strait of (Estrecho de Magallanes) .....54 00 s 71 00 w
Maule, *river* .....35 19 s 72 25 w
Medio, Point .....27 11 s 70 59 w
Melimoyu, Mount .....44 05 s 72 52 w
Monte Verde, *historical site* .....41 56 s 72 30 w
Navarino Island .....55 05 s 67 40 w
Ojos del Salado, Mount .....27 06 s 68 32 w
Pacific Ocean .....36 00 s 74 00 w
Parinacota, Volcano .....18 10 s 69 09 w
Pehuenche Pass .....35 59 s 70 24 w
Penas, Gulf of, .....47 22 s 74 50 w
Puyehue Pass .....40 42 s 71 57 w
Ranco, Lake .....40 14 s 72 24 w
Reina Adelaida Archipelago .....52 10 s 74 25 w
Riesco Island .....53 00 s 72 30 w
Rincón, Mount .....24 02 s 67 20 w
San Pedro, Point .....25 30 s 70 38 w
San Valentín, Mount .....46 30 s 73 20 w
Sapaleri, Mount .....22 49 s 67 11 w
Sarmiento, Mount .....54 27 s 70 50 w
Southern Ice Cap (Campo de Hielo Sur) .....48 45 s 73 30 w
Taitao Peninsula .....46 30 s 74 25 w
Tamarugal Plain .....21 00 s 69 25 w
Tetas, Point .....23 31 s 70 38 w
Tierra del Fuego, *island* .....54 00 s 69 00 w
Tórtolas, Mountain of the .....29 56 s 69 54 w
Tronador, Mount .....41 10 s 71 54 w
Tupungato, Mount .....33 22 s 69 47 w
Valdivia, *river* .....39 52 s 73 23 w
Vicente Perez Rosales National Park .....40 50 s 72 10 w
Volcan Isluga National Park .....19 30 s 68 40 w
Wellington Island .....49 20 s 74 40 w

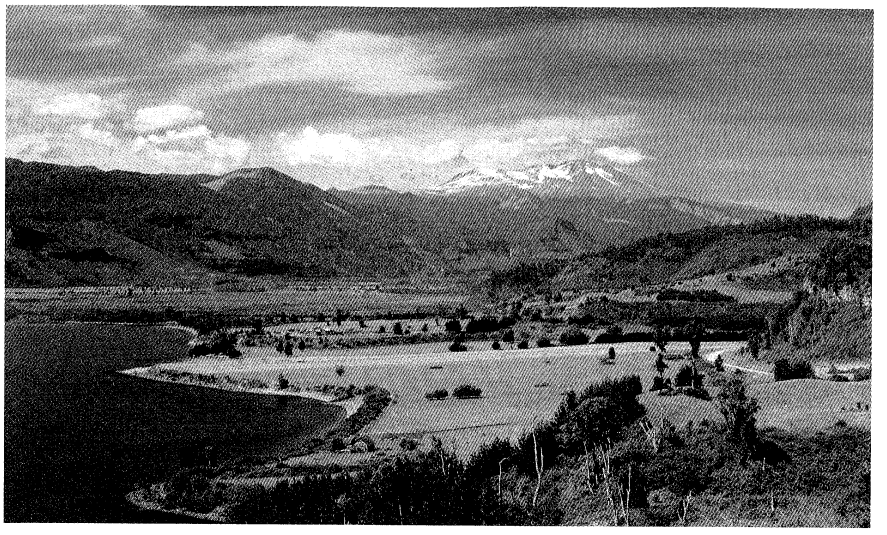

Zones of the Chilean Andes

and arid, with heights generally between 16,500 and 19,500 feet (5,000 and 6,000 metres). Most of the higher summits are extinct volcanoes, such as the Llullaillaco, 22,109 feet; Licancábur, 19,409 feet; and Ojos del Salado, 22,614 feet. After the last glaciation the melting waters collected in shallow lakes in the intermediate elevated basins. Today these salt lake basins (*salares*), the most noted of which is the Atacama Salt Flat, are evaporating to the point of disappearing. Farther south the mountains decrease somewhat in height, but in central Chile, between latitudes 32° and 34°30′ S, they heighten again, with peaks reaching 21,555 feet at Mount Tupungato and 17,270 feet at Maipo Volcano. All of these summits are capped by eternal snow that feeds the numerous rivers of central Chile. Winter sports are pursued in the Andes near Santiago.

Most of the highest mountains between 34°30′ and 42° S are volcanoes, ranging between 8,700 and 11,500 feet. Some of them are extinct while others are still active. Among them are Copahue, Llaima, Osorno, and the highest, Mount Tronador, at an elevation of 11,453 feet. Their perfect conical shapes reflecting on the quiet waters in the Lake District provide some of the most splendid scenery in temperate South America. In southern Chile, below latitude 42° S, the Andes lose elevation and their summits become more separated as a consequence of the Quaternary glacial erosion.

Farther south is Chilean Patagonia, a loosely defined area that includes the subregion of Magallanes and sometimes Chilean Tierra del Fuego. There significant heights are still reached: Mount San Valentín is more than 12,000 feet high, and Mount Darwin in Tierra del Fuego reaches almost 8,000 feet. Reminders of the last ice age are the perfectly U-shaped glacial troughs, sharp-edged mountains, Andean lakes, and some 7,000 square miles of continental ice masses. The Southern Ice Cap, between 48°30′ and 51°30′ S, is the largest in the Southern Hemisphere, with the exception of Antarctica.

*The intermediate depression.* The intermediate depression between the Andes and the coastal ranges is mostly flanked by fault lines. A natural receptacle for materials coming from the Andes, the depression has been filled by



Farmland on Lake Puyehue and, in the right background, Puyehue Volcano, Osorno province, the Lake District.
Chip and Rosa Maria Peterson

alluvial, fluvioglacial, or moraine sediments, depending on the region. In northern Chile it appears as a plateau with elevations between 2,000 and 4,000 feet. Saline sediments that washed down during the Tertiary and Quaternary periods created the rich nitrate deposits found in the Tamarugal Plain and Carmen Salt Flat, where the once-bustling mining towns of María Elena, Pedro de Valdivia, and Baquedano are located. In north-central Chile, extending southward out of the desert region, the depression is interrupted by east–west mountain spurs that create fertile transverse valleys. The Aconcagua River Valley, a transverse valley farther south, marks the beginning of central Chile.

The Central Valley

The alluvial deposits from the numerous Andean rivers in central Chile have provided mineral-rich soils that support the flourishing Mediterranean-type agriculture of the Central Valley of the intermediate depression. These soils and abundant water resources, along with a temperate climate, make the Central Valley the most populated and productive area in Chile. In south-central Chile the intermediate depression is formed by mixtures of fluvial and alluvial depositions, making this region suitable for growing grain and for pastures that support an important dairy industry.

South of the Bío-Bío River dense forests replace open scrub woodland moraines and lakes are common, and the intermediate depression descends to sea level at Puerto Montt. In the extreme south only the Andes and the summits of the coastal ranges are visible because the intermediate depression submerges or is replaced by intracoastal channels and fjords.

*The coastal cordilleras.* In most of northern and central Chile coastal ranges form a ridge between the intermediate depression and the Pacific coast. These mountains, which are seldom higher than 6,500 feet, display smooth forms or flattened summits, since they are considerably older than the Andes. In north-central and central Chile the coastal ranges are built of Paleozoic and Mesozoic granites and metamorphic rocks that were uplifted during the Andean folding phase. In south-central and southern Chile the coastal ranges consist of early Paleozoic metamorphic and igneous rocks, which is evidence of an even earlier folding phase. The coastal ranges were never glaciated, and their former dense vegetation has been destroyed by humans. In places where intensive agriculture has been practiced, the soil is severely eroded and has been depleted of organic and mineral nutrients. Only in the evergreen forests in the Cordillera de Nahuelbuta south of Concepción and the coastal ranges south of Valdivia are the soils well preserved.

On the western margins of the coastal ranges, sea advances during the Tertiary Period deposited thick sediments. During the Quaternary Period sea level changes and continued continental uplift created several coastal terraces in the Tertiary layers, and wave erosion shaped Chile's abrupt coastal line, which has few good natural harbours.

**Drainage.** Most of Chile's rivers originate in the Andes and flow westward to the Pacific Ocean, draining the intermediate depression and the coastal ranges. They are therefore quite short. While their steep gradients and turbulent flow make them unsuitable for navigation—the lower courses of the south-central rivers are an exception—they are particularly useful for hydroelectric power. In areas where water flow is subjected to seasonal variations that hamper agricultural development, dams have been built in order to regulate the rivers and to establish hydroelectric plants.

The rivers

The rivers of Chile have differing physical characteristics that are related to the climatic region in which they are located. In the parched northern region they are fed by the summer rains that fall on the Chilean-Bolivian Altiplano; their volumes are so small that they are either absorbed by the soil or evaporate before reaching the sea. Only the Loa River, the longest Chilean river at some 275 miles, empties into the Pacific Ocean.

The rivers of central Chile have more regular flows and volumes. During the winter months (May–August) they are fed by heavy frontal rains, resulting in frequent flooding of the riverine communities. In late spring (October–November) the rivers receive the runoff from the snow that has accumulated during the winter in the high Andes. This runoff proves quite beneficial for commercial and subsistence crop irrigation. In south-central Chile south of the Bío-Bío River, the steady flow is maintained by constant rains, although there is a slack in discharge during the summer months (December–March). In Chilean Patagonia and Tierra del Fuego intense year-round rains and snowstorms combine to keep the rivers well fed, but their extremely steep drainage into the Pacific renders them totally unusable for commercial purposes.

**Soils.** The geologic variety and diverse origin of surface sediments cause the soils of Chile to vary greatly in character from north to south. In the northern desert region saline soils, made up of gravel and sand cemented with calcium sulfate, alternate with alkali-rich soils, which are difficult to cultivate even with irrigation because of their surface salt accumulations. In river oases salinity also becomes a limiting factor for agriculture. In the transverse valleys of north-central Chile fertile alluvial soils have developed on fluvial deposits, while between the rivers soils are dry and infertile. Within the Central Valley the alluvial soils have developed over fluviovolcanic deposits, which is the reason for their mineral and organic richness. In areas of widespread recent volcanic activity, andosol soils (nutrient-rich soils that develop over volcanic ash)

Rich Central Valley soils

are common. Under good aeration these soils of the Central Valley have excellent agricultural potential, but if the volcanic soils are too permeable, they can be used only for coniferous plantations. In the Lake District the extreme impermeability of the soils leads to the formation of humid soils (*trumaos*). In the southernmost Andes, under conditions of permanent rainfall and cold temperatures, lithosols covered by a thin layer of andosols are the rule: only rain forests grow on such soils. On the archipelagos of Chilean Patagonia and Tierra del Fuego the low terrain is carpeted by moorland soils that support only low shrubs and bog plants of no economic value or potential. Soils at high elevation are characterized by rankers (thin organic soils overlying a rocky substratum) supporting growths of Antarctic beeches.

**Climate.** The extension of Chile across some 38 degrees of latitude encompasses nearly all climates, with the exception of the humid tropics. The Pacific Ocean, the cold Peru (Humboldt) Current, the South Pacific anticyclone winds, and the Andes Mountains constitute the major climatic controls.

The permanent chilling effect of the Peru Current and the constantly blowing southwesterlies emanating from the South Pacific anticyclone determine a temperate climate for most of northern and central Chile. Only the extreme south, unaffected by these controls, is characterized by a cold and humid climate. Temperatures drop in a regular pattern from north to south; the principal cities average the following annual mean temperatures: Arica 64° F (18° C), Antofagasta 61° F (16° C), Santiago 57° F (14° C), Puerto Montt 52° F (11° C), and Punta Arenas 43° F (6° C). During winter, when the polar front advances northward, temperatures drop, though not drastically, owing to the temperate action of the ocean. If snow falls in central Chile, it does not stay on the ground for more than a few hours. During summer, cooling sea winds keep temperatures down and there are no heat waves. The highest monthly means register in the northern desert.

Temperatures of key cities

Annual precipitation differs remarkably from the dry extreme north to the very humid extreme south. North of 27° S latitude there is practically no rainfall. In the north-central region frontal rains in winter account for increasing precipitation: the annual rainfall in Copiapó is less than one inch (21 millimetres). In Santiago the annual rainfall is 13 inches, and along the Central Valley it increases gradually southward until it reaches 73 inches in Puerto Montt, where precipitation occurs throughout the year. The coast of central and south-central Chile is more humid than the Central Valley. In Valparaíso annual precipitation amounts to 15 inches, rising to 52 inches in Concepción and reaching about 90 inches in Valdivia. Farther south, where the westerlies reach their maximum intensity and the polar front is always present, precipitation highs unequaled by any other nontropical region in the world have been recorded; there, San Pedro Point, at latitude 48° S, receives about 160 inches annually. Still farther south, in the rain shadow that occurs on the eastern slopes of the southern Andes, precipitation diminishes drastically, occurring mostly as snow during winter. Punta Arenas, in Chilean Patagonia, receives only 18 inches annually.

Considering all climatic factors and meteorological characteristics, three large climatic regions may be distinguished in Chile: the northern desert, the central Mediterranean zone, and the humid-cool southern region.

Climatic regions

*The northern desert.* This region experiences an aridity that is primarily caused by the dry subsidence created by the South Pacific high pressure cell and the stabilizing action of the cold Peru Current. Although the air along the coast is abnormally humid, it never reaches saturation point; at most, there is a development of coastal fogs (*garúa* or *camanchaca*). Besides the lack of rain, drainage systems, and permanent vegetation, the Chilean desert is characterized by relatively moderate daytime temperatures, the variations in which are dependent upon the direct heat of the Sun; during the night, temperatures may approach the freezing point. In the piedmont oasis of Los Canchones the daily temperature fluctuates up to 47° F (26° C). The interior of the Atacama Desert, which makes up a large portion of the southern part of the desert region, is reported to receive the highest solar radiation in the world.

*Mediterranean central Chile.* The climate of central Chile is characteristic of mid-latitudinal temperate areas. The seasons are well accentuated. Winters are cool and humid as a consequence of continuous passages of fronts and depressions; cloudy days are common. In spring, when there are fewer fronts and the depressions vanish, steady southwest winds and clear skies dominate. During summer, when anticyclonic conditions are established, the days are warm, though not stifling, and without rain. These weather conditions are ideal for the Mediterranean agricultural products that grow so well in central Chile, such as grapes, peaches, plums, honeydew melons, and apricots. Autumn is still sunny and dry, suitable for the ripening of grains, mainly wheat, and vegetables. With the onset of winter, the fronts and depressions return and the accompanying rains last from May to August.

*Southern Chile.* The southern segments of Chile are always under the influence of the polar front and of cyclonic depressions. In addition, the permanently blowing westerlies batter the margins of the continent with oceanic air masses that lower temperatures and cause heavy rainfall along the Pacific coast. Around Cape Horn the westerlies reach their maximum intensity and storms abound. Before the era of steam power, the passage from the Atlantic to the Pacific via Cape Horn was a most feared venture.

**Plant and animal life.** The vegetation of Chile, like the climate and soils, is arranged in latitudinal belts. Only in the Andes is altitude a determining factor. In the northern desert region the vegetation has adapted to the lack of rain and to the salinity of the soils. The tamarugo, a spiny acacia tree, does well in the dry interior desert. Near the coast, and kept alive by the coastal fogs, varieties of cacti as well as shrubs and spiny brambles occur. In the high plateaus of northern Chile hardy species, such as llareta, and grasses, such as ichu and tola, support the Indian population and their llama herds. In semiarid north-central Chile some of the cacti continue, and hardwoods, such as the espino or algarrobo, and shrubs, such as *Adesmia,* become more common. In the more humid and temperate region of central Chile grows a particular vegetal formation called *matorral,* in which hardwoods, shrubs, cacti, and green grass are mixed. Most of this dense growth is disappearing because of the rural population's overexploitation of it for firewood. South of the Bío-Bío River, mixed deciduous forest and evergreen trees are common. Many unique species are found in these humid forests, the most conspicuous being the rauli, or southern cedar, the roble beech, the ulmo (an evergreen shrub), and the evergreen laurel. On the western slopes of the Andes the magnificent monkey puzzle tree, or Chile pine, forms dense stands. A dense rain forest, rich in timber species, grows in the humid Lake District and extends southward. The Antarctic beech, the Chilean cedar, and the giant alerce dominate these often impenetrable southern woods. On the rainy islands of Chilean Patagonia and Tierra del Fuego, the growth of large trees is inhibited by the constant winds and low temperatures. There, only dwarf versions of southern beech and hard grasses are found. In eastern Chilean Patagonia the cold steppes are primarily composed of grasses and herbs that provide grazing for livestock.

Unique trees of the humid forests

The animal life of Chile lacks the diversity of other countries in South America. The barrier of the Andes has restricted animal migrations, and the northern desert has proved a formidable obstacle to the southward migration of tropical Andean fauna. Among the terrestrial animals, the most abundant and varied are the rodents. The chinchilla, the degu, and the mountain viscacha are Andean rodents famed for their fine furs. *Monito de monte,* a marsupial, lives in the deciduous forests and rain forests of the south. The nutria, or coypu (*coipo*) is a water rodent common in the streams of Chile. Among the ruminants are the guanaco, the only survivor of the Paleocamelides (ancient predecessors of the camel family), and its domesticated relatives, the llama, the alpaca, and the vicuña, the latter known for the high-quality wool produced from its

silky fleece; the Indians of the Altiplano make wide use of it. Guanacos are still found from northern Chile to Chilean Patagonia. Two members of the deer family are the huemul, a rarely seen inhabitant of the southern Andes that is represented on the national coat of arms, and the pudu, the smallest known deer. Carnivores are not in great abundance. The puma is the largest, and other feline predators include the *guiña* and the *colocolo*. Among the canids are the Andean wolf and the long-tailed fox. The avian fauna is relatively more diverse, the country being

Exotic bird life

host to wintering migratory birds. Some exotic birds like parrots and flamingos appear over northern and central Chile. Throughout the Chilean Andes there still lives, though reduced in number, the condor, a large scavenger. In Chilean Patagonia is found the carancha, a bird of prey that attacks lambs. Amphibians abound, the most curious being Darwin's frog, discovered by Charles Darwin in south-central Chile. Chile's geographic isolation accounts for the absence of poisonous reptiles and spiders.

**Settlement patterns.** Climatic characteristics and historic events have strongly influenced settlement patterns and population distribution in contemporary Chile.

The Central Nucleus

The early settlement by Spaniards occurred in the temperate part of the country, known as the Central Nucleus, or Zona Central, where the agriculture, industry, and main population centres developed. The area's traditional agriculture developed on the basis of large landed estates, the haciendas, which covered about three-quarters of Chile's arable land. The agrarian reform initiated by the Christian Democratic president Eduardo Frei Montalva in 1965, and continued by the Socialist president Salvador Allende Gossens into the early 1970s, resulted in a redistribution of the land. Agrarian productivity to boost exports was accentuated.

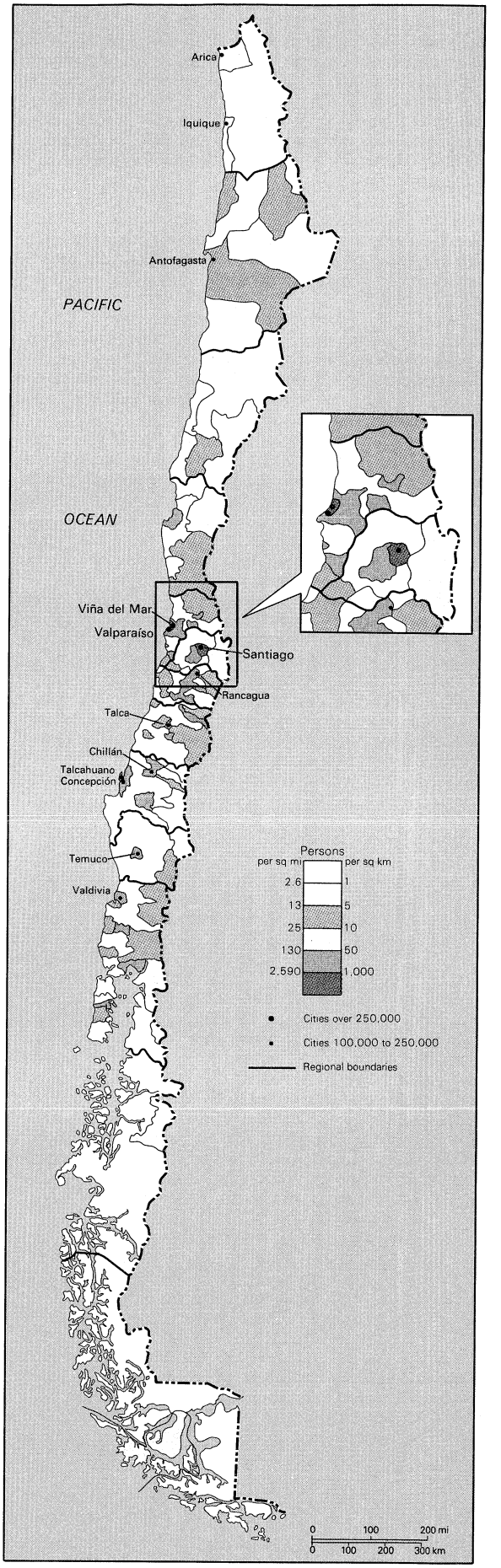
In the Central Nucleus are the major cities of Chile. Santiago was founded there and grew into the country's major metropolis. Seventy miles west of Santiago is the port city of Valparaíso and the neighbouring resort city of Viña del Mar, which form the second largest population centre of Chile. In the Central Valley, south of the Santiago basin, stretches a series of secondary cities, the development of which has been tied to the agricultural success of central Chile. Among them are Rancagua, Curicó, Talca, Chillán, and Los Angeles. All of these cities are connected by rail and the Pan-American Highway.

Most of Chile's cities were founded during the colonial era, and they were arranged around a central square (*plaza de armas*). The original buildings were made of adobe (sun-dried brick) and wood, materials that would deteriorate or burn. Most of the colonial buildings fell prey to earthquakes and fires; much rebuilding took place and the cities of central Chile have become showcases of modern urbanization, high population density, and bustling commercial and industrial activities. On the coast of the southern Central Nucleus lies Concepción and its port city of Talcahuano, both industrial centres.

Norte Chico, the semiarid north-central part of Chile, developed in close association with the Central Nucleus. Agricultural production and mining characterize this region, of which La Serena, near the coast, and the port of Coquimbo are the major centres. The population is primarily concentrated in the irrigated valleys of the Copiapó, Huasco, Elqui, and Limarí rivers or else dispersed in the mountains, where there are mining activities. The main cities, somewhat smaller than those of central Chile, are located in the valleys: they include Copiapó, in the valley of that name, the most important mining centre of the country during the 19th century; Vallenar, Ovalle, and Vicuña. Agriculture, goat raising, and iron and copper mining are the main economic activities. From this region come the famous pisco (a white brandy distilled from sun-dried grapes), fine wines, and high-quality fruits for export.

La Frontera

During colonial times, the fringe of territory at the southern extreme of the Central Nucleus was bitterly contested by Spaniards and Araucanians, the original Indian population, which gave the northern part of south-central Chile its name, La Frontera ("The Border"). After the pacification of the Araucanians in the 1880s, the area was gradually settled by Chileans and by European colonists




Population density of Chile.

who had already begun immigrating there in the 1850s. It developed in modern times as a region of grain growing and commercial pine forestry for cellulose manufacture. The regional capital is Temuco, and in the surrounding countryside still live—in rather precarious conditions—a concentration of Araucanians, locally called Mapuche.

Colonization of the Lake District, located south of La Frontera, began after 1850 with immigrants from Germany, Switzerland, and Belgium. Homesteads, rather than large haciendas as in the Central Nucleus, became the pattern of rural settlement. Although the land has been consolidated in recent times, land fragmentation is still visible. The largest city of this region is Valdivia, founded in early colonial times. This once active industrial centre for footwear, textiles, brewing, and shipbuilding declined after most of its manufacturing installations were destroyed by a 1960 earthquake. Osorno and Puerto Montt are other regional centres, specializing in dairy and flour production. The scenic piedmont lakes and the snow-capped volcanoes attract a steady flow of tourists.

The extreme north and the extreme south could be considered the population and resource frontiers. Both are sparsely populated and rich in natural resources. Settlement of the arid Norte Grande in northernmost Chile began in the middle of the 19th century in response to the exploitation of minerals in the interior. A string of coastal cities emerged as export centres for nitrates, borax, and copper. Iquique, once an exporter of nitrates, has become the capital of Chile's fish meal industry. Antofagasta, the railroad terminus to Oruro, Bolivia, is an active administrative and trading centre and an export facility for the Chuquicamata copper mine. Arica, which acts as a port for Bolivia at the end of the railroad to La Paz, supports fish meal plants and oversees the agricultural production of the Azapa Valley. Once the automobile assembly centre of Chile, Arica has lost its prominence as an industrial city. The only city of significance in the interior of the Norte Grande is Calama, adjacent to the Chuquicamata copper mine, the world's largest open-pit mine. Still, the rest of the area remains picturesque. Old Indian towns, scattered oases, and spectacular desert scenery attract tourists. At the Shrine of La Tirana, on the Tamarugal Plain, Indian and mestizo pilgrims from northern Chile, Bolivia, and southern Peru gather for a colourful festival each July.

Norte Grande

The extreme south encompasses three natural units: the Chiloé island group, the Channels region, and Chilean Patagonia and Tierra del Fuego. Chiloé and its neighbouring islands are among the most undeveloped regions of the country; rudimentary agriculture and algae (used in making confectionary products) and shellfish gathering are the main activities. The small towns of Castro and Ancud are the main population centres of the mostly rural habitat. The Channels region is characterized by islands, separated by glacially carved channels, where colonization has been unsuccessfully attempted since the 1920s. Outlying towns such Puerto Aisén and Coihaique are the only population centres. The region of Magallanes, hinged on the Strait of Magellan, is the most developed area of Chilean Patagonia and Tierra del Fuego. Sheep raising estancias (ranches), which have exported wool since the late 19th century, and oil and natural gas, which have been exploited since 1945, are the pillars of its economy. These activities, combined with meat-packing plants and the trading functions of Punta Arenas, have made this one of the more modernized parts of Chile.

The extreme south

## THE PEOPLE

The Chileans are racially a mixture of Europeans and American Indians. The first miscegenation occurred during the 16th and 17th centuries between the indigenous tribes, including the Atacameños, Diaguitas, Picunches, Araucanians (Mapuches), Huilliches, Pehuenches, and Cuncos, and the conquistadores from Spain. Basque families who migrated to Chile in the 18th century vitalized the economy and joined the old Castilian aristocracy to become the political elite that still dominates the country. Few blacks were brought to Chile as slaves during colonial times because a tropical plantation economy, common in much of the New World, did not develop.

After independence and during the republican era, English, Italian, and French merchants established themselves in the growing cities of Chile and incidentally joined the political or economic elites of the country. The official encouragement of German and Swiss colonization in the Lake District during the second half of the 19th century was exceptional. The censuses of the late 19th century showed that foreigners—principally Spaniards, Argentines, French, Germans, and Italians—formed scarcely more than 1 percent of the total population. At the turn of the century, small numbers of displaced eastern European Jews and Christian Syrians and Palestinians fleeing the Ottoman Empire arrived in Chile. Today they spearhead financial and small manufacturing operations.

Racial mixture

The population displays a strong sense of cultural identity, which can be traced to the predominance of the Spanish language, the Roman Catholic religion, and the comparative isolation of Chile from the rest of South America. The Araucanian Indians form the only significant ethnic minority.

The trend of age-group distribution of Chile's more than 12,000,000 people, with increasingly larger numbers in the older brackets, reflects a progressive maturing of the population. Life expectancy rose from 57 years in 1960 to about 70 years by the early 1980s. These demographic changes reflect both improved health-care conditions and modernization of the life-style by the predominantly urban population. Also ascribed to the same factors is the dramatic decline during the late 20th century in infant mortality and in the fertility rate. Chile's crude death rate is lower than that of most of its South American neighbours.

The large cities and the industrial centres of central Chile attract a steady flow of internal migrants. Most of them head for the capital city of Santiago, with the rest going primarily to Valparaíso–Viña del Mar and to Concepción–Talcahuano. These migrants emanate mostly out of the rural regions of the Central Valley and north-central Chile. The northern coastal cities receive some migrants from Santiago and Valparaíso and also from the small villages in the far north. Chiloé has been losing its population to Punta Arenas and the agrarian areas of the Lake District, and even to Argentina, where Chilotes work on estancias or in the mines of Patagonia. After 1973, hundreds of thousands of Chileans left the country for political reasons to live in exile.

Internal migration

## THE ECONOMY

The Chilean economy is based on the exploitation of agricultural, fishing, forest, and mining resources. Chile developed historically on the basis of a few agricultural and mineral exports, as was common in Latin America. Many manufactured products had to be imported, and land, wealth, and power were concentrated in the hands of a small aristocracy. Although there have been land reforms and development of manufacturing, many of Chile's economic problems in the 20th century are related to the country's early economic structure.

During the 19th century the Chilean economy grew on the basis of exported agricultural products, copper, and nitrates. After the nitrate market dropped during World War I, Chile's economy took a sharp downturn, intensifying the effect on the country of the Great Depression. These events turned Chile toward more socialistic programs that featured strong government control of the economy. An attempt was made to develop import substitution industries so as to lessen dependence on imported products. Industrial growth was placed in the hands of the Corporación de Fomento de la Producción (Corfo; the Development Corporation). Agrarian reforms were instituted, and the government assumed greater control of industry, especially during the administrations of Pedro Aquirre Cerda (1938–41) and Salvador Allende Gossens (1970–73), when many banks, copper mines, and business firms were nationalized. The economy at first improved under these policies, inflation going down and the gross domestic product increasing. The government, however, was unable to establish a sound tax base to match the expanding economy; by 1973 conditions were deteriorating rapidly and a military coup overthrew the government.

Economic development

The new regime instituted more conservative, free-market programs and reversed many of the previous governments' acts. The country has continued to face severe economic problems, reflected in periodic high inflation, fluctuating trade policies, unemployment, and heavy dependence on a single major export, copper, in an unstable market.

**Resources.** A geographically varied country, Chile is rich in mineral deposits, natural forests, sea resources, and energy sources.

*Mineral resources, noncarboniferous.* Mining, historically the mainstay of the Chilean economy, has been catalytic to both external commerce and domestic industrial development. Copper, molybdenum, iron, nitrates, and other concentrated minerals make up a large part of the total value of national exports.

Metals account for the highest percentage in mining, copper being primary. Chile is the world's largest producer and exporter of copper. Copper mines are located in northern Chile (Chuquicamata and El Salvador) and along the Andes of north-central Chile (especially El Teniente and Andina). Small-scale extractions are carried out by individuals, or *pirquineros,* who operate in the uplands of north-central Chile and in the coastal ranges of central Chile. Medium-sized activity lies in the hands of companies with a larger investment capacity and with their own treatment plant. Large-scale mining was developed with U.S. capital at the beginning of the 20th century.

Copper

Copper plays the role in the Chilean economy that was occupied by nitrates prior to World War I. The large U.S. corporations were tranformed into mixed-ownership enterprises during the late 1960s and totally nationalized during the early 1970s, when the mining and sales were turned over to the Corporación Nacional del Cobre de Chile (Codelco). A drop in world market prices influences production and sales and creates financial hardship.

Iron-ore mining in El Tofo and El Romeral, both in north-central Chile, is significant, and manganese, silver and gold, and molybdenum (a metal derived from the large copper deposits) are also mined. Among nonmetallic minerals, sulfur, gypsum, lithium, and limestone are moderately exploited. Nitrate deposits occur in the northern interior desert. Their economic value, so important during the 19th century, has decreased, but the production of iodine, a by-product of nitrate, is of major importance.

*Energy resources.* Hydroelectric potential and installed capabilities, as well as coal and moderate oil and natural gas reserves, furnish Chile with good energy resources. The steady flow of the Andean rivers has been utilized by the Empresa Nacional de Electricidad (Endesa; National Electric Company) to produce hydroelectricity. Hydroelectric development has been extended to the coastal mountain ranges. Prior to the installation of Chile's huge electrical system, most of the country's energy, including electricity, was obtained from soft coal, mined since the 19th century in the Gulf of Arauco, south of Concepción. Petroleum and natural gas are extracted on Tierra del Fuego and along the northern shore of the Strait of Magellan and are shipped to refineries in central Chile. Production, however, meets only about half of the country's oil requirements.

Hydroelectric potential

*Forestry resources.* South of the Bío-Bío river, climatic conditions favour the growth of natural forests. The primary species used for lumber and paneling are the coigue, oak, rauli, ulmo, tepa (laurel tree), and monkey puzzle tree. Pine for the manufacture of paper and pulp is taken from forests in central Chile and the Bío-Bío region.

*Fishing resources.* Since 1974, after the collapse of the fishing industry of Peru, Chile has become the chief fishery of South America, and it is one of the foremost fishing nations of the world. Sardines, jack mackerel, chub mackerel, hake, and anchovy constitute most of the catch. The principal products are fish meal and fish oil, which are shipped to Europe and the United States for the production of animal feed and industrial oil. The fish-processing plants—all privately owned—are mainly located in the northern cities of Iquique, Arica, and Antofagasta.

**Agriculture.** While good climatic conditions and abundant water resources favour Chile's agriculture, outdated land-tenure patterns, managerial ineptitude, and inadequate price policies have combined to make agriculture one of the most inefficient sectors of the economy. Employing approximately one-sixth of the labour force, agriculture generates less than one-tenth of the gross domestic product. To meet expenditures and credit payments abroad, the military government that took over in 1973 strongly encouraged exports of agricultural commodities by private national and international companies. Within the framework of this policy, Chile increased remarkably the export of fresh fruit, canned vegetables, and wines.

In temperate central Chile the primary crops are cereals (chiefly wheat) followed by grapes, potatoes, corn (maize), apples, beans, rice, and a variety of vegetables. Industrial crops, such as sugar beets and sunflower seeds for cooking oil, are also common.

Stock raising has been one of the most underdeveloped activities in rural areas, partly because of poor technology and inefficient breeding. Cattle are the major livestock. There has been, however, some expansion in poultry, lamb, and pork production, as well as that of beef.

**Industry.** An estimated one-eighth of the economically active population is employed in the industrial sector, which accounts for about one-fifth of the gross domestic product. Factories are concentrated in the principal urban centres—Santiago, Valparaíso, and Concepción. Light industries produce appliances, chemical products, food products, textiles and clothing, and construction materials.

Larger industrial complexes are located in the San Vicente harbour of Concepción; they include the Huachipato iron and steel mill, fish-processing factories, and a petroleum refinery associated with a petrochemical complex. Another such refinery is situated in Concón, at the mouth of the Aconcagua River. Pulp and paper mills thrive in the vicinities of the Bío-Bío and Laja rivers. (C.N.Ca.)

**Trade and finance.** Chile's principal markets for mining and agricultural commodities are Europe, the United States, and East Asia. Most imports are from the United States, Venezuela, Argentina, Brazil, West Germany, and Japan. The balance of payments, generally unfavourable since the 1950s because of increased foreign expenditures and payment of external loans, showed occasional improvement after 1976 but with considerable fluctuation.

Balance of payments

The Banco Central de Chile, established in 1925, is the official bank of the nation; it implements the internal banking policies of the government and also conducts foreign trade. In 1989 the bank became an autonomous institution entirely responsible for the country's financial and exchange-rate policies. The Banco del Estado de Chile is also a state entity, but it functions as a private commercial bank. National private banks as well as international banks from Europe, the United States, and Asia operate freely in the country.

Within the Chilean economic system there is collaboration between the private and public sectors, with the private sector contributing an increasing percentage of the total annual investment. Private businesses are generally organized as joint-stock companies (similar to U.S. corporations) that participate in all areas of economic activity.

**Transportation.** The country's length and physical barriers constrain communication and traffic flow. Only the sea offers an expeditious means of transportation, which was taken advantage of during the 19th century when Chile owned one of the largest merchant fleets of Latin America. Chile's overall economic decline during the early 20th century and the supplanting of maritime transport with overland means signified its demise. Eventually only international transport was conducted by ship. The main port of entry is Valparaíso. San Antonio, the port for Santiago, exports copper and agricultural commodities. Other busy ports, such as Antofagasta and Arica, serve the international trade with Bolivia. Chañaral, Huasco, Guayacán, and Tocopilla transport minerals. The port of Talcahuano serves the industrial complex of Concepción.

The development of an overland transportation system began with two railway systems initiated around the turn of the century: the northern network, between La Calera (near Valparaíso) and Iquique, now in disuse, and the southern network, between La Calera and Puerto Montt. The most traveled sections connect Santiago with Valparaíso and Santiago with Puerto Montt; both sections

are electrified, making them more competitive with road transportation. The railway system is controlled by the Empresa de los Ferrocarriles del Estado (State Railway Enterprise). International railroads connect Arica and La Paz (Bolivia), Antofagasta and Oruro (Bolivia), and Los Andes and Mendoza (Argentina). A railbus transports passengers over the short route between Arica and Tacna (Peru).

Highways

Chile's rapid motorization has brought enhanced highway transportation for passengers and goods. The backbone of the Chilean road system is the paved Pan-American Highway, which connects Arica with Quellón on Chiloé Island, a distance of more than 2,100 miles. From this main artery secondary routes connect numerous cities, including Santiago, with the ports of San Antonio and Valparaíso, Bulnes with Concepción, and Los Lagos with Valdivia. The most important international paved road connects Santiago with Mendoza (Argentina). All-weather roads connect Iquique with Oruro (Bolivia), Antofagasta with Salta (Argentina), La Serena with San Juan (Argentina), Osorno with San Carlos de Bariloche (Argentina), and Punta Arenas with Río Gallegos (Argentina).

Air transport, serving mostly passengers between the cities at both extremes of the country and some towns of difficult access, such as El Salvador and Coihayque, is supplied by the state-owned Línea Aérea Nacional de Chile (LAN; National Airline of Chile) and by Línea Aérea del Cobre (Ladeco; Copper Airline). A tourist service is maintained by LAN between Santiago and Easter Island, in the Pacific, with the flight continuing to Papeete, Tahiti. All major South American lines, plus others from the United States and Europe, handle the flow of international passengers to the Arturo Merino Benítez airport near Santiago. Chacalluta, northeast of Arica, is another major airport.

## ADMINISTRATION AND SOCIAL CONDITIONS

**Government.** The Republic of Chile, inaugurated in 1821, has had a long history of representative democracy, with only a few short-lived exceptions. Historically, Chile has been renowned for its political freedom. From September 1973 to March 1990, however, a military junta headed by Gen. Augusto Pinochet Ugarte presided over the longest period of authoritarian dictatorship in Chilean history. The country is governed in accordance with the constitution of 1981, approved by a plebiscite called by General Pinochet to change the constitution of 1925. The 1981 document placed the administration of the state into the hands of the president and permitted Pinochet to hold office until 1990. The president appoints the state ministers, who are unimpeachable and totally in his confidence.

Under the 1981 constitution a presidential candidate chosen by the junta was to be put up for approval in a national plebiscite and, if approved, to serve as president until 1997. The junta nominated Pinochet, who was rejected in the 1988 plebiscite, and the first presidential elections since the 1973 coup were held in December 1989.

The bicameral National Congress was dissolved at the time of the 1973 coup, after which legislative functions were carried on by the junta, assisted by legislative commissions. The 1981 constitution allows for a bicameral legislature consisting of an upper chamber, or Senado, and a lower chamber of representatives, or Camara de Diputados, to be elected by direct popular vote. These two bodies remained in recess until the elections of December 1989.

The constitution of 1981

The justices and prosecutors of the Supreme Court and the Courts of Appeals are appointed by the president from a list of nominees proposed by the Supreme Court. Judges are career functionaries of the Ministry of Justice. The composition of the lower courts is similarly determined.

Local government is carried on through 12 administrative regions and the capital, Santiago. The regions are divided administratively into provinces, which in turn are divided into communes. The president appoints the intendents (*intendentes*) who head the administrations of the regions and Santiago. The intendents govern with the aid of a regional council, which may include the governors of the constituent provinces and representatives of various other private and public institutions within the region. The provincial governors, like the intendents, serve at the sole pleasure of the president. The communes are administered by a municipal corporation (*municipalidad*) composed of a mayor (*alcalde*) and a communal council. The mayor is appointed by the regional council from a list of three candidates submitted by the communal council; in the case of some larger urban centres, the mayor is appointed directly by the president. The councilmen (*regidores*) are elected by popular vote for four-year terms.

In September 1973, when the junta suspended all activity by political parties and outlawed Marxist parties, the political spectrum extended from right through centre to left. At the extreme right stood the National Party (Partido Nacional)—an alliance formed in 1965 that included the former Liberal Party (Partido Liberal) and United Conservative Party (Partido Conservador Unido)—and the Radical Democrats (Democracia Radical)—the right wing of the once-populist Radical Party (Partido Radical; see below), from which it split in 1969. The centre was occupied by the Christian Democratic Party (Partido Demócrata Cristiano), which, since the congressional elections of 1965, had attracted the most voters. With many of its partisans coming from the extensive Chilean bureaucracy, the social democratic Radical Party had been another important centrist party, but after 1965 it drifted to the left, where in 1973 it was reduced to a minority partner. The core of the left consisted of the Socialist Party (Partido Socialista) and the Soviet-oriented Communist Party of Chile (Partido Comunista Chileno). Several Marxist groups affiliated with the two major powers of the left included the Unified Movement of Popular Action (Movimiento de Acción Popular Unitaria), the Christian Left (Izquierda Cristiana), and the extremely militant Movement of the Revolutionary Left (Movimiento de Izquierda Revolucionaria).

After 1980 General Pinochet eased his severe repression of political activism, and in 1981 some political exiles were allowed to return to Chile, but political activity was still heavily restricted. The rightist groups rallied in a movement called the National Democratic Accord (Acuerdo Democrático Nacional). The parties of the centre, led by the Christian Democrats, formed the Democratic Alliance (Alianza Democrática). The exiled leaders of the left split into the Socialist Convergence (Convergencia Socialista), which in 1983 merged into the Chilean Socialist Bloc (Bloque Socialista Chileno), and the Popular Democratic Movement (Movimiento Democrático Popular), dominated by the Communist Party of Chile. Reluctant to join forces with the Socialist Convergence in opposing the Pinochet government, the Christian Democratic Party isolated itself from the right and the left, and this climate of dissension weakened the opposition. The National Democratic Accord and the Popular Democratic Movement were dissolved in 1985.

Political parties

To organize opposition to Pinochet in the 1988 plebiscite, 16 centrist and leftist parties formed the Command for No (Comando por el No), which after Pinochet's defeat became the strongest opposition group and was renamed the Coalition of Parties for Democracy (Condertación de los Partidos por la Democracia). In July 1989, constitutional amendments worked out between the government and the opposition were approved in a national referendum; included was the revocation of Article Eight, which banned Marxist parties. Two months later the government declared, with some restrictions, that all political exiles were permitted to return to Chile. (C.N.Ca./Ed.)

**Education.** Chile's educational system, structured along the lines of 19th-century French and German models and highly regarded among Latin-American countries, is divided into eight years of free and compulsory basic (primary) education, four years of optional secondary or vocational education, and additional (varying) years of higher education. Education is received by most primary-school-age individuals, accounting for the low illiteracy rate of 4 percent for persons 12 years of age and over. After the 1973 coup education was required to adopt values consistent with the junta's Christian, anti-Marxist values. Private schools, which are run by religious congregations, ethnic groups (such as German, French, Italian, and Israeli), and private educators have relatively high enrollments.

University education in Chile is of considerable renown throughout Latin America. The major institution is the

Chilean universities

University of Chile (originally founded in 1738), with campuses in Santiago, Arica, Talca, and Temuco. The University of Santiago of Chile and the Federico Santa Marta Technical University, in Valparaíso, are technical universities patterned after the German model. Private universities are the Catholic University of Chile in Santiago, the Catholic University of Valparaíso, the University of the North in Antofagasta, the University of Concepción, and the Southern University of Chile in Valdivia.

**Health and welfare.** Social welfare and labour legislation evolved earlier in Chile than it did in other Latin-American countries, and they have reached a high level of development. Legislation was passed in the early part of the 20th century that regulated labour contracts, workers' health, and accident insurance. In successive years the social security system expanded in an attempt to cover all labour sectors. All workers were eventually covered by the Social Insurance System, maintained through contributions of employers, employees, and the state. Since 1973 the military government has changed social security into an individual savings scheme in which workers invest with private companies.

Health care also developed remarkably during the first half of the 20th century by means of state health plans managed by the National Health Service, a subsidiary of the Ministry of Public Health. An increasing number of facilities, equipment, and qualified personnel have reduced morbidity and infant mortality, eradicated tuberculosis, and brought infectious diseases under control. A movement by the Pinochet government to modify the state-administered public health system by introducing a profit-oriented private health system began in 1980. It offered the option of private health care to those who could afford it.

Public health

## CULTURAL LIFE

Language and a common history have promoted cultural homogeneity in the country. Even the Araucanians and certain Aymara minorities in the north share the values of the Chilean identity, while continuing to cherish their own cultural heritage. Chileans have always displayed a high degree of tolerance toward the customs and traditions of minority groups, as well as toward Christian and non-Christian religious practices.

The flavour of local custom and tradition in Chile is readily observable in the numerous colourful religious festivals that take place at various localities throughout the country. Hundreds of thousands of spectators are drawn to these processions.

Literature

*The arts.* Literature, poetry in particular, is the most significant of the creative arts in Chile. Two Chilean poets, Gabriela Mistral and Pablo Neruda, won the Nobel Prize for Literature (1945 and 1971, respectively), and the poetry of Vicente Huidobro and Nicanor Parra, also of the 20th century, is recognized in the world of Hispanic literature. Fiction, on the other hand, has not been a successful genre, perhaps because of its marked parochialism. Manuel Rojas enjoyed, during the 1950s and 1960s, a degree of international popularity, and more recently the novels of Isabel Allende have become highly acclaimed not only in Latin America but also, in translation, in Europe and North America.

Much of the fine and performing arts of Chile is centred in Santiago, and the main season for cultural events is between March and November. The most famed Chilean musician has long been the pianist Claudio Arrau. Composers such as Enrique Soro and Juan Orrego are noted in the Latin-American world of music, but they never achieved world recognition. The Chilean National Symphony Orchestra and several chamber music ensembles keep European musical culture alive in Chile. Dance and opera are highlighted by the Municipal Ballet and Opera and the National Ballet of the University of Chile. Contemporary folk music, particularly *tonadas* (poetic tunes accompanied by guitar), had its halcyon days in the 1960s and early 1970s, when protest and social-content songs were fashionable. Violeta Parra, who died in 1969, excelled in this style.

Performing arts

Santiago in particular is a hub of art galleries where the works of Chile's artists are displayed and sold. The country, however, has produced few artists of high acclaim. The painter Roberto Matta Echaurren and the sculptor Marta Colvin are among those of significance.

*Cultural institutions.* The country, and Santiago in particular, is rich in museums of fine arts; modern, folk, colonial, and pre-Columbian art; natural history; and Chilean national history. The Museum of National History is of particular note, and others include the Museum of Fine Arts, the Museum of Contemporary Art, and the Museum of Natural Science, all in Santiago. The main library, the National Library of Chile, ranks among the largest in Latin America.

Museums

*Recreation.* There is ample recreational and sports opportunity in Chile; the people can engage in most such activities common to Western cultures. The Pacific beaches are notably beautiful, but the cold water encourages more sunbathing than swimming. Viña del Mar is a particularly well-known summer resort, and the scenery of the Lake District to the south attracts many tourists. As in many Latin-American countries, football (soccer) arouses a particular devotion among the populace, and crowds of up to 80,000 attend matches in Santiago. In this mountainous country skiing is enjoyed by devotees who flock to ski resorts, such as those at Portillo and Farellones (near Santiago) and those near Chillán to the south.

*Press and broadcasting.* The degree of literacy and the demand for national and international information keeps a large number of journals and magazines in publication. Prior to the 1973 military coup, practically all political groups published their own daily or weekly journals. After the coup only journals that refrained from criticizing the government were allowed and censorship was strict and implacable. After 1981, books of political content or dissent were allowed to be published, provided the author was not suspected of being a Marxist. Radio and television stations followed policies of focusing attention away from poignant socioeconomic and political problems of the country. By tradition the stations have been operated by the universities but as commercial, profit-oriented enterprises. In 1967 a government channel was founded, which was used by subsequent administrations to disseminate propaganda.

State control of media

For statistical data on the land and people of Chile, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL. (C.N.Ca.)

# History

## PRECOLONIAL PERIOD

At the time of the Spanish conquest of Chile in the mid-16th century, at least 500,000 Indians inhabited the region. Nearly all of the scattered tribes were related in race and language, but they lacked any central governmental organization. The groups in northern Chile lived by fishing and by farming in the oases. In the 15th century they fell under the influence of expanding civilizations from Peru, first the Chincha and then the Quechua, who formed part of the extensive Inca Empire. Those invaders also tried unsuccessfully to conquer central and southern Chile.

Original inhabitants

The Araucanian Indian groups were dispersed throughout southern Chile. These mobile peoples lived in family clusters and small villages. A few engaged in subsistence agriculture, but most thrived from hunting, gathering, fishing, trading, and warring. The Araucanians resisted the Spanish as they had the Incas, but fighting and disease reduced their numbers by two-thirds during the first century after the Europeans arrived.

The Spanish conquest of Chile began in 1536–37, when forces under Diego de Almagro, associate and subsequent rival of Francisco Pizarro, invaded the region as far south as the Maule River in search of an "Otro Peru" ("Another Peru"). Finding neither a high civilization nor gold, the Spaniards decided to return immediately to Peru. The discouraging reports brought back by Almagro's men forestalled further attempts at conquest until 1540–41, when Pizarro, after the death of Almagro, granted Pedro de Valdivia license to conquer and colonize the area. Valdivia, with about 150 companions, including his mistress, Inés Suárez, the only Spanish woman in the company, entered

Chile in late 1540 and founded Santiago (Feb. 12, 1541). For the next two decades the settlers lived a precarious existence and were constantly threatened by the Indians, who resisted enslavement. Before the safety of the colony was guaranteed, land was apportioned to the conquerors, and thus began the system of large estates. The estates were later institutionalized through the *mayorazgo,* a practice of transmitting estates by entail.

The establishment of estates

Valdivia did not undertake the conquest of the region south of the Bío-Bío River until 1550. In that year Concepción was founded, and preparations were made to move southward. During the next two years settlements and forts were established in La Frontera, but in 1553 the Araucanian Indians, under a skilled military chieftain named Lautaro, rose in a revolt that led to the capture and death of Valdivia and to the beginning of a costly struggle. The Araucanians, often referred to as the Apache of South America, kept the struggle alive until the 1880s by successfully adapting their way of life and military tactics to changing conditions.

Although Concepción was destroyed on several occasions, it remained as the Spanish outpost in the south as did La Serena, founded in 1544, in the north. The province of Cuyo held the same position east of the Andes until 1776, when it was made a part of the newly created Viceroyalty of the Río de la Plata. The conquest of Chile was finally consolidated during the late 1550s under Gov. Don García Hurtado de Mendoza. Before the end of the 16th century English pirates and freebooters, including Sir Francis Drake and Thomas Cavendish, and later Dutch adventurers harassed the coast in search of sudden wealth and as part of a prolonged effort to force Spain to permit neutral nations to trade with its New World colonies.

## COLONIAL PERIOD

Because only quite limited amounts of precious metal were found in Chile, the settlers early turned their attention to agriculture. They grew a wide variety of cereals, vegetables, and fruits; raised livestock; and consumed nearly all of their production locally. Largely because of the poverty of the colony, there were never more than a few thousand black slaves; and, because the Indians proved to be an unreliable source of labour, the settlers often had to work the fields themselves. The lack of mineral wealth also made the area unattractive to Spaniards, and at the end of the 16th century there were no more than 5,000 Spanish settlers in the entire colony. In this regard it should be pointed out that, beginning in 1600 and continuing until trade restrictions were relaxed in the late colonial period, Chile was a "deficit area" in the empire, and the Spanish crown had to provide an annual subsidy to meet the expense of maintaining officials in Santiago and an army on the Araucanian frontier.

Chile lived under the same administrative and religious systems as its neighbours, but because the colony was poor, there was until the 18th century a tendency to send mediocre officials to preside over its destinies. The Spanish crown and the Roman Catholic Church combined to limit the colonists' administrative experience and economic development. The power of the captain-general, the highest royal official in the colony, was absolute. Appeals to the viceroy in Peru or the king in Spain were always possible, at least in theory. Chilean trade was tightly controlled from Peru. The influence of the Catholic Church in secular affairs was always significant and frequently decisive.

The most apparent social development after 1600 was the rapid growth of a mestizo (mixed Indian and European) group, which gives present-day Chile its homogeneous ethnic character. By the end of the colonial period, when the population reached an estimated 500,000 (not including unsubjugated Indians), approximately 300,000 were mestizos and about 150,000 were Creoles (native-born persons of European descent). About 20,000 were *peninsulares* (recently arrived Spaniards), perhaps 15,000 were blacks, and a handful were recently emancipated Indians. Society was highly structured, with *peninsulares* at the top, followed by Creoles, mestizos, Indians, and African slaves. At the end of the colonial period, the vast majority of the population was concentrated in the Aconcagua Valley and the Central Valley (extending from Santiago to Concepción), which together form "the cradle of Chilean nationality."

Growth of the mestizo group

Education in colonial Chile was almost a complete monopoly of the Catholic clergy and reinforced the society's strong class differences. In 1758, however, courses were opened in the Royal and Pontifical University of San Felipe at Santiago and attracted students from the Spanish colonies across the Andes. Nonetheless, intellectual life in Chile developed slowly. The colony did not have a printing press until shortly before it won independence from Spain in 1818, and the paucity of contacts with the outside world reinforced its insularity.

## STRUGGLE FOR INDEPENDENCE

Despite the colony's isolation, its inhabitants at the start of the 19th century were affected by developments elsewhere. The most significant of those developments were the winning of independence by the 13 Anglo-American colonies and by Haiti, the French Revolution, and the inability of Spain to defend its system in America, as indicated by the British invasion of the La Plata region and increased contraband trade on the part of British and U.S. citizens. Finally and decisively came the intervention of Napoleon in Spain, an act that in 1808 threw Chile and the other colonies on their own resources and led them to take the first steps toward greater autonomy and self-government. In Chile the initial move toward independence was made on Sept. 18, 1810, when a *cabildo abierto* (open town meeting) in Santiago, attended by representatives of privileged groups whose vaguely defined objectives included a change in administration, accepted the resignation of the President-Governor and in his place elected a junta composed of local leaders.

From 1810 to 1813 the course of the patriots was relatively peaceful because they were able to maintain themselves without formal ties to the Viceroyalty of Lima. Trade restrictions were relaxed; steps were taken toward the eventual abolition of slavery; a newspaper was established to publicize the beliefs of the patriots; and education was promoted, including the founding of the National Institute. However, the embers of civil strife were also fanned. The Creoles were divided over how far the colony should go toward self-government. José Miguel Carrera and his brothers, whose desire for complete independence was equaled if not surpassed by their personal ambition, inflamed the issues. Meanwhile, Spain had taken steps to reassert its control over the colony. At the Battle of Rancagua, on Oct. 1 and 2, 1814, it reestablished its military supremacy and ended what has been called *la patria vieja* ("The old fatherland").

The Carrera brothers

Following the defeat at Rancagua, patriot leaders, among them the Carrera brothers and Bernardo O'Higgins, future director-dictator of Chile, migrated to Argentina. There O'Higgins won the support of José de San Martín, who, with the support of the revolutionary government in Buenos Aires, was raising an army to free the southern portion of the continent by first liberating Chile and then attacking Peru from the sea. The Carreras continued their spirited agitation for independence in Buenos Aires and the United States.

Meanwhile, many of those who remained in Chile suffered from the harsh rule of Spain's inept representatives and became convinced that absolute independence was necessary. In January 1817 San Martín's well-drilled army, with O'Higgins as one of its commanders, began its march across the Andes; and on Feb. 12, 1817, the patriot forces defeated the royalists on the hill of Chacabuco, which opened the way to Santiago. O'Higgins was proclaimed supreme director of Chile, although the act of declaring Chile's independence was not taken until a year later (Feb. 12, 1818), on the first anniversary of Chacabuco; and the decisive defeat of Spain on the Chilean mainland (Spain held the island of Chiloé until 1826) did not come until the Battle of Maipú, on April 5, 1818. Before emancipation was assured, O'Higgins began the creation of the Chilean navy, which by late 1818 was in the process of clearing the Chilean coast of Spanish vessels.

The liberation of Chile

Chile was free, but its inherent weaknesses were every-

where manifest. The Creoles remained bitterly divided between O'Higgins and the Carreras. Two of the Carrera brothers had been executed in Mendoza, Arg., in 1818; and José Miguel Carrera suffered the same fate in the same city in 1821. The elite groups were dedicated to the retention of those institutions on which such things as law, property, family, and religion were founded. The masses, who had been little more than spectators in the conflicts between 1810 and 1818, were excluded from government.

(J.J.Jo./P.W.D.)

### CHILE FROM 1818 TO 1920

The Chilean oligarchy had little sympathy with O'Higgins, who favoured reducing their privileges. They accepted him, however, because he was supported by the army and because of dangers posed by Spaniards still in Peru and in parts of Chile (Valdivia and the island of Chiloé) and by internal guerrillas loyal to the Spanish monarchy. Opposition to O'Higgins began to make itself heard once the Chilean-Argentine army expelled the Spaniards from Peru; it increased after 1822, when the Chileans succeeded in driving the remaining Spaniards from Chile. O'Higgins' attempt, by means of a new constitution, to concede a larger political role to the oligarchy did not increase his support, and general unrest and poor harvests forced him to abdicate in 1823.

O'Higgins' abdication

The years 1823–30 were troubled by an internal political split between the oligarchy and the army; 30 successive governments held office, and a variety of political experiments were tried. Rivalries developed between federalists and centralizers and between authoritarians and liberals. To the political chaos were added financial and economic disorder and an increase in lawlessness that tended to strengthen the authoritarian members of the oligarchy. Rival political factions were eliminated in 1829 when authoritarians, with the help of a part of the army, were able to install a junta (collegial government) that nominated José Tomás de Ovalle as provisory president. Actual power, however, was held by Diego Portales, who, as either a cabinet member or a private citizen, in fact ruled as a virtual dictator.

**The conservative hegemony, 1830–61.** During the next 30 years, Chile established its own definitive organization, made possible by a compromise among the members of the oligarchy. Portales played an important role in the compromise, and a new constitution achieved as a result (1833) remained the basis of Chilean political life until 1925. It created a strong central government, responsive to the influence of the landowning class, which controlled the parliament.

The establishment of this new political structure united the different factions that brought Ovalle and later Joaquín Prieto to power. The new government was strengthened by a successful war against the Peruvian-Bolivian Confederation (1836–39), during which it broadened its support by reinstating army officers ousted when the conservatives had seized power in 1829–30.

*Economic prosperity.* The government of Prieto and the succeeding governments of Manuel Bulnes and of Manuel Montt dedicated themselves to developing the economy. Their first and most pressing need was to reestablish the state finances, exhausted by the war. To this end, measures were taken to expand the principal source of state income—foreign trade. A free port was created at Valparaíso to encourage trade by foreign, especially British, merchants. These measures, however, would not have worked if Chilean products had not found new markets abroad. The discovery of gold in California (1848) and in Australia (1853) assured Chilean grain a vast market as the populations of those two areas expanded. The production of silver and copper increased in response to European demand, thereby increasing the wealth of the state and the dominant class. The economic development helped overcome political disagreements and aided the consolidation of internal peace.

Political stability and economic prosperity opened the way to modernization: the construction of the first railroads began, new roads were opened, and the harbours were improved. The government tried also to develop education, though largely for upper-class children. The University of Chile was founded, and foreign scholars were recruited to foster geologic, botanical, and economic studies. The development of commerce attracted numerous foreign entrepreneurs (British, French, and North American), who came to dominate the import-export trade.

Modernization

*Political diversification.* The increase of wealth that especially favoured the oligarchy and foreign merchants also contributed to a diversification of the ruling class; the development of mining production in the north and of agriculture in the south created new fortunes, whose owners soon made their entry into the political world. An attempted coup d'état, the "revolution of 1851," failed but was an indication of the political awakening of these new elements. A new development among younger members of the traditional oligarchy was the growth of liberalism and the appearance of political clubs around the middle of the century.

The impact of these forces was felt inside the political establishment, so much so that a minor conflict between the state and the church over the right to make ecclesiastical appointments was sufficient to break the unity of the dominant political class. The oligarchy was divided into two groups: conservatives, who defended the traditional privileges of the church; and nationalists, who maintained the supremacy of the state. A part of each group, dissatisfied by the authoritarian government of President Montt, united and created a separate faction, the liberals.

**The widening of liberal influence, 1861–91.** The period after 1860, known as the "Liberal Republic," saw the emergence of many rival political groups whose common characteristic—following an unsuccessful armed insurrection by radicals in 1859—was an attempt to gain power by peaceful means.

*Political factions.* After 1855 the conservative element, supporting the hegemony of the church, had allied with the liberals in opposing President Montt. The radicals joined the alliance against Montt. José Joaquín Pérez (1861–71), though elected with the support of the "nationalists," governed with the help of the liberal-conservative alliance. A division in the dominating political classes occurred about 1872, when the liberals started to draw away from the conservatives; the liberals succeeded in ending the Roman Catholic Church's monopoly in religious matters.

*European influences.* The fight to secularize the state opened the country to European influences in cultural activities and civil reforms. Young members of the economic and political oligarchy began to travel and study in Europe. They brought back many political, literary, and scientific ideas.

This new political and cultural opening toward Europe was linked to closer economic relations, especially with Great Britain, Chile's main trading partner. The British began to invest directly in Chile, supplying the capital needed to bring about the construction of railroads and the modernization of ports and public services. The increase of imports and the payment of interest from loans aggravated an already weak balance of payments and resulted in a continuing devaluation of the Chilean peso in relation to the British pound sterling.

Balance-of-payments crisis

*The War of the Pacific (1879–83).* The need to improve its balance of payments attracted Chile to saltpetre mines situated along the Chilean border in the Bolivian province of Antofagasta and in the Peruvian provinces of Tarapacá and Arica. Ill-defined borders and oppressive measures allegedly taken against the Chilean migrant population in these territories furnished Chile with a pretext for invasion. Chile defeated the Peruvian-Bolivian army and annexed these provinces.

The War of the Pacific had broad repercussions. France, Germany, and especially Britain had strong interests in the saltpetre mines, and they threatened to intervene. The United States, hoping to restrict European influence, offered to resolve the conflict by mediation; Chile refused the U.S. offer, fearing that it would have to give up its territorial gains. German support of the Chilean position further impeded European intervention.

The war weakened Chilean finances, and the economic situation continued to worsen. During the presidency of

José Manuel Balmaceda (1886–91) the government tried to claim the revenues from the saltpetre mines and thus to assert major responsibility in economic matters. Nearly all of the oligarchy, however, was looking for a weaker, rather than a stronger, central power and objected to this attempt to strengthen the executive. The clash was resolved in a brief civil war, which ended with Balmaceda's abdication of the presidency.

**Political development, 1891–1920.** The coalition that overthrew Balmaceda resulted from a large political regrouping of all those who wanted to strengthen the parliament; thus, after the civil war Chile's presidential republic was converted into a parliamentary republic. This meant that the oligarchy, which had extended itself into commerce and banking, needed only to assure itself of control of parliament—and thus of the various ministries—to dominate the political life of the country. In order to remain in office, governments now had to have the confidence of the parliament. What emerged was a continual struggle for power among the factions, which began to organize themselves as real political parties.

Organization of political parties

*Growth of the middle and lower classes.* The period between 1891 and 1920 was one of intense political activity that saw the formation of new political parties and tendencies that tried to express the political desires of the middle and lower classes. The development of a state bureaucracy and the growth of the railroads and of commerce favoured the formation of social groups with urban concerns, rarely linked to the landed oligarchy, and increasingly aware of their possible political roles.

An active working class developed in the saltpetre mines, in the large public utility enterprises (railways, gas, electricity), and in the many factories that began to appear in the urban centres, especially in Santiago. The first strikes to obtain better salaries and working conditions occurred during this period.

*Formation of new political parties.* The radical political faction—born as a dissenting wing of the liberals and striving toward the secularization of the country—became the Radical Party in 1888 and tended progressively to voice the concerns of the growing middle class.

The Democratic Party (Partido Democrático; formed 1887) was led by Malaquías Concha, who spoke for the needs of the artisans and a part of the urban workers. Founded by former radicals, this party differed from the Radical Party only in the particular emphasis it gave to the labour movement.

Marxist ideology had begun to spread among Chilean workers. The first socialist group, founded in 1897, advocated anarchism and a worker-controlled economy. It became the Socialist Party in 1901 but had a fleeting life. The increase of strikes and dissatisfaction of the miners, however, led to the formation (1912) in the mining region of a new Worker's Socialist Party (Partido Obrero Socialista), which influenced workers and university students and advocated an international class struggle; it became the Communist Party in 1922.

*Decline of the ruling class.* The radicalization of the parties of the left was caused largely by the ruling class's neglect of Chile's complex economic and social problems. The ruling class, concerned with protecting its own interests, failed to introduce needed reforms, and as a result the political instability already evident in the late 19th century grew worse. The traditional Liberal and Conservative parties were unable to adapt to the country's changing situation.

Deteriorating economy

Along with the growing political and social problems, the economic situation also worsened. Loans obtained from Britain and, after 1916, from the United States served more to pay the interest on previous debts and to cover state expenses than to allow productive investments. The country consumed more than it produced, and this was translated into an annual inflation rate of more than 10 percent and to the constant devaluation of the currency in relation to the pound sterling and the dollar. Agrarian production barely kept pace with home consumption, but the large landowners were unable to introduce techniques to increase it. Industrial development lagged because of insufficient capital.

## CHILE AFTER 1920

**Political uncertainty, 1920–38.** In the decade following World War I, falling saltpetre sales and rising inflation fueled dissatisfaction among the middle and working classes. They supported the election of the reformist president Arturo Alessandri Palma in 1920. When the legislature blocked his initiatives, discontent spread to middle-class army officers. They intervened in 1924 to force parliamentary passage of his social reforms. Alessandri resigned but the military returned him to power in 1925. In that year the army backed Alessandri's installation of a new constitution, which lasted until 1973. It established a presidential republic, separated church and state, and codified the new labour and welfare legislation.

New presidential republic

In the period between 1924 and 1932, 21 cabinets were formed and dissolved. These were years of profound crises, marked by attempts to create a new political structure by replacing the oligarchy with a new political elite. Under the military dictatorship of Carlos Ibáñez del Campo (1927–31), new economic reforms were tried: new industrial products were developed, the saltpetre mines were partially nationalized, public works were begun, and public education was improved. But these reforms did not touch the economic power of the oligarchy, which remained the principal political force.

*Effects of the world depression.* The world depression of the 1930s was difficult for Chile's economy because the international demand and the prices for saltpetre and copper plummeted. Chile was forced to reduce imports, which in turn reduced national production. Incomes diminished, while public expenditures grew.

The economic crisis, accompanied by the fall of Ibáñez, permitted the traditional political forces to regain power. They remained in office only briefly, from July 1931 to June 1932, under the presidency of Juan Esteban Montero Rodríguez, because the crisis was so strong that every attempted improvement failed. Power was then gained by a civilian-military coalition that formed the Socialist Republic (from June to September 1932), which spawned the modern Socialist Party. By the end of 1932, however, new elections returned Arturo Alessandri Palma to the presidency.

Return of Alessandri

*Return to constitutional normality.* Alessandri's second term (1932–38) was characterized by a return to constitutional normality and by the return to power of the old ruling class. Alessandri tried to restore state finances, badly weakened by the crisis. His economic measures attempted to increase mining and industrial production. Public works eased part of the existing unemployment. Social discomfort diminished, but it did not disappear.

**The Radical presidencies, 1938–52.** The return to constitutional government did not resolve Chile's serious problems. The discontent of the workers and especially of the middle class was manifested in the 1938 presidential election. The Radical candidate, Pedro Aguirre Cerda, won with the support of a coalition of the left.

*The presidencies of Aguirre Cerda and Ríos.* The period of Radical presidencies can be divided into two parts, separated by 1946. The first part included the presidencies of Aguirre Cerda (1938–41) and Juan Antonio Ríos (1942–46). Aguirre Cerda represented the middle class; his triumph came through the support of a popular front, which included the Radical, Socialist, and Communist parties and also the left-inspired Confederation of Chilean Workers.

Aguirre Cerda's program included measures for increasing industrial output. The Development Corporation (Corporación de Fomento de la Producción; Corfo) was created in 1939 to reduce imports and thus diminish the trade deficit by developing industry, mainly to produce consumer and intermediate goods.

Effects of World War II

During World War II Chile remained neutral until, in 1942, in a common action with other Latin-American countries, it declared war on Germany, Italy, and Japan. World War II and the Korean War of the early 1950s benefited Chile's economy; an increased demand for copper permitted a rise in incomes, which facilitated the expansion of public education and aided industrial development, thus helping to increase production.

*The presidency of Gabriel González Videla.* During the period from 1946 to 1952, the president was Gabriel González Videla, also of the Radical Party, who gained a plurality with the support of the Communists. The Socialist Party denounced an offer of alliance, however, and the popular front could not be reconstituted. González Videla's first cabinets, between 1946 and 1948, included Communist ministers; but the international Cold War and Chile's internal troubles soon pushed González Videla toward the right. After 1948 he outlawed the Communists and ruled with the support of the Liberal Party.

Economic links with the United States, which had grown after the economic crisis of the 1930s, were strengthened after World War II; U.S. investments in Chile increased from $414,000,000 in 1945 to $540,000,000 in 1950, largely in copper production. By 1952 the United States had loaned $342,000,000 to the Chilean government. The exchange of technicians and professors helped tighten technical and cultural links between the two countries.

Recovery of the political right

The presidency of González Videla saw the strong political recovery of the right. The Radical presidents had failed to transform Chile's economic and social situations. Between 1940 and 1952 Chile's population rose from 5,000,000 to 6,350,000, with the strongest increase in urban areas, which accounted for 52 percent of the total population in 1940 and 60 percent in 1952. Production rose during this period by a rate very close to the rise in population. But social inequities were not reduced.

**Political stagnation, 1952–64.** Various conditions explain the victory in 1952 of the former dictator Gen. Carlos Ibáñez del Campo. Under Radical rule the middle class had affirmed its political importance without injuring the economic power of the landed oligarchy, but the lower classes fell farther behind the middle and upper strata. In 1949 the vote was granted to women, and the electorate thus expanded from 631,257 in 1946 to about 1,000,000 in 1952. President Ibáñez was the candidate of a heterogeneous front based on his personal charisma, but he was not the choice of particular political parties.

Ibáñez had promised to rule with a strong hand and if necessary eliminate the parliament; but during his six years as president, he ruled with the support of the traditional right, which prevented any attempt at reform. Ibáñez retained the policy of state intervention in the economy and industrial matters inaugurated by the Radical cabinets.

*The presidency of Jorge Alessandri Rodríguez.* Ibáñez was succeeded (1958–64) by the son of Arturo Alessandri Palma, Jorge Alessandri Rodríguez, who won the support of the Conservative and Liberal parties. To satisfy popular demands without altering profoundly the structures of the country, he launched a public works program that helped absorb the masses of unemployed. At the same time, he tried to reduce the high inflation rate (about 60–70 percent yearly), to augment productivity by reducing taxes on business enterprises, and to stimulate industrial growth by expanding the home market through public expenditure.

The government placed restrictions on salary increases; salaries thus rose more slowly than prices, which continued to increase by about 30 percent yearly. This alienated the voters, and the government had to call for the support of the Radical Party.

*New political groupings.* Popular discontent helped revive the Marxist-inspired Socialist and Communist parties and produced an electoral loss of the parties of the right that corresponded with the rise of those of the left. The Christian Democratic Party, a centrist reform party founded in 1957, enjoyed the biggest increase—from 9 percent in 1957 to 15 percent in 1961. The Christian Democratic Party grew out of the Conservative Party. In 1938 a group of young conservatives had left their party to form the National Falange (Falange Nacional). In 1957 the National Falange fused with the Social Christian Party (which had also seceded from the Conservatives) to form the Christian Democratic Party, whose program tended toward serious reforms in the archaic economic and social structures. The Communist Party regained strength peacefully through an alliance with the Socialist Party, which believed that election was not the only way to power and which rejected alliances with the non-Marxist left.

The Christian Democratic Party

At the end of Alessandri Rodríguez' rule the right-wing parties were so weakened that their electoral strength was practically cut in half in the 1965 elections; in order to remain on the political scene, they joined together to form the National Party. The centrist Radical Party also lost support. A common point existed between the Christian Democratic Party and the Marxist parties—the wish to weaken the old economic and political oligarchy and to try to rescue the country from its chronic underdevelopment by more decisive action in the agrarian sectors.

**A period of change, 1964–73.** In the election of 1964 the Christian Democratic candidate, Eduardo Frei Montalva, won 56 percent of the votes. Support from the right-wing parties helped him defeat the Marxist coalition.

*The presidency of Frei Montalva.* Frei's program, synthesized in the slogan "Revolution in Liberty," promised a series of reforms for developing the country by raising the incomes of the lower classes. To attain this aim, Frei and the Christian Democrats instituted a program of "Chileanization," by which the state took control of copper, Chile's principal resource, acquiring 51 percent of the shares of the large U.S. copper companies in Chile. They thus intended to increase incomes, with which they planned to permit industries to develop; they also planned a vast agrarian reform by which to reduce the imports of agricultural products. Frei also promised decisive state intervention and reform in banking. The Frei administration, at least during its first years, counted on strong support from the middle class. But the government alienated some of the middle class by trying also to obtain the support of the peasants and of the urban underemployed, until then on the margin of the political scene.

In 1967, with the support of the Socialist and Communist parties, an agrarian reform law was approved that enabled the government to expropriate uncultivated land and to limit the land that could be conserved by each owner. Peasant cooperatives were to be established on these lands, and the state was empowered to teach the peasants better farming techniques. Agrarian reform, however, proceeded slowly because of its costly emphasis on better housing and agricultural equipment and on an irrigation system. By 1970 about 5,000,000 acres had been expropriated.

Agrarian reforms

*The socialist experiment.* The reformist program of the Frei government gave poorer people the incentive to take an active role in political life. This increase in political participation brought about further radicalization not only of the Communist and Socialist parties but also of some of the Radicals and Christian Democrats. In 1969 this cluster of parties and left-wing groups formed the Popular Unity (Unidad Popular) coalition, proposing as its presidential candidate Salvador Allende Gossens, a Socialist and an avowed Marxist; he was elected president in 1970.

The Popular Unity program envisaged the eventual transition to socialism, which was to be accomplished through the end of domination of mining and finance by foreign capital, expanded agrarian reform, and more equal distribution of income favouring the poorer classes. The accomplishments of this program were responsible for the advance of Popular Unity in the municipal elections of 1971 and in the congressional elections of 1973.

Between 1970 and 1972, however, toleration of the Popular Unity government by the middle class declined as a consequence of difficulties in the economy, which featured a complex and not always consistent reorganization resulting from the nationalization of U.S.-owned copper mines—the main resource of economic production—and of a number of heavy industries. Difficulties in maintaining production levels were further augmented by boycotts on the side of foreign capital, mainly American, and the reduction of agricultural production as a consequence of agrarian reform. Inflation and stagnation of production were propitious to the growth and regrouping of the forces that opposed the socialist experiment. The oligarchy, the right-wing National Party, and the centre Christian Democrats finally joined their efforts and supported the antigovernment trends in the armed forces.

**The military dictatorship, from 1973.** On Sept. 11, 1973, the armed forces staged a coup d'état. Allende died during an assault on the presidential palace, and a junta

composed of three generals and an admiral, with Gen. Augusto Pinochet Ugarte as president, was installed. At the outset the junta received the support of the oligarchy and of a sizable part of the middle-class and the right-wing movements. This support by moderate political forces, including many Christian Democrats, can be explained by their belief that a dictatorship represented a transitional stage necessary to restoring the status quo as it had been before 1970. Very soon they were to concede that the military officers in power had their own political objectives, including the repression of all left-wing and centre political forces. The Christian Democratic, National, and Radical Democracy parties were declared to be in "indefinite recess," and the Communists, Socialists, and Radicals were proscribed. In 1977 the traditional parties were dissolved, and a private enterprise economy was instated.

The Pinochet regime

The policies of the military government, though encouraging the development of free enterprise and a new entrepreneurial class, caused unemployment, a decline of real wages, and, as a consequence, a worsening of the standard of living of the lower and middle classes. Political and social conditions were complicated by a developing international economic crisis. In 1980 a new constitution, as well as an eight-year extension of Pinochet's presidential term, was approved by a tightly controlled plebiscite. The document included specific transitional provisions for a return to civilian constitutional government over the same eight-year period. (M.A.Ca./P.W.D.)

Pinochet's economic policies called for extensive borrowing and a return to a free market system. Economic growth did not occur as expected; the national debt increased, the price of copper on the international market dropped, and inflation and unemployment rose. Several opposition parties, the Christian Democratic Party being the largest, formed a new centre-left coalition, the Democratic Alliance (Alianza Democrática; AD). Broad participation in national days of protest organized by leftist political groups after 1983 demonstrated widespread discontent. The Roman Catholic Church also began openly to support the opposition. In August 1984, 11 parties of the right and centre signed an accord, worked out by the Archbishop of Santiago, Raúl Cardinal Silva Henríquez, calling for elections to be scheduled before 1989. Additional pressure came from the United States and other countries that had supported Chile economically but now showed signs of impatience with Pinochet's resistance to the resumption of democratic processes and with the numerous reports of human rights violations attributed to his regime.

Growing opposition to Pinochet's policies

The economic and political climate continued to be volatile in the late 1980s, with political factions exerting increasing pressure for governmental change, acts of terrorism multiplying, and the economy, though showing some signs of recovery, remaining basically unstable and precipitating strikes and protests from the labour sector. Pinochet, despite occasional concessions, showed little sign of relinquishing his control or relaxing his restrictive policies. On Oct. 5, 1988, however, in a constitutionally mandated plebiscite, voters rejected Pinochet, who had been chosen as the sole candidate for president by the junta in August. The first presidential and legislative elections since the 1973 coup were scheduled for December 1989.

In the presidential election, Christian Democrat Patricio Aylwin Azócar, leader of the Coalition of Parties for Democracy, won by a large margin over his closest opponent, Hernán Büchi Buc, a former finance minister and the government-endorsed candidate. The coalition also gained a majority in the lower chamber and nearly half the seats in the upper chamber. Aylwin, who took office in March 1990, pledged to support Chile's free-market system but also to emphasize social and political change. The 1981 constitution, however, did not allow the new government to replace Pinochet as army commander for eight years. (Ed.)

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

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

BIBLIOGRAPHY

*Physical and human geography:* General descriptive information on the land and people of Chile is available in ANDREA T. MERRILL (ed.), *Chile: A Country Study,* 2nd ed. (1982); PEDRO CUNILL GRAU, *Geografía de Chile* (1970, reissued 1978); and GEORGE MCCUTCHEN MCBRIDE, *Chile: Land and Society* (1936, reprinted 1971). An excellent collection on contemporary Chilean geography is INSTITUTO GEOGRÁFICO MILITAR (CHILE), *Geografía de Chile* (1983– ); more than 20 vol. had appeared to 1990. Good coverage of Chile is found in PRESTON E. JAMES, C.W. MINKEL, and EILEEN W. JAMES, *Latin America,* 5th ed. (1986); and HAROLD BLAKEMORE and CLIFFORD T. SMITH (eds.), *Latin America: Geographical Perspectives,* 2nd ed. (1983). Useful atlases are INSTITUTO GEOGRÁFICO MILITAR (CHILE), *Atlas geográfico de Chile* (1985), and *Atlas de la República de Chile,* 2nd ed. (1983). Statistical information may be found in the annuals CHILE. INSTITUTO NACIONAL DE ESTADÍSTICAS, *Compendio estadístico,* and *Anuario de demografía.*

The economy of Chile is described by MARKOS J. MAMALAKIS, *The Growth and Structure of the Chilean Economy: From Independence to Allende* (1976); and UNIVERSIDAD DE CHILE. INSTITUTO DE ECONOMIA, *La economía de Chile en el período 1950 & 1963,* 2 vol. (1964). Other economic issues are discussed in WORLD BANK, *Chile: An Economy in Transition,* 2 vol. (1979); CORPORACIÓN DE FOMENTO DE LA PRODUCCIÓN (CHILE), *Geografía económica de Chile: primer apendice* (1966); and *Economic Report of Chile* (annual). The neoliberalism of Pinochet is presented in EDWARD NELL (ed.), *Free Market Conservatism: A Critique of Theory and Practice* (1984).

Works on political development and ideology include FRANCISCO J. MORENO, *Legitimacy and Stability in Latin America: A Study of Chilean Political Culture* (1969); RICARDO DONOSO, *Las ideas políticas en Chile,* 3rd ed. (1975); and GERMÁN URZÚA VALENZUELA, *Los partidos políticos chilenos, los fuerzas políticas* (1968). The policies of the Pinochet regime are discussed in AUGUSTO VARAS (ed.), *Transición a la democracia: América Latina y Chile* (1984); GENARO ARRIAGADA HERRERA, *La política militar de Pinochet* (1985); and ALEJANDRO FOXLEY, *After Authoritarianism* (1985), a short working paper. See also PAUL WILLIAM GARBER and PHILIP CHARLES GARBER (trans.), *The Political Constitution of Chile,* translated from Spanish (1981).

Works on the social conditions include ARTURO VALENZUELA and J. SAMUEL VALENZUELA (eds.), *Chile: Politics and Society* (1976); BARBARA STALLINGS, *Class Conflict and Economic Development in Chile: 1958–1973* (1978); BRIAN H. SMITH, *The Church and Politics in Chile: Challenges to Modern Catholicism* (1982); and CESAR CAVIEDES, *The Politics of Chile: A Sociogeographical Assessment* (1979). The military's role in Chilean politics is treated in J. SAMUEL VALENZUELA and ARTURO VALENZUELA (eds.), *Military Rule in Chile: Dictatorship and Oppositions* (1986); CESAR CAVIEDES, *The Southern Cone: Realities of the Authoritarian State in South America* (1984); and FREDERICK M. NUNN, *The Military in Chilean History: Essays on Civil-Military Relations, 1810–1973* (1976), and *Chilean Politics, 1920–1931: The Honorable Mission of the Armed Forces* (1970). Contemporary culture is discussed in HERNÁN GODOY URZÚA, *Apuntes sobre la cultura en Chile* (1982).

*History:* Overviews are provided by BRIAN LOVEMAN, *Chile: The Legacy of Hispanic Capitalism* (1979); and ROBERT D. TALBOTT, *A History of the Chilean Boundaries* (1974). See also JORDI FUENTES *et al., Diccionario histórico de Chile,* 8th ed. (1984). The political significance of Chile's mineral resources is discussed in HAROLD BLAKEMORE, *British Nitrates and Chilean Politics, 1886–1896: Balmaceda and North* (1974). Land tenure and reform issues are analyzed in THOMAS C. WRIGHT, *Landowners and Reform in Chile: The Sociedad Nacional de Agricultura, 1919–1940* (1981); and BRIAN LOVEMAN, *Struggle in the Countryside: Politics and Rural Labor in Chile, 1919–1973* (1976). The role of labour unions is treated in ALAN ANGELL, *Politics and the Labour Movement in Chile* (1972).

Works on various periods of Chilean history include ARNOLD J. BAUER, *Chilean Rural Society from the Spanish Conquest to 1930* (1975); SIMON COLLIER, *Ideas and Politics of Chilean Independence 1808–1833* (1967); WILLIAM F. SATER, *Chile and the War of the Pacific* (1986); FREDERICK B. PIKE, *Chile and the United States, 1880–1962: The Emergence of Chile's Social Crisis and the Challenge to United States Diplomacy* (1963); PAUL W. DRAKE, *Socialism and Populism in Chile, 1932–52* (1978); ARTURO VALENZUELA, *The Breakdown of Democratic Regimes: Chile* (1978); STEFAN DE VYLDER, *Chile 1970–73: The Political Economy of the Rise and Fall of the Unidad Popular* (1974; U.S. title, *Allende's Chile,* 1976); PAUL E. SIGMUND, *The Overthrow of Allende and the Politics of Chile, 1964–1976* (1977); ROBERT J. ALEXANDER, *The Tragedy of Chile* (1978); and FEDERICO G. GIL, RICARDO LAGOS E., and HENRY A. LANDSBERGER (eds.), *Chile at the Turning Point: Lessons of the Socialist Years, 1970–1973* (1979; originally published in Spanish, 1977).

(C.N.Ca./M.A.Ca./P.W.D.)

# China

China (in full People's Republic of China, Wade–Giles romanization Chung-hua Jen-min Kung-ho-kuo, Pinyin Zhonghua Renmin Gongheguo) is the largest of all Asian countries and has the largest population of any country in the world. Occupying nearly the entire East Asian landmass, it stretches for about 3,100 miles (5,000 kilometres) from east to west and 3,400 miles from north to south and covers an area of about 3,696,100 square miles (9,572,900 square kilometres), which is approximately one-fourteenth of the land area of the Earth. Among the major countries of the world, China is surpassed in area only by the Soviet Union and Canada, and it is about twice the size of Europe, excluding the Soviet Union.

China's land frontier is about 12,400 miles in length, and its coastline extends for some 8,700 miles. The country is bounded by the Mongolian People's Republic to the north; the Soviet Union and North Korea to the northeast; the Yellow Sea and the East China Sea to the east; the South China Sea to the southeast; Vietnam, Laos, Burma, India, Bhutan, and Nepal to the south; Pakistan to the southwest; and Afghanistan and the Soviet Union to the west. In addition to the 11 countries that border directly on it, China also faces South Korea and Japan, across the Yellow Sea, and the Philippines, which lie beyond the South China Sea.

China has 30 administrative units directly under the central government; these consist of 22 provinces, five autonomous regions, and three municipalities (Peking, Shanghai, and Tientsin). The island province of Taiwan, which has been under separate administration since 1949, is discussed in the article TAIWAN. Peking, the capital of the People's Republic, is also the cultural, economic, and communications centre of the nation. Shanghai is the largest industrial and commercial city, and it is also the leading port.

Within China's boundaries exists a highly diverse and complex country. Its topography encompasses the highest and one of the lowest places on Earth, and its relief varies from nearly impenetrable mountainous terrain to vast coastal lowlands. China's climate ranges from extremely dry, desertlike conditions in the northwest to tropical monsoon in the southeast, and the country has the greatest contrast in temperature between its northern and southern borders of any nation in the world.

The diversity of China's relief and climate provides for one of the world's widest arrays of ecological niches, and these have been filled by a vast number of plant and animal species. Indeed, practically all types of Northern Hemisphere plants, except those of the polar tundra, are found in China; and despite the continuous inroads of humans over the millennia, China still is home to some of the world's most exotic animals.

Probably the single most identifiable characteristic of China to the people of the rest of the world is the size of its population. More than one-fifth of mankind is of Chinese nationality. The great majority of the population is Chinese (Han), and thus China is often characterized as a racially homogeneous country; but few countries have as wide a variety of indigenous peoples as does China. Even among the Han there are cultural and linguistic differences between regions; for example, the only point of linguistic commonality between two individuals from different parts of China may be the written Chinese language. Because China's population is so enormous, the population density of the country is also often thought to be uniformly high, but vast areas of China are either uninhabited or are sparsely populated.

With more than 4,000 years of recorded history, China is one of the few existing countries that also flourished economically and culturally in the earliest stages of world civilization. Indeed, despite the political and social upheavals that frequently have ravaged the country, China is unique among nations in its longevity and resilience as a discrete politico-cultural unit. Much of China's cultural development has been accomplished with relatively little outside influence, the introduction of Buddhism from India constituting a major exception. Even when the country was penetrated by such "barbarian" peoples as the Manchus, these groups soon became largely absorbed into the fabric of Han Chinese culture.

This relative isolation from the outside world made possible over the centuries the flowering and refinement of the Chinese culture, but it also left China ill prepared to cope with that world when, from the mid-19th century, it was confronted by technologically superior foreign nations. There followed a century of decline and decrepitude, as China found itself relatively helpless in the face of a foreign onslaught. The trauma of this external challenge became the catalyst for a revolution that began in the early 20th century against the old regime and culminated in the establishment of a Communist government in 1949. This event reshaped global political geography, and China has since come to rank among the most influential countries in the world.

Central to China's long-enduring identity as a unitary country is the province, or *sheng* ("secretariat"). The provinces are traceable in their current form to the T'ang dynasty (AD 618–907). Over the centuries, provinces gained in importance as centres of political and economic authority and increasingly became the focus of regional identification and loyalty. Provincial power reached its peak in the first two decades of the 20th century, but since the establishment of Communist rule in China this power has been curtailed by a strong central leadership in Peking. Nonetheless, while the Chinese state has remained unitary in form, the vast size and population of China's provinces—which are comparable to large and midsize nations—dictate their continuing importance as a level of subnational administration.

This article first discusses the physical and human geography of the country as a whole. This is followed by the history of China from prehistory to the present day. Finally, the 30 administrative units, organized by region, are treated individually in detail. Although the province-level municipalities of Peking, Shanghai, and Tientsin are listed in the outline below in their respective regions, the discussions of these cities appear in the articles PEKING, SHANGHAI, and TIENTSIN.

The article is divided into the following sections:


Physical and human geography 37
  The land 37
    Relief
    Drainage
    Soils
    Climate
    Plant and animal life
    Settlement patterns
  The people 50
    Ethnic and linguistic groups
    Population growth
    Population distribution
    Internal migration
  The economy 53
    Resources
    Agriculture
    Industry
    Finance
    Trade
    Administration of the economy

Transportation
Administration and social conditions 60
Government and party
Administration
Armed forces
Justice
Health and welfare
Education
Cultural life 64
History 65
Prehistory 65
Archaeology in China
Early man
Neolithic Period
The first historical dynasty: the Shang 68
The Chou and Ch'in dynasties 70
The history of the Chou (1111–255 BC)
Social, political, and cultural changes
The Ch'in Empire (221–206 BC)
The Han dynasty 75
Dynastic authority and the succession of emperors
The administration of the Han Empire
Relations with other peoples
Cultural developments
The Six Dynasties 81
Political developments
Intellectual and religious trends
The Sui dynasty 85
Wen-ti's institutional reforms
Integration of the south
Foreign affairs under Yang-ti
The T'ang dynasty 87
Early T'ang (618–626)
The period of T'ang power (626–755)
The late T'ang (755–907)
Cultural developments
Social change
The Five Dynasties and the Ten Kingdoms 95
The Wu-tai (Five Dynasties)
The Shih-kuo (Ten Kingdoms)
The "barbarians": Tangut, Khitan, and Juchen 97
The Tangut
The Khitan
The Juchen
The Sung dynasty 98
Pei (Northern) Sung (960–1127)
Nan (Southern) Sung (1127–1279)
Sung culture
The Yüan, or Mongol, dynasty 106
The Mongol conquest of China
China under the Mongols
The Ming dynasty 112
Political history
Government and administration
Foreign relations
Economic policy and developments
Culture
The Ch'ing dynasty 119
Early Ch'ing
Late Ch'ing
The republican period 131
The development of the republic (1912–20)
The interwar years (1920–37)
The war against Japan (1937–45)
Civil war (1945–49)
The People's Republic of China 143
Establishment of the People's Republic
The Cultural Revolution, 1966–76
China after the death of Mao
Northeast China 154
Heilungkiang 154
Kirin 156
Liaoning 158
North China 161
Honan 161
Hopeh 163
Peking 166
Shansi 166
Shantung 168
Tientsin 171
Lower Yangtze Valley 172
Anhwei 172
Hupeh 174
Kiangsu 177
Shanghai 179
South China 179
Chekiang 179
Chuang Autonomous Region of Kwangsi 182
Fukien 184
Hai-nan 187
Hunan 188
Kiangsi 191
Kwangtung 194
Taiwan 197
Southwest China 197
Kweichow 197
Szechwan 199
Yunnan 203
Western China 206
Tibet Autonomous Region 206
Tsinghai 212
Uighur Autonomous Region of Sinkiang 213
Northwest China 216
Hui Autonomous Region of Ningsia 216
Inner Mongolia Autonomous Region 217
Kansu 219
Shensi 222
Bibliography 225


# PHYSICAL AND HUMAN GEOGRAPHY

## The land

### RELIEF

Three relief levels

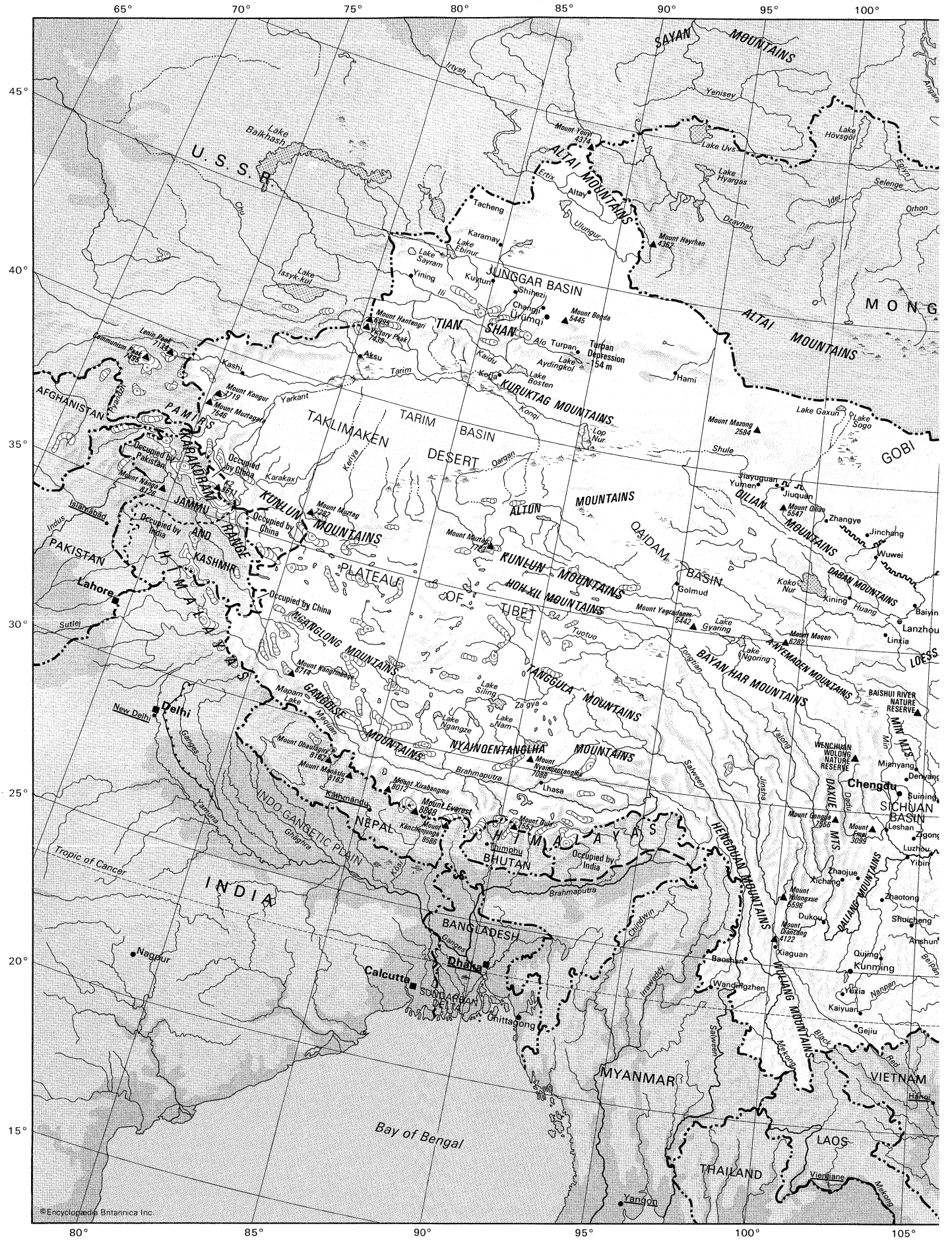
The relief of China is high in the west and low in the east; consequently, the direction of flow of the major rivers is generally eastward. The surface may be divided into three steps, or levels. The first level is represented by the Plateau of Tibet, which is located in both Tibet and Tsinghai and which, with an average height of well over 13,000 feet (4,000 metres) above sea level, is the loftiest highland area in the world. The western part of this region, the Ch'iang-t'ang, has an average height of 16,500 feet and is known as the "roof of the world."

The second step lies to the north of the Kunlun and Ch'i-lien mountains and (farther south) to the east of the Ch'iung-lai and Ta-liang mountains. Here the mountains descend sharply to heights of between 6,000 and 3,000 feet, after which basins intermingle with plateaus. This step includes the Mongolian Plateau, the Tarim Basin, the Loess Plateau (loess is a yellow-gray dust deposited by the wind), the Szechwan Basin, and the Yunnan–Kweichow highland region.

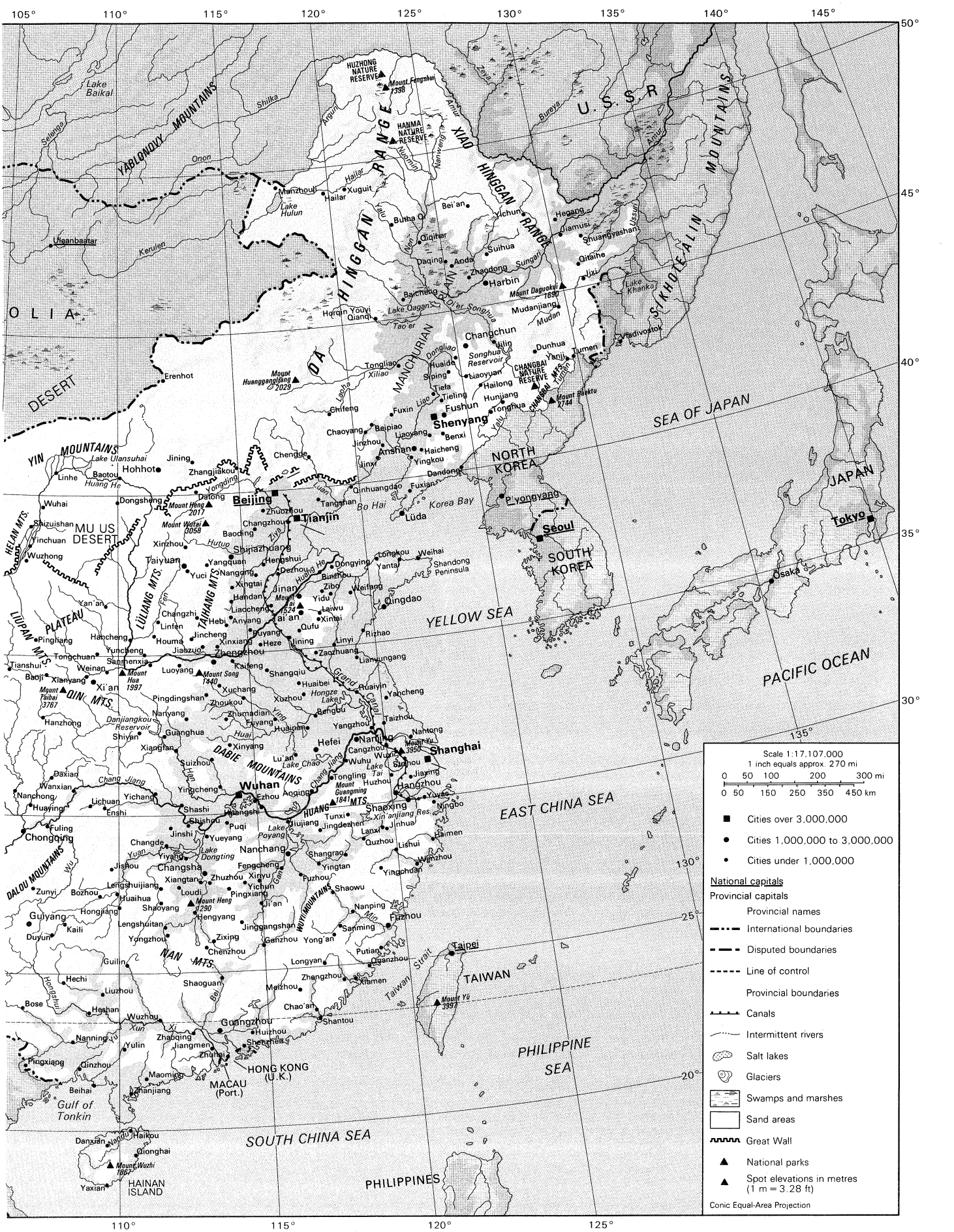
The third step extends from the east of the Ta-lou, T'ai-hang, and Wu mountain ranges and from the eastern perimeter of the Yunnan–Kweichow highland region to the China Sea. Almost all of this area is made up of hills and plains lying below 1,500 feet.

The most remarkable feature of China's relief is the vast extent of its mountain chains; the mountains, indeed, have exerted a tremendous influence on the country's political, economic, and cultural development. At a rough estimate, about one-third of the total area of China consists of mountains. China has the world's highest mountain and the world's highest and largest plateau, in addition to possessing extensive coastal plains. The five major landforms—mountain, plateau, hill, plain, and basin—are all well represented. China's complex natural environment and rich natural resources are closely connected with the varied nature of its relief.

The topography of China is marked by many splendours. Mt. Everest (Chu-mu-lang-ma Feng; 29,028 feet [8,848 metres] high), at the border between China and Nepal, is the highest peak in the world. By contrast, the lowest part of the Turfan Depression in the Uighur Autonomous Region of Sinkiang—Lake Ai-ting—is 505 feet below sea level. The coast of China contrasts greatly between South and North. To the south of the bay of Hang-chou, the coast is rocky and indented with many harbours and off-



65°
70°
75°
80°
85°
90°
95°
100°
105°
45°
40°
35°
30°
25°
20°
15°
SAYAN MOUNTAINS
Irtysh
Yenisey
Angara
Lake Balkhash
U.S.S.R.
Chu
Lake Issyk-kul
Mount Youyi 4374
ALTAI MOUNTAINS
Ertix
Altay
Ulungur
Lake Uvs
Lake Hyargas
Lake Hövsgöl
Selenge
Ider
Orhon
Dzavhan
Mount Hayrhan 4362
Tacheng
Karamay
Lake Ebinur
Lake Sayram
Yining
Ili
Kuytun
JUNGGAR BASIN
Shihezi
Changji
Ürümqi
Mount Bogda 5445
Mount Hantengri 6995
Victory Peak 7439
TIAN SHAN
Alo
Turpan
Turpan Depression -154 m
Lake Aydingkol
Lake Bosten
Kaidu
Kofla
Aksu
Tarim
Kashi
Lenin Peak 7134
Communism Peak 7495
Mount Kongur 7719
Mount Muztagata 7546
Yarkant
KURUKTAG MOUNTAINS
Konqi
Lop Nur
Hami
MONG
ALTAI MOUNTAINS
Lake Gaxun
Lake Sogo
Mount Mazong 2584
Shule
GOBI
AFGHANISTAN
Pyandzh
PAMIRS
KARAKORAM RANGE
Occupied by Pakistan
Occupied by China
Mount Nanga 8126
Karakax
Keriya
TAKLIMAKEN DESERT
TARIM BASIN
Qarqan
Islamabad
Indus
JAMMU AND KASHMIR
Occupied by India
Occupied by China
KUNLUN MOUNTAINS
Mount Muztag 7282
ALTUN MOUNTAINS
Mount Muztag 7723
QAIDAM BASIN
Jiayuguan
Yumen
Jiuquan
QILIAN MOUNTAINS
Mount Qilian 5547
Zhangye
Jinchang
Wuwei
DABAN MOUNTAINS
Koko Nur
Xining
Huang
Baiyin
Lanzhou
Linxia
LOESS
PAKISTAN
Lahore
HIMALAYAS
Sutlej
PLATEAU OF TIBET
KUNLUN MOUNTAINS
HOH XIL MOUNTAINS
Golmud
Mount Yagradagze 5442
Lake Gyaring
Lake Ngoring
Mount Maqen 6282
A NYEMAQEN MOUNTAINS
BAYAN HAR MOUNTAINS
Tongtian
Tuotuo
NGANGLONG MOUNTAINS
Occupied by China
Mount Kangrinboqe 6714
Mapam Lake
Lake Siling
Za'gya
TANGGULA MOUNTAINS
Lake Ngangze
Lake Nam
NYAINQENTANGLHA MOUNTAINS
Mount Nyainqentanglha 7088
Lhasa
Brahmaputra
Salween
Yalong
Jinsha
BAISHUI RIVER NATURE RESERVE
MIN MTS.
Min
WENCHUAN WOLONG NATURE RESERVE
Mianyang
Denyang
Chengdu
Suining
SICHUAN BASIN
DAXUE MTS.
Dadu
Mount Gongga 7556
Mount Emei 3099
Leshan
Zigong
Luzhou
Yibin
Zhaojue
Xichang
Zhaotong
Shuicheng
Anshun
DALIANG MOUNTAINS
Mount Yulongxue 5596
Dukou
Mount Diancang 4122
Xiaguan
Qujing
Kunming
Yuxia
Kaiyuan
Gejiu
Nanpan
Beipan
Black
Red
Mekong
HENGDUAN MOUNTAINS
WULIANG MOUNTAINS
Baoshan
Wandingzhen
Irrawaddy
Chindwin
Salween
MYANMAR
VIETNAM
Hanoi
LAOS
Vientiane
THAILAND
Yangon
Bay of Bengal
Delhi
New Delhi
Ganges
Yamuna
Mount Dhaulagiri 8167
Mount Manaslu 8163
Kathmandu
Mount Xixabangma 8012
Mount Everest 8848
Mount Kanchenjunga 8586
NEPAL
INDO-GANGETIC PLAIN
Ghaghra
Kosi
Mount Gula 7553
Thimphu
BHUTAN
HIMALAYAS
Occupied by India
Brahmaputra
Tropic of Cancer
INDIA
Nagpur
BANGLADESH
Ganges
Dhaka
Calcutta
SUNDARBAN DELTA
Chittagong
©Encyclopædia Britannica Inc.


105°
110°
115°
120°
125°
130°
135°
140°
145°
50°
45°
40°
35°
30°
25°
20°
130°
135°
110°
115°
120°
125°
Lake Baikal
Selenga
Onon
YABLONOVY MOUNTAINS
Shilka
Argun
HUZHONG NATURE RESERVE
Mount Fengshui 1398
HANMA NATURE RESERVE
Amur
Zeya
Bureya
U.S.S.R
Nenjiang
Nuomin
XIAO HINGGAN RANGE
Hailar
Lake Hulun
Manzhouli
Xuguit
Ulaanbaatar
Kerulen
DA HINGGAN RANGE
Bei'an
Ichun
Hegang
Jiamusi
Shuangyashan
Ussuri
SIKHOTE-ALIN MOUNTAINS
Nen
Qiqihar
Butha Qi
Suihua
Qitaihe
Lixi
Daqing
Anda
Zhaodong
Sungari
Harbin
Mount Daguoshi 1690
Lake Khanka
Baicheng
Er
Songhua
Mudanjiang
Mudan
Vladivostok
Horqin Youyi Qianqi
Lake Oagan
Tao'er
Changchun
Jilin
Dunhua
Yanji
Tumen
Tongliao
Xiliao
Liao
Huaide
Siping
Liaoyuan
Songhua Reservoir
CHANGBAI MOUNTAINS NATURE RESERVE
Tumen
Mount Paektu 2744
Erenhot
DESERT
Mount Huanggangliang 2029
Lenke
MANCHURIAN PLAIN
Tieling
Hailong
Hunjiang
Tonghua
Yalu
Fushun
Shenyang
Chifeng
Fuxin
Chaoyang
Beipiao
Liaoyang
Benxi
SEA OF JAPAN
Jinzhou
Anshan
Haicheng
Yingkou
Jinxi
NORTH KOREA
Dandong
Fuxian
YIN MOUNTAINS
Lake Ulansuhai
Jining
Zhangjiakou
Chengde
Hohhot
Linhe
Baotou
Huang He
Yongding
Juan
Qinhuangdao
Bo Hai
Korea Bay
P'yongyang
JAPAN
Tokyo
Dongsheng
Wuhai
Datong
Mount Heng 2017
Beijing
Tangshan
Lüda
Seoul
Zhuozhou
Tianjin
Mount Wutai 3058
Changzhou
Baoding
Zhi
Shizuishan
MU US DESERT
HELAN MTS.
Yinchuan
Xinzhou
Hutuo
Shijiazhuang
Wuzhong
Taiyuan
Yangquan
Hengshui
Dezhong
Dongying
Longkou
Yantai
Weihai
Shandong Peninsula
Yuci
Nangong
Hong He
Binzhou
Zibo
Weifang
Xingtai
Jinan
Handan
Mount Tai 1524
Yidu
Qingdao
Osaka
Xi
Yan'an
Changzhi
Liaocheng
Laiwu
Tai'an
Xintai
LÜLIANG MTS.
Fen
TAIHANG MTS.
YELLOW SEA
LIUPAN MTS.
PLATEAU
Hebi
Anyang
Qufu
Linfen
Puyang
Pinghiang
Hancheng
Houma
Jincheng
Xinxiang
Heze
Jining
Linyi
Rizhao
Yuncheng
Jiaozuo
Zhengzhou
Zaozhuang
Tongchuan
Tianshui
Weinan
Mount Hua 1997
Luoyang
Mount Song 1440
Kaifeng
Shangqiu
Lianyungang
Grand Canal
Huaibei
Huaiyin
Yancheng
Baoji
Xianyang
Xi'an
Mount Taibai 3767
QIN MTS.
Xuchang
Xuzhou
Hongze Lake
Pingdingshan
Zhoukou
Bengbu
PACIFIC OCEAN
Nanyang
Zhumadian
Yang
Taizhou
Danjiangkou Reservoir
Fuyang
Huainan
Yangzhou
Nantong
Hanzhong
Guanghua
Huai
Shiyan
Xiangfan
Xinyang
Hefei
Nanjing
Wuyi Mt.
Shanghai
Lu'an
Cangzhou
Lake Chao
Suizhou
DABIE MOUNTAINS
Wuhu
Suzhou
Baxian
Chang Jiang
Wanxian
Yingcheng
Wuhan
Ezhou
Anqing
Tongling
Lake Tai
Jiaxing
Nanchong
Yichang
Lichuan
Huaying
Enshi
Mount Guangming 1841
Huzhou
Hangzhou
Shaoxing
Ningbo
Yuyao
HUANG MTS.
Shashi
Huangshi
Tunxi
Xin'anjiang Res.
EAST CHINA SEA
Fuling
Chongqing
Shishou
Puqi
Jiujiang
Jingdezhen
Jinhua
Lake Poyang
Jinshi
Yueyang
Changde
Yuan
Yiyang
Lake Dongting
Nanchang
Lanxi
Quzhou
Lishui
Haimen
Wenzhou
Changsha
Jishou
Fengcheng
Shangrao
Yingtan
Yingchuan
DALOU MOUNTAINS
Wu
Xiangtan
Zhuzhou
Xinyu
Fuzhou
Lengshuijiang
Loudi
Yichun
Pingxiang
Shaowu
Zunyi
Bozhou
Huaihua
Ji'an
Mount Heng 1290
Nanping
Min
Guiyang
Hongjiang
Shaoyang
Hengyang
Jinggangshan
WUYI MOUNTAINS
Sanming
Kaili
Lengshuitan
Duyun
Yongzhou
Zixing
Ganzhou
Yong'an
Putian
Taipei
Strait
NAN MTS.
Chenzhou
Guilin
Longyan
Quanzhou
Hechi
Shaoguan
Zhangzhou
Xiamen
TAIWAN
Liuzhou
Meizhou
Taiwan
Mount Yu 3997
Bose
Hezhou
Chao'an
Shantou
Wuzhou
Guangzhou
Nanning
Yulin
Zhaoqing
Huizhou
Shenzhen
Pingxiang
Qinzhou
Jiangmen
Zhuhai
HONG KONG (U.K.)
MACAU (Port.)
Maoming
Beihai
Zhanjiang
Gulf of Tonkin
PHILIPPINE SEA
SOUTH CHINA SEA
Danxian
Haikou
Qionghai
Mount Wuzhi 1867
Yaxian
HAINAN ISLAND
PHILIPPINES
Scale 1:17,107,000
1 inch equals approx. 270 mi
0 50 100 200 300 mi
0 50 150 250 350 450 km
Cities over 3,000,000
Cities 1,000,000 to 3,000,000
Cities under 1,000,000
National capitals
Provincial capitals
Provincial names
International boundaries
Disputed boundaries
Line of control
Provincial boundaries
Canals
Intermittent rivers
Salt lakes
Glaciers
Swamps and marshes
Sand areas
Great Wall
National parks
Spot elevations in metres
(1 m = 3.28 ft)
Conic Equal-Area Projection

## MAP INDEX

Entries in this index are transcribed from Chinese in the Pinyin romanization system. This system was adopted by the Chinese government beginning in 1956 and since 1979 has been the officially prescribed system in publishing for English-speaking countries. In addition, the index contains a number of cross-references from Pinyin to certain conventional English names and spellings used in the *Britannica*. In this index and the supplement that follows it, the term "local" designates a name used locally but not administratively.

### Political subdivisions


Anhui . . . . . 32 00 N 117 00 E
Beijing Municipality . . . . 40 15 N 116 30 E
Fujian . . . . . 26 00 N 118 00 E
Gansu . . . . . 38 00 N 102 00 E
Guangdong . . . . . 23 00 N 113 00 E
Guangxi, Zhuang Autonomous Region of . . . . . 24 00 N 109 00 E
Guizhou . . . . . 27 00 N 107 00 E
Hainan . . . . . 19 00 N 109 00 E
Hebei . . . . . 39 00 N 116 00 E
Heilongjiang . . . . 48 00 N 129 00 E
Henan . . . . . 34 00 N 114 00 E
Hubei . . . . . 31 00 N 112 00 E
Hunan . . . . . 28 00 N 112 00 E
Inner Mongolia Autonomous Region . . . . . 44 00 N 112 00 E
Jiangsu . . . . . 33 00 N 120 00 E
Jiangxi . . . . . 28 00 N 116 00 E
Jilin . . . . . 43 00 N 126 00 E
Liaoning . . . . . 41 00 N 123 00 E
Ningxia, Hui Autonomous Region of . . . . . 37 00 N 106 00 E
Qinghai . . . . . 36 00 N 96 00 E
Shaanxi . . . . . 36 00 N 109 00 E
Shandong . . . . . 36 00 N 119 00 E
Shanghai Municipality . . . . 31 14 N 121 28 E
Shanxi . . . . . 37 00 N 112 00 E
Sichuan . . . . . 30 00 N 103 00 E
Tianjin Municipality . . . . 39 08 N 117 12 E
Tibet Autonomous Region . . 32 00 N 90 00 E
Xinjiang, Uygur Autonomous Region of . . . . . 42 00 N 86 00 E
Yunnan . . . . . 25 00 N 102 00 E
Zhejiang . . . . . 29 00 N 120 00 E


### Cities and towns


Aksu . . . . . 41 09 N 80 15 E
Altay . . . . . 47 52 N 88 07 E
Anda (locally Sartu) . . . . . 46 24 N 125 19 E
Anqing . . . . . 30 31 N 117 02 E
Anshan . . . . . 41 07 N 122 57 E
Anshun . . . . . 26 15 N 105 56 E
Anyang (locally Zhangde) . . . . . 36 05 N 114 21 E
Baicheng . . . . . 45 37 N 122 49 E
Baiyin . . . . . 36 32 N 104 12 E
Baoding . . . . . 38 52 N 115 29 E
Baoji . . . . . 34 23 N 107 09 E
Baoshan . . . . . 25 07 N 99 09 E
Baotou . . . . . 40 36 N 109 59 E
Bei'an . . . . . 48 16 N 126 36 E
Beihai . . . . . 21 29 N 109 06 E
Beijing . . . . . 39 56 N 116 24 E
Beipiao . . . . . 41 48 N 120 44 E
Bengbu . . . . . 32 57 N 117 20 E
Benxi . . . . . 41 20 N 123 45 E
Binzhou . . . . . 37 27 N 118 20 E
Bose . . . . . 23 54 N 106 37 E
Bozhou . . . . . 27 23 N 109 18 E
Butha Qi (locally Zalantun) . . . . . 48 00 N 123 43 E
Cangzhou . . . . . 38 19 N 116 52 E
Changchun . . . . . 43 52 N 125 21 E
Changde . . . . . 29 02 N 111 14 E
Changji . . . . . 44 01 N 87 19 E
Changsha . . . . . 28 12 N 112 58 E
Changzhi . . . . . 36 11 N 113 06 E
Changzhou . . . . . 31 47 N 119 58 E
Chao'an (locally Chaozhou) . . . . 23 41 N 116 38 E
Chaoyang . . . . . 41 33 N 122 25 E
Chengde . . . . . 40 58 N 117 53 E
Chengdu . . . . . 30 40 N 104 04 E
Chenzhou . . . . . 25 48 N 113 02 E
Chifeng . . . . . 42 17 N 118 53 E
Chongqing (locally Yuzhou) . 29 34 N 106 35 E
Dandong . . . . . 40 08 N 124 24 E
Danxian (locally Nada) . . . . . 19 31 N 109 33 E
Daqing . . . . . 46 36 N 125 00 E
Datong . . . . . 40 05 N 113 18 E
Daxian . . . . . 31 16 N 107 31 E
Denyang . . . . . 31 08 N 104 24 E
Dezhou . . . . . 37 27 N 116 18 E
Dongsheng . . . . . 39 49 N 109 59 E
Dongying . . . . . 37 30 N 118 31 E
Dukou . . . . . 26 33 N 101 44 E
Dunhua . . . . . 43 21 N 128 13 E
Duyun . . . . . 26 16 N 107 31 E
Enshi . . . . . 30 18 N 109 29 E
Erenhot . . . . . 43 45 N 112 02 E
Ezhou . . . . . 30 24 N 114 50 E
Fengcheng . . . . . 28 12 N 115 46 E
Fuling . . . . . 29 43 N 107 24 E
Fushun . . . . . 41 52 N 123 53 E
Fuxian (locally Wafangdian) . . . 39 38 N 122 00 E
Fuxin . . . . . 42 06 N 121 46 E
Fuyang . . . . . 32 54 N 115 49 E
Fuzhou . . . . . 26 05 N 119 18 E
Fuzhou . . . . . 28 01 N 116 20 E
Ganzhou . . . . . 25 51 N 114 56 E
Gejiu . . . . . 23 23 N 103 09 E
Golmud . . . . . 36 22 N 94 55 E
Guanghua (locally Laohekou) . . . . . 32 22 N 111 40 E
Guangzhou . . . . . 23 07 N 113 15 E
Guilin . . . . . 25 17 N 110 17 E
Guiyang . . . . . 26 35 N 106 43 E
Haicheng . . . . . 40 52 N 123 00 E
Haikou . . . . . 20 03 N 110 19 E
Hailar . . . . . 49 12 N 119 42 E
Hailong (locally Meichekou) . . . . 42 32 N 125 38 E
Haimen . . . . . 28 41 N 121 27 E
Hami . . . . . 42 48 N 93 27 E
Hancheng . . . . . 35 28 N 110 29 E
Handan . . . . . 36 35 N 114 29 E
Hangzhou . . . . . 30 15 N 120 10 E
Hanzhong . . . . . 33 08 N 107 02 E
Harbin . . . . . 45 45 N 126 39 E
Hebi . . . . . 35 57 N 114 13 E
Hechi (locally Jinchengjang) . . 24 42 N 108 02 E
Hefei . . . . . 31 51 N 117 17 E
Hegang . . . . . 47 24 N 130 22 E
Hengshui . . . . . 37 43 N 115 42 E
Hengyang . . . . . 26 54 N 112 36 E
Heshan . . . . . 23 42 N 108 48 E
Heze (locally Caozhou) . . . . . 35 14 N 115 27 E
Hohhot . . . . . 40 47 N 111 37 E
Hongjiang . . . . . 27 07 N 109 56 E
Horqin Youyi Qianqi (locally Ulan Hot) . . . . . 46 05 N 122 05 E
Houma . . . . . 35 36 N 111 21 E
Huaibei . . . . . 33 57 N 116 45 E
Huaide (locally Gongzhuling) . . . 43 30 N 124 49 E
Huaihua (locally Yushuwan) . . . . 27 33 N 109 57 E
Huainan . . . . . 32 40 N 117 00 E
Huaiyin (locally Wangying) . . . . . 33 35 N 119 02 E
Huangshi . . . . . 30 13 N 115 06 E
Huaying . . . . . 30 14 N 106 40 E
Huizhou . . . . . 23 05 N 114 24 E
Hunjiang (locally Badaojiang) . . . . 41 54 N 126 26 E
Huzhou (locally Wuxing) . . . . . 30 52 N 120 06 E
Jiamusi . . . . . 46 50 N 130 21 E
Ji'an . . . . . 27 08 N 115 00 E
Jiangmen . . . . . 22 35 N 113 05 E
Jiaozuo . . . . . 35 15 N 113 13 E
Jiaxing . . . . . 30 46 N 120 45 E
Jiayuguan . . . . . 39 49 N 98 18 E
Jilin . . . . . 43 51 N 126 33 E
Jinan . . . . . 36 40 N 117 00 E
Jinchang . . . . . 38 24 N 102 06 E
Jincheng . . . . . 35 30 N 112 50 E
Jingdezhen . . . . . 29 16 N 117 11 E
Jinggangshan . . . 26 37 N 114 05 E
Jinhua . . . . . 29 07 N 119 39 E
Jining . . . . . 35 24 N 116 33 E
Jining . . . . . 40 57 N 113 02 E
Jinshi . . . . . 29 40 N 111 45 E
Jinxi . . . . . 40 45 N 120 50 E
Jinzhou . . . . . 41 07 N 121 06 E
Jishou . . . . . 28 19 N 109 43 E
Jiujiang . . . . . 29 44 N 115 59 E
Jiuquan (locally Suzhou) . . . . . 39 46 N 98 34 E
Jixi . . . . . 45 18 N 130 58 E
Kaifeng . . . . . 34 51 N 114 21 E
Kaili . . . . . 26 35 N 107 55 E
Kaiyuan . . . . . 23 42 N 103 14 E
Karamay . . . . . 45 30 N 84 55 E
Kashi . . . . . 39 29 N 75 58 E
Korla . . . . . 41 44 N 86 09 E
Kunming . . . . . 25 04 N 102 41 E
Kuytun . . . . . 44 25 N 85 00 E
Laiwu . . . . . 36 41 N 118 28 E
Lanxi . . . . . 29 13 N 119 28 E
Lanzhou . . . . . 36 03 N 103 41 E
Lengshuijiang . . . 27 41 N 111 25 E
Lengshuitan . . . . 26 27 N 111 35 E
Leshan . . . . . 29 34 N 103 44 E
Lhasa . . . . . 29 39 N 91 06 E
Lianyungang (locally Xinpu) . . 34 36 N 119 13 E
Liaocheng . . . . . 36 26 N 115 58 E
Liaoyang . . . . . 41 17 N 123 11 E
Liaoyuan . . . . . 42 55 N 125 09 E
Lichuan . . . . . 30 18 N 108 51 E
Linfen . . . . . 36 05 N 111 31 E
Linhe . . . . . 40 50 N 107 30 E
Linxia . . . . . 35 28 N 102 55 E
Linyi . . . . . 37 05 N 118 20 E
Lishui . . . . . 28 27 N 119 54 E
Liuzhou . . . . . 24 19 N 109 24 E
Longkou . . . . . 37 38 N 120 18 E
Longyan . . . . . 25 11 N 117 00 E
Loudi . . . . . 27 45 N 111 59 E
Lu'an . . . . . 31 45 N 116 29 E
Lüda (locally Dalian) . . . . . 38 55 N 121 39 E
Luoyang . . . . . 34 41 N 112 28 E
Luzhou . . . . . 28 53 N 105 23 E
Manzhouli . . . . . 49 36 N 117 26 E
Maoming . . . . . 21 39 N 110 54 E
Meizhou . . . . . 23 51 N 117 18 E
Mianyang . . . . . 31 28 N 104 46 E
Mudanjiang . . . . . 44 35 N 129 36 E
Nanchang . . . . . 28 41 N 115 53 E
Nanchong . . . . . 30 48 N 106 04 E
Nangong . . . . . 37 22 N 115 22 E
Nanjing . . . . . 32 03 N 118 47 E
Nanning . . . . . 22 49 N 108 19 E
Nanping . . . . . 26 38 N 118 10 E
Nantong . . . . . 32 02 N 120 53 E
Nanyang . . . . . 33 00 N 112 32 E
Ningbo . . . . . 29 53 N 121 33 E
Pingdingshan . . . 33 44 N 113 18 E
Pingliang . . . . . 35 32 N 106 41 E
Pingxiang . . . . . 22 06 N 106 44 E
Pingxiang . . . . . 27 37 N 113 51 E
Puqi . . . . . 29 43 N 113 53 E
Putian . . . . . 25 26 N 119 01 E
Puyang . . . . . 35 42 N 114 59 E
Qingdao . . . . . 36 04 N 120 19 E
Qinhuangdao . . . 39 56 N 119 37 E
Qinzhou . . . . . 21 57 N 108 37 E
Qionghai (locally Jiaji) . . . . . 19 15 N 110 26 E
Qiqihar . . . . . 47 22 N 123 57 E
Qitaihe . . . . . 45 48 N 130 53 E
Quanzhou . . . . . 24 54 N 118 35 E
Qufu . . . . . 35 36 N 116 59 E
Qujing . . . . . 25 36 N 103 49 E
Quzhou . . . . . 28 54 N 118 48 E
Rizhao . . . . . 35 26 N 119 27 E
Sanmenxia . . . . . 34 50 N 111 05 E
Sanming . . . . . 26 14 N 117 35 E
Shanghai . . . . . 31 14 N 121 28 E
Shangqiu (locally Zhuji) . . . . . 34 27 N 115 39 E
Shangrao . . . . . 28 26 N 117 58 E
Shantou . . . . . 23 22 N 116 40 E
Shaoguan . . . . . 24 48 N 113 35 E
Shaowu . . . . . 27 18 N 117 30 E
Shaoxing . . . . . 30 00 N 120 35 E
Shaoyang (locally Baoqing) . . . . . 27 15 N 111 28 E
Shashi . . . . . 30 19 N 112 14 E
Shenyang . . . . . 41 48 N 123 27 E
Shenzhen . . . . . 22 32 N 114 08 E
Shihezi . . . . . 44 18 N 86 02 E
Shijiazhuang . . . . 38 03 N 114 29 E
Shishou . . . . . 29 43 N 112 24 E
Shiyan . . . . . 32 34 N 110 47 E
Shizuishan . . . . . 39 10 N 106 45 E
Shuangyashan . . 46 40 N 131 21 E
Shuicheng . . . . . 26 36 N 104 51 E
Siping . . . . . 43 10 N 124 20 E
Suihua . . . . . 46 39 N 126 59 E
Suining . . . . . 30 32 N 105 32 E
Suizhou . . . . . 31 36 N 113 03 E
Suzhou . . . . . 31 18 N 120 37 E
Tacheng . . . . . 46 45 N 82 57 E
Tai'an . . . . . 36 12 N 117 07 E
Taiyuan . . . . . 37 52 N 112 33 E
Taizhou . . . . . 32 29 N 119 55 E
Tangshan . . . . . 39 38 N 118 11 E
Tianjin . . . . . 39 08 N 117 12 E
Tianshui . . . . . 34 35 N 105 43 E
Tiefa . . . . . 42 30 N 123 25 E
Tieling . . . . . 42 18 N 123 49 E
Tongchuan . . . . . 35 05 N 109 05 E
Tonghua . . . . . 41 41 N 125 55 E
Tongliao . . . . . 43 37 N 122 16 E
Tongling . . . . . 30 56 N 117 50 E
Tumen . . . . . 42 58 N 129 49 E
Tunxi . . . . . 29 43 N 118 19 E
Turpan . . . . . 42 56 N 89 10 E
Ürümqi . . . . . 43 48 N 87 35 E
Wandingzhen . . . 24 05 N 98 04 E
Wanxian . . . . . 30 49 N 108 24 E
Weifang . . . . . 36 43 N 119 06 E
Weihai . . . . . 37 30 N 122 06 E
Weinan . . . . . 34 30 N 109 30 E
Wenzhou . . . . . 28 01 N 120 39 E
Wuhai . . . . . 39 47 N 106 52 E
Wuhan . . . . . 30 35 N 114 16 E
Wuhu . . . . . 31 21 N 118 22 E
Wuwei (locally Liangzhou) . . . . 37 58 N 102 48 E
Wuxi . . . . . 31 35 N 120 18 E
Wuzhong . . . . . 38 00 N 106 10 E
Wuzhou . . . . . 23 29 N 111 19 E
Xiaguan . . . . . 25 34 N 100 14 E
Xiamen . . . . . 24 27 N 118 05 E
Xi'an . . . . . 34 16 N 108 54 E
Xiangfan . . . . . 32 03 N 112 05 E
Xiangtan . . . . . 27 51 N 112 54 E
Xianyang . . . . . 34 22 N 108 42 E
Xichang . . . . . 27 53 N 102 18 E
Xingtai . . . . . 37 03 N 114 30 E
Xining . . . . . 36 37 N 101 46 E
Xintai . . . . . 35 54 N 117 44 E
Xinxiang . . . . . 35 19 N 113 52 E
Xinyang (locally Pingqiao) . . . . . 32 03 N 114 05 E
Xinyu . . . . . 27 48 N 114 56 E
Xinzhou . . . . . 38 24 N 112 44 E
Xuchang . . . . . 34 01 N 113 49 E
Xuguit (locally Yakeshit) . . . . . 49 17 N 120 44 E
Xuzhou . . . . . 34 16 N 117 11 E
Yan'an . . . . . 36 36 N 109 28 E
Yancheng . . . . . 33 23 N 120 08 E
Yangquan . . . . . 37 54 N 113 36 E
Yangzhou . . . . . 32 24 N 119 26 E
Yanji . . . . . 42 53 N 129 31 E
Yantai . . . . . 37 32 N 121 24 E
Yaxian (locally Sanya) . . . . . 18 14 N 109 29 E
Yibin . . . . . 28 46 N 104 34 E
Yichang . . . . . 30 42 N 111 17 E
Yichun . . . . . 27 50 N 114 24 E
Yichun . . . . . 47 42 N 128 54 E
Yidu . . . . . 36 41 N 118 28 E
Yinchuan . . . . . 38 28 N 106 19 E
Yingcheng . . . . . 30 57 N 113 33 E
Yingchuan . . . . . 27 51 N 119 22 E
Yingkou . . . . . 40 40 N 122 17 E
Yingtan . . . . . 28 14 N 117 00 E
Yining . . . . . 43 54 N 81 21 E
Yiyang . . . . . 28 36 N 112 20 E
Yong'an . . . . . 25 58 N 117 22 E
Yongzhou . . . . . 26 14 N 111 37 E
Yuci . . . . . 37 42 N 112 44 E

Yueyang . . . . . 29 23 N 113 06 E
Yulin . . . . . 22 38 N 110 09 E
Yumen (locally Laojunmiao) . . . . 39 50 N 97 44 E
Yuncheng . . . . . 35 01 N 110 59 E
Yuxia . . . . . 24 28 N 104 20 E
Yuyao . . . . . 30 03 N 121 09 E
Zaozhuang . . . . . 34 53 N 117 34 E
Zhangjiakou . . . . 40 50 N 114 56 E
Zhangye . . . . . 38 56 N 100 27 E
Zhangzhou . . . . . 24 31 N 117 40 E
Zhanjiang . . . . . 21 12 N 110 23 E
Zhaodong . . . . . 46 05 N 125 59 E
Zhaojue . . . . . 28 02 N 102 52 E
Zhaoqing . . . . . 23 03 N 112 27 E
Zhaotong . . . . . 27 19 N 103 43 E
Zhengzhou . . . . . 34 45 N 113 40 E
Zhoukou . . . . . 33 38 N 114 38 E
Zhuhai (locally Xiangzhou) . . . . 22 17 N 113 34 E
Zhumadian . . . . . 32 58 N 114 03 E
Zhuozhou . . . . . 39 30 N 115 58 E
Zhuzhou . . . . . 27 50 N 113 09 E
Zibo (locally Zhangdian) . . . . 36 48 N 118 03 E
Zigong . . . . . 29 24 N 104 47 E
Zixing . . . . . 25 58 N 113 24 E
Zunyi . . . . . 27 42 N 106 55 E


**Physical features and points of interest**


Alo, *river* . . . . . 42 42 N 89 12 E
Altai Mountains . . . 48 00 N 90 00 E
Altun Mountains . . 38 00 N 88 00 E
Amur (Heilong), *river* . . . . . 48 26 N 135 00 E
A'nyemaqen Mountains . . . . 34 30 N 100 00 E
Argun (Ergun), *river* . . . . . 53 20 N 121 28 E
Aydingkol, Lake . . 42 40 N 89 15 E
Baishui River Nature Reserve . 32 36 N 104 45 E
Bayan Har Mountains . . . . . 34 20 N 97 00 E
Bei, *river* . . . . . 23 02 N 112 58 E
Beibu, see Tonkin, Gulf of
Beipan, *river* . . . . 25 07 N 106 01 E
Black (Lixian), *river* . . . . . 23 30 N 101 00 E
Bo Hai . . . . . 38 30 N 120 00 E
Bogda, Mount . . . 43 45 N 88 32 E
Bosten, Lake . . . . 42 00 N 87 00 E
Brahmaputra (Yarlung), *river* . 29 25 N 90 00 E
Chang Jiang, *river* . . . . . 31 48 N 121 10 E
Changbai Mountains . . . . . 41 40 N 128 00 E
Changbai Nature Reserve . . . . . 42 00 N 128 00 E
Chao, Lake . . . . . 31 31 N 117 33 E
Da Hinggan Range . . . . . 49 00 N 122 00 E
Da Yunhe, see Grand Canal
Daban Mountains . . . . 37 00 N 102 10 E
Dabie Mountains . 31 15 N 115 00 E
Dadu, *river* . . . . . 29 32 N 103 44 E
Daguokui, Mount . 45 19 N 129 50 E
Daliang Mountains . . . . . 28 00 N 103 00 E
Dalou Mountains . 28 00 N 106 40 E
Danjiangkou Reservoir . . . . . 32 37 N 111 30 E
Daxue Mountains . . . . 30 30 N 101 30 E
Diancang, Mount . 25 42 N 100 02 E
Di'er Songhua, *river* . . . . . 45 26 N 124 39 E
Dong, see East China Sea
Dongbei, see Manchurian Plain
Dongliao, *river* . . . 43 24 N 123 42 E
Dongting, Lake . . 29 18 N 112 45 E
East China (Dong) Sea . . . . 29 00 N 125 00 E
Ebinur, Lake . . . . 44 55 N 82 55 E
Emei, Mount . . . . 29 32 N 103 21 E
Ergun, see Argun
Ertix, *river* . . . . . 47 52 N 84 16 E
Everest (Qomolangma), Mount . . . . . 27 59 N 86 56 E
Fen, *river* . . . . . 35 36 N 110 42 E
Fengshui, Mount . 52 25 N 123 21 E
Gan, *river* . . . . . 29 12 N 116 00 E
Gangdise Mountains . . . . . 31 00 N 82 00 E
Gaxun, Lake . . . . 42 25 N 101 00 E
Gobi Desert . . . . 43 00 N 105 00 E
Godwin-Austen, see K2
Gongga, Mount . . 29 34 N 101 53 E
Grand Canal (Da Yunhe) . . . . . 39 54 N 116 44 E
Guangming, Mount . . . . . 30 09 N 118 11 E
Gula (Kula, Bhutan), Mount . . . . . 28 14 N 90 36 E
Gyaring, Lake . . . 34 52 N 97 30 E
Hailar, *river* . . . . . 49 30 N 117 50 E
Hainan Island . . . 19 00 N 109 00 E
Han, *river* . . . . . 30 34 N 114 17 E
Hanma Nature Reserve . . . . . 49 50 N 123 30 E
Hantengri, Mount . . . . . 42 12 N 80 15 E
Heilong, see Amur
Helan Mountains . 39 00 N 106 00 E
Heng, Mount . . . . 27 18 N 112 41 E
Heng, Mount . . . . 39 42 N 113 45 E
Hengduan Mountains . . . . . 27 30 N 99 00 E
Herlen, see Kerulen
Himalayas, *mountains* . . . . . 29 00 N 83 00 E
Hoh Xil Mountains . . . . . 35 20 N 91 00 E
Hongshui, *river* . . 23 47 N 109 33 E
Hongze Lake . . . . 33 18 N 118 41 E
Hua, Mount . . . . . 34 27 N 110 05 E
Huai, *river* . . . . . 33 12 N 115 33 E
Huang, *river* . . . . 36 05 N 103 20 E
Huang, see Yellow Sea
Huang He, *river* . . 37 32 N 118 19 E
Huang Mountains . . . . 30 10 N 118 07 E
Huanggangliang, Mount . . . . . 43 33 N 117 32 E
Huangtu, see Loess Plateau
Hulun, Lake . . . . 49 00 N 117 27 E
Hutuo, *river* . . . . . 38 14 N 116 05 E
Huzhong Nature Reserve . . . . . 52 05 N 123 40 E
Ili, *river* . . . . . 45 24 N 74 08 E
Jinsha, *river* . . . . 28 46 N 104 38 E
Junggar Basin . . . 45 00 N 88 00 E
K2 (Qogir, Mount; Godwin-Austen), *mountain* . . . . . 35 53 N 76 30 E
Kaidu, *river* . . . . . 41 55 N 86 38 E
Kangrinboqe, Mount . . . . . 31 04 N 81 19 E
Karakax, *river* . . . 38 06 N 80 24 E
Karakoram Range . . . . . 36 00 N 76 00 E
Keriya, *river* . . . . . 38 30 N 82 10 E
Kerulen (Herlen), *river* . . . . . 48 48 N 117 00 E
Khanka (Xingkai), Lake . . . . . 45 00 N 132 24 E
Koko Nur (Qinghai), *lake* . 37 00 N 100 20 E
Kongur, Mount . . 38 40 N 75 21 E
Konqi, *river* . . . . . 41 48 N 86 47 E
Kula, Bhutan, see Gula, Mount
Kunlun Mountains . . . . . 36 00 N 84 00 E
Kuruktag Mountains . . . . . 41 30 N 90 00 E
Lancang, see Mekong
Laoha, *river* . . . . . 43 24 N 120 39 E
Liao, *river* . . . . . 40 39 N 122 12 E
Liupan Mountains . . . . . 35 40 N 106 15 E
Lixian, see Black
Loess (Huangtu) Plateau . . . . . 37 00 N 108 00 E
Lop Nur, *lake* . . . 40 30 N 90 30 E
Luan, *river* . . . . . 39 20 N 119 10 E
Lüliang Mountains . . . . . 37 45 N 111 25 E
Manchurian (Dongbei) Plain . 44 00 N 124 00 E
Mapam Lake . . . . 30 40 N 81 25 E
Maqen, Mount . . . 34 55 N 99 18 E
Maquan, *river* . . . 29 36 N 84 09 E
Mazong, Mount . . 41 33 N 97 10 E
Mekong (Lancang), *river* . 23 30 N 100 15 E
Min, *river* . . . . . 26 05 N 119 32 E
Min, *river* . . . . . 28 46 N 104 38 E
Min Mountains . . . 33 35 N 103 00 E
Mu Us Desert . . . 38 45 N 109 10 E
Mudan, *river* . . . . 46 18 N 129 31 E
Muztag, Mount . . 35 55 N 80 20 E
Muztag, Mount . . 36 25 N 87 25 E
Muztagata, Mount . . . . . 38 17 N 75 07 E
Nam, Lake . . . . . 30 42 N 90 35 E
Nan, see South China Sea
Nan Mountains . . 25 00 N 112 00 E
Nandu, *river* . . . . 20 04 N 110 22 E
Nanpan, *river* . . . 24 56 N 106 12 E
Nanweng, *river* . . 51 10 N 125 59 E
Nen, *river* . . . . . 45 26 N 124 39 E
Nganglong Mountains . . . . . 32 00 N 83 00 E
Ngangze, Lake . . 31 01 N 86 56 E
Ngoring, Lake . . 34 55 N 98 00 E
Nu, see Salween
Nuomin, *river* . . . . 48 21 N 124 32 E
Nyainqentanglha, Mount . . . . . 30 12 N 90 33 E
Nyainqentanglha Mountains . . . . 30 10 N 90 00 E
Pamirs, *region* . . . 38 00 N 73 00 E
Pobedy, see Victory Peak
Poyang, Lake . . . 29 00 N 116 25 E
Qagan, Lake . . . . 45 14 N 124 17 E
Qaidam Basin . . . 37 00 N 93 00 E
Qarqan, *river* . . . . 39 30 N 88 15 E
Qilian, Mount . . . . 39 12 N 98 35 E
Qilian Mountains . 38 30 N 100 00 E
Qin Mountains . . . 34 00 N 108 00 E
Qing Zang, see Tibet, Plateau of
Qinghai, see Koko Nur
Qogir, Mount, see K2
Qomolangma, see Everest, Mount
Red (Yuan), *river* . 23 17 N 104 20 E
Salween (Nu), *river* . . . . . 25 00 N 98 52 E
Sayram, Lake . . . 44 36 N 81 07 E
Shandong Peninsula . . . . . 37 00 N 121 00 E
Shengli, see Victory Peak
Shule, *river* . . . . . 40 20 N 92 50 E
Sichuan Basin . . . 30 00 N 105 00 E
Siling, Lake . . . . . 31 50 N 89 00 E
Sogo, Lake . . . . . 42 20 N 101 20 E
Song, Mount . . . . 34 31 N 113 00 E
Songhua Reservoir . . . . . 43 30 N 126 51 E
South China (Nan) Sea . . . . . 20 00 N 115 00 E
Sungari (Songhua), *river* . . . . . 47 42 N 132 30 E
Tabyn-Bogdo, see Youyi, Mount
Tai, Lake . . . . . 31 15 N 120 10 E
Tai, Mount . . . . . 36 30 N 117 20 E
Taibai, Mount . . . 33 57 N 107 40 E
Taihang Mountains . . . . . 37 00 N 114 00 E
Taiwan Strait . . . . 24 00 N 119 00 E
Taklimaken Desert . . . . . 39 00 N 83 00 E
Tanggula Mountains . . . . . 33 00 N 92 00 E
Tao'er, *river* . . . . . 45 42 N 124 05 E
Tarim, *river* . . . . . 41 05 N 86 40 E
Tarim Basin . . . . . 41 00 N 84 00 E
Tavan Bogd, see Youyi, Mount
Tian Shan, *mountains* . . . . 42 00 N 80 00 E
Tibet (Qing Zang), Plateau of . . . . 33 00 N 92 00 E
Tongtian, *river* . . . 33 26 N 96 36 E
Tonkin (Beibu), Gulf of . . . . . 20 00 N 108 00 E
Tumen, *river* . . . . 42 18 N 130 41 E
Tuotuo, *river* . . . . 34 03 N 93 06 E
Turpan Depression . . . . 42 45 N 89 00 E
Ulansuhai, Lake . 40 56 N 108 49 E
Ulungur, *river* . . . 46 58 N 87 28 E
Ussuri (Wusuli), *river* . . . . . 48 28 N 135 02 E
Victory (Pobedy, U.S.S.R.; Shengli, China) Peak . . . . . 42 03 N 80 11 E
Wenchuan Wolong Nature Reserve . . . . . 31 51 N 102 50 E
Wu, *river* . . . . . 29 43 N 107 24 E
Wuliang Mountains . . . . . 24 00 N 101 00 E
Wusuli, see Ussuri
Wutai, Mount . . . 39 04 N 113 28 E
Wuyi Mountains . 27 00 N 117 00 E
Wuzhi, Mount . . . 18 54 N 109 40 E
Xi, *river* . . . . . 23 05 N 114 23 E
Xiao Hinggan Range . . . . . 48 45 N 127 00 E
Xiliao, *river* . . . . . 43 24 N 123 42 E
Xin'anjiang Reservoir . . . . . 29 27 N 119 06 E
Xingkai, see Khanka, Lake
Xixabangma, Mount . . . . . 28 21 N 85 47 E
Xun, *river* . . . . . 23 28 N 111 18 E
Yagradagze, Mount . . . . . 35 09 N 95 39 E
Yalong, *river* . . . . 26 37 N 101 48 E
Yalu, *river* . . . . . 39 55 N 124 20 E
Yalu, *river* . . . . . 46 56 N 123 30 E
Yarkant, *river* . . . 40 28 N 80 52 E
Yarlung, see Brahmaputra
Yellow (Huang) Sea . . . . . 36 00 N 124 00 E
Yin Mountains . . . 41 30 N 109 00 E
Ying, *river* . . . . . 32 30 N 116 31 E
Yongding, *river* . . 39 20 N 117 04 E
Youyi (Tabyn-Bogdo, U.S.S.R.; Tavan Bogd, Mongolia), Mount . . . . . 49 08 N 87 45 E
Yu, *river* . . . . . 22 49 N 108 25 E
Yu, Mount . . . . . 31 40 N 120 42 E
Yuan, *river* . . . . . 29 35 N 112 58 E
Yuan, see Red
Yulongxue, Mount . . . . . 27 09 N 100 12 E
Za'gya, *river* . . . . 31 55 N 88 58 E
Ziya, *river* . . . . . 39 09 N 117 10 E

**INDEX SUPPLEMENT**

This index supplement shows Pinyin equivalents for geographic names rendered in conventional spelling or in the Wade-Giles romanization system. Most English-language reference works, including the *Macropædia* and the *Micropædia,* use conventional spellings for Chinese political subdivisions and a combination of conventional and Wade-Giles spellings for cities and towns as well as physical features and points of interest. The supplement thus serves as a bridge between this standardized usage in text and the map.

| Conventional/Wade-Giles | Pinyin |
|---|---|
| **Political subdivisions** | |
| Anhwei (An-hui) | Anhui |
| Chekiang (Che-chiang) | Zhejiang |
| Fukien (Fu-chien) | Fujian |
| Hai-nan | Hainan |
| Heilungkiang (Hei-lung-chiang) | Heilongjiang |
| Honan (Ho-nan) | Henan |
| Hopeh (Ho-pei) | Hebei |
| Hunan (Hu-nan) | Hunan |
| Hupeh (Hu-pei) | Hubei |
| Inner Mongolia (Nei-meng-ku) Autonomous Region | Nei Mongol Autonomous Region |
| Kansu (Kan-su) | Gansu |
| Kiangsi (Chiang-hsi) | Jiangxi |
| Kiangsu (Chiang-su) | Jiangsu |
| Kirin (Chi-lin) | Jilin |
| Kwangsi (Kuang-hsi), Chuang Autonomous Region of | Guangxi, Zhuang Autonomous Region of |
| Kwangtung (Kuang-tung) | Guangdong |
| Kweichow (Kuei-chou) | Guizhou |
| Liaoning (Liao-ning) | Liaoning |
| Ningsia (Ning-hsia), Hui Autonomous Region of | Ningxia, Hui Autonomous Region of |
| Peking (Pei-ching) Municipality | Beijing Municipality |
| Shanghai (Shang-hai) Municipality | Shanghai Municipality |
| Shansi (Shan-hsi) | Shanxi |
| Shantung (Shan-tung) | Shandong |
| Shensi (Shen-hsi) | Shaanxi |
| Sinkiang (Hsin-chiang), Uighur (Wei-wu-erh) Autonomous Region of | Xinjiang, Uygur Autonomous Region of |
| Szechwan (Ssu-ch'uan) | Sichuan |
| Tibet (Hsi-tsang) Autonomous Region | Xizang Autonomous Region |
| Tientsin (T'ien-ching) Municipality | Tianjin Municipality |
| Tsinghai (Ch'ing-hai) | Qinghai |
| Yunnan (Yün-nan) | Yunnan |
| **Cities and towns** | |
| A-k'o-su | Aksu |
| A-le-t'ai | Altay |
| Amoy (Hsia-men) | Xiamen |
| An-ch'ing | Anqing |
| An-shan | Anshan |
| An-shun | Anshun |
| An-yang (locally Chang-te) | Anyang (locally Zhangde) |
| Canton (Kuang-chou) | Guangzhou |
| Chan-chiang | Zhanjiang |
| Ch'ang-chi | Changji |
| Chang-chia-k'ou, see Kalgan | |
| Ch'ang-chih | Changzhi |
| Chang-chou | Zhangzhou |
| Ch'ang-chou | Changzhou |
| Ch'ang-ch'un | Changchun |
| Ch'ang-sha | Changsha |
| Ch'ang-te | Changde |
| Chang-yeh | Zhangye |
| Ch'ao-an (locally Ch'ao-chou) | Chao'an (locally Chaozhou) |
| Chao-ch'ing | Zhaoqing |
| Chao-chüeh | Zhaojue |
| Chao-tung | Zhaodong |
| Chao-t'ung | Zhaotong |
| Ch'ao-yang | Chaoyang |
| Chefoo (Yen-t'ai) | Yantai |
| Ch'en-chou | Chenzhou |
| Cheng-chou | Zhengzhou |
| Ch'eng-te | Chengde |
| Ch'eng-tu | Chengdu |
| Chi-an | Ji'an |
| Ch'i-ch'i-ha-erh, see Tsitsihar | |
| Chi-hsi | Jixi |
| Chi-lin (conventional Kirin) | Jilin |
| Chi-nan (conventional Tsinan) | Jinan |
| Chi-ning | Jining |
| Chi-shou | Jishou |
| Ch'i-t'ai-ho | Qitaihe |
| Chia-hsing | Jiaxing |
| Chia-mu-ssu | Jiamusi |
| Chia-yü-kuan | Jiayuguan |
| Chiang-men | Jiangmen |
| Chiao-tso | Jiaozuo |
| Ch'ih-feng | Chifeng |
| Chin-ch'ang | Jinchang |
| Chin-ch'eng | Jincheng |
| Chin-chou | Jinzhou |
| Ch'in-chou | Qinzhou |
| Chin-hsi | Jinxi |
| Chin-hua | Jinhua |
| Ch'in-huang-tao | Qinhuangdao |
| Chin-shih | Jinshi |
| Ch'ing-hua-chen | Huaying |
| Ching-kang-shan | Jinggangshan |
| Ch'ing-tao, see Tsingtao | |
| Ching-te-chen | Jingdezhen |
| Chiu-chiang | Jiujiang |
| Chiu-ch'üan (locally Su-chou) | Jiuquan (locally Suzhou) |
| Ch'iung-hai (locally Chia-chi) | Qionghai (locally Jiaji) |
| Cho-chou | Zhuozhou |
| Chou-k'ou | Zhoukou |
| Ch'ü-ching | Qujing |
| Chu-chou | Zhuzhou |
| Ch'ü-chou | Quzhou |
| Ch'ü-fu | Qufu |
| Chu-hai (locally Hsiang-chou) | Zhuhai (locally Xiangzhou) |
| Chu-ma-tien | Zhumadian |
| Ch'üan-chou | Quanzhou |
| Chungking (Ch'ung-ch'ing; locally Yü-chou) | Chongqing (locally Yuzhou) |
| Dairen, see Lü-ta | |
| En-shih | Enshi |
| Erh-lien-hao-t'e | Erenhot |
| Feng-ch'eng | Fengcheng |
| Fu-chou | Fuzhou |
| Fu-hsien (locally Wa-fang-tien) | Fuxian (locally Wafangdian) |
| Fu-hsin | Fuxin |
| Fu-ling | Fuling |
| Fu-shun | Fushun |
| Fu-yang | Fuyang |
| Ha-erh-pin, see Harbin | |
| Ha-mi | Hami |
| Hai-ch'eng | Haicheng |
| Hai-k'ou | Haikou |
| Hai-lung (locally Mei-ho-k'ou) | Hailong (locally Meichekou) |
| Hai-men | Haimen |
| Hailar (Hai-la-erh) | Hailar |
| Han-ch'eng | Hancheng |
| Han-chung | Hanzhong |
| Han-tan | Handan |
| Hang-chou | Hangzhou |
| Harbin (Ha-erh-pin) | Harbin |
| Heng-shui | Hengshui |
| Heng-yang | Hengyang |
| Ho-ch'ih (locally Chin-ch'eng-chiang) | Hechi (locally Jinchengjang) |
| Ho-fei | Hefei |
| Ho-kang | Hegang |
| Ho-pi | Hebi |
| Ho-shan | Heshan |
| Ho-tse (locally Ts'ao-chou) | Heze (locally Caozhou) |
| Hou-ma | Houma |
| Hsi-an, see Sian | |
| Hsi-ch'ang | Xichang |
| Hsi-ning | Xining |
| Hsia-kuan | Xiaguan |
| Hsia-men, see Amoy | |
| Hsiang-fan | Xiangfan |
| Hsiang-t'an | Xiangtan |
| Hsien-yang | Xianyang |
| Hsin-chou | Xinzhou |
| Hsin-hsiang | Xinxiang |
| Hsin-yang (locally P'ing-ch'iao) | Xinyang (locally Pingqiao) |
| Hsin-yü | Xinyu |
| Hsing-t'ai | Xingtai |
| Hsü-ch'ang | Xuchang |
| Hsü-chou, see Suchow | |
| Hu-chou (locally Wu-hsing) | Huzhou (locally Wuxing) |
| Hu-ho-hao-t'e | Hohhot |
| Huai-hua (locally Yü-shu-wan) | Huaihua (locally Yushuwan) |
| Huai-nan | Huainan |
| Huai-pei | Huaibei |
| Huai-te (locally Kung-chu-ling) | Huaide (locally Gonzhuling) |
| Huai-yin (locally Wang-ying) | Huaiyin (locally Wangying) |
| Huang-shih | Huangshi |
| Hui-chou | Huizhou |
| Hun-chiang (locally Pa-tao-chiang) | Hunjiang (locally Badaojiang) |
| Hung-chiang | Hongjiang |
| I-ch'ang | Yichang |
| I-ch'un | Yichun |
| I-ning (conventional Kuldja) | Yining |
| I-pin | Yibin |
| I-tu | Yidu |
| I-yang | Yiyang |
| Jin-chao | Rizhao |
| K'a-shih, see Kashgar | |
| K'ai-feng | Kaifeng |
| K'ai-li | Kaili |
| K'ai-yuan | Kaiyuan |
| Kalgan (Chang-chia-k'ou) | Zhiangjiakou |
| Kan-chou | Ganzhou |
| Karamai, see K'o-la-ma-i | |
| Kashgar (K'a-shih) | Kashi |
| Kirin, see Chi'lin | |
| Ko-chiu | Gejiu |
| K'o-erh-ch'in-yu-i-ch'ien-ch'i (locally Wu-lan-hao-t'e) | Horqin Youyi Qianqi (locally Ulan Hot) |
| Ko-erh-mu | Golmud |
| K'o-la-ma-i (conventional Karamai) | Karamay |
| K'u-erh-le | Korla |
| Kuang-chou, see Canton | |
| Kuang-hua (locally Lao-ho-k'ou) | Guanghua (locally Laohekou) |
| Kuei-lin | Guilin |
| K'uei-t'un | Kuytun |
| Kuei-yang | Guiyang |
| Kuldja, see I-ning | |
| K'un-ming | Kunming |
| Lai-wu | Laiwu |
| Lan-chou | Lanzhou |
| Lan-hsi | Lanxi |
| Le-shan | Leshan |
| Leng-shui-chiang | Lengshuijiang |
| Leng-shui-t'an | Lengshuitan |
| Lhasa (La-sa) | Lhasa |
| Li-ch'uan | Lichuan |
| Li-shui | Lishui |
| Liao-ch'eng | Liaocheng |
| Liao-yang | Liaoyang |
| Liao-yüan | Liaoyuan |
| Lien-yün-kang (locally Hsih-p'u) | Lianyungang (locally Xinpu) |

Lin-fen ..... Linfen
Lin-ho ..... Linhe
Lin-hsia ..... Linxia
Lin-i ..... Linyi
Liu-chou ..... Liuzhou
Lo-yang ..... Luoyang
Lou-ti ..... Loudi
Lu-an ..... Lu'an
Lu-chou ..... Luzhou
Lü-ta (conventional Dairen, locally Ta-lien) ..... Lüda (locally Dalian)
Lung-k'ou ..... Longkou
Lung-yen ..... Longyan
Man-chou-li ..... Manzhouli
Mao-ming ..... Maoming
Mei-chou ..... Meizhou
Mien-yang ..... Mianyang
Mu-tan-chiang ..... Mudanjiang
Nan-ch'ang ..... Nanchang
Nan-ching, see Nanking
Nan-ch'ung ..... Nanchong
Nan-kung ..... Nangong
Nan-ning ..... Nanning
Nan-p'ing ..... Nanping
Nan-t'ung ..... Nantong
Nan-yang ..... Nanyang
Nanking (Nan-ching) ..... Nanjing
Ning-po ..... Ningbo
O-cheng ..... Ezhou
Pai-ch'eng ..... Baicheng
Pai-se ..... Bose
Pai-yin ..... Baiyin
Pang-pu ..... Bengbu
Pao-chi ..... Baoji
Pao-shan ..... Baoshan
Pao-ting ..... Baoding
Pao-t'ou ..... Baotou
Pei-an ..... Bei'an
Pei-hai ..... Beihai
Pei-piao ..... Beipiao
Peking (Pei-ching) ..... Beijing
Pen-hsi ..... Benxi
Pin-chou ..... Binzhou
P'ing-hsiang ..... Pingxiang
P'ing-liang ..... Pingliang
P'ing-ting-shan ..... Pingdingshan
Po-chou ..... Bozhou
Pu-t'e-ha-ch'i (locally Cha-lan-t'un) ..... Butha Qi (locally Zalantun)
P'u-t'ien ..... Putian
P'u-yang ..... Puyang
San-men-hsia ..... Sanmenxia
San-ming ..... Sanming
Sha-shih ..... Shashi
Shan-t'ou, see Swatow
Shang-ch'iu (locally Chu-chi) ..... Shangqiu (locally Zhuji)
Shang-jao ..... Shangrao
Shanghai (Shang-hai) ..... Shanghai
Shao-hsing ..... Shaoxing
Shao-kuan ..... Shaoguan
Shao-wu ..... Shaowu
Shao-yang (locally Pao-ch'ing) ..... Shaoyang (locally Baoqing)
Shen-chen ..... Shenzhen
Shen-yang ..... Shenyang
Shih-chia-chuang ..... Shijiazhuang
Shih-ho-tzu ..... Shihezi
Shih-shou ..... Shishou
Shih-tsui-shan ..... Shizuishan
Shih-yen ..... Shiyan
Shuang-ya-shan ..... Shuangyashan
Shui-ch'eng ..... Shuicheng
Sian (Hsi-an) ..... Xi'an
Ssu-p'ing ..... Siping
Su-chou ..... Suzhou
Suchow (Hsü-chou) ..... Xuzhou
Sui-chou ..... Suizhou
Sui-hua ..... Suihua
Sui-ning ..... Suining
Swatow (Shan-t'ou) ..... Shantou
T'a-ch'eng ..... Tacheng
Ta-ch'ing ..... Daqing
Ta-hsien ..... Daxian
Ta-t'ung ..... Datong
T'ai-an ..... Tai'an
T'ai-chou ..... Taizhou
T'ai-yüan ..... Taiyuan
Tan-hsien (locally Na-ta) ..... Danxian (locally Nada)
Tan-tung ..... Dandong
T'ang-shan ..... Tangshan
Te-chou ..... Dezhou
Te-yang ..... Deyang
T'ieh-fa ..... Tiefa
T'ieh-ling ..... Tieling
T'ien-shui ..... Tianshui
Tientsin (T'ien-chin) ..... Tianjin
Ts'ang-chou ..... Cangzhou
Tsao-chuang ..... Zaozhuang
Tsinan, see Chi-nan
Tsingtao (Ch'ing-tao) ..... Qingdao
Tsitsihar (Ch'i-ch'i-ha-erh) ..... Qiqihar
Tsun-i ..... Zunyi
Tu-k'ou ..... Dukou
T'u-lu-p'an (conventional Turfan) ..... Turpan
T'u-men ..... Tumen
Tu-yün ..... Duyun
T'un-hsi ..... Tunxi
Tun-hua ..... Dunhua
T'ung-ch'uan ..... Tongchuan
T'ung-hua ..... Tonghua
T'ung-liao ..... Tongliao
T'ung-ling ..... Tongling
Tung-sheng ..... Dongsheng
Tung-ying ..... Dongying
Turfan, see T'u-lu-p'an
Tzu-hsing ..... Zixing
Tzu-kung ..... Zigong
Tzu-po (locally Chang-tien) ..... Zibo (locally Zhangdian)
Urumchi, see Wu-lu-mu-ch'i
Wan-hsien ..... Wanxian
Wan-ting-chen ..... Wandingzhen
Wei-fang ..... Weifang
Wei-hai ..... Wiehai
Wei-nan ..... Weinan
Wen-chou ..... Wenzhou
Wu-chou ..... Wuzhou
Wu-chung ..... Wuzhong
Wu-hai ..... Wuhai
Wu-han ..... Wuhan
Wu-hsi ..... Wuxi
Wu-hu ..... Wuhu
Wu-lu-mu-ch'i (conventional Urumchi) ..... Ürümqi
Wu-wei (locally Liang-chou) ..... Wuwei (locally Liangzhou)
Ya-hsien (locally San-ya) ..... Yaxian (locally Sanya)
Yang-ch'üan ..... Yangquan
Yen-an ..... Yan'an
Yen-ch'eng ..... Yancheng
Yen-chi ..... Yanji
Yen-t'ai, see Chefoo
Yin-ch'uan ..... Yinchuan
Ying-ch'en ..... Yingcheng
Ying-ch'uan ..... Yingchuan
Ying-k'ou ..... Yingkou
Ying-t'an ..... Yingtan
Yü-hsia ..... Yuxia
Yü-lin ..... Yulin
Yü-men (locally Lao-cun-miao) ..... Yumen (locally Laojunmiao)
Yü-tz'u ..... Yuci
Yü-yao ..... Yuyao
Yüeh-yang ..... Yueyang
Yün-ch'eng ..... Yuncheng
Yung-an ..... Yong'an
Yung-chou ..... Yongzhou

**Physical features and points of interest**

A-erh-chin Mountains ..... Altun Mountains
A-la-kou, *river* ..... Alo
A-ni-ma-ch'ing (conventional Amne Machin) Mountains ..... A'nyemaqen Mountains
Ai-pi, Lake ..... Ebinur, Lake
Ai-ting, Lake ..... Aydingkol, Lake
Altai (A-erh-t'ai) Mountains ..... Altay Mountains
Amur (Hei-lung), *river* ..... Heilong
Ang-lung Mountains ..... Nganglong Mountains
Ang-tse, Lake ..... Ngangze, Lake
Argun (O-erh-ku-na), *river* ..... Ergun
Black (Li-hsien), *river* ..... Lixian
Brahmaputra (Ya-lu-tsang-pu), *river* ..... Yarlung
Cha-chia-tsang-pu, *river* ..... Za'gya
Ch'a-kan, Lake ..... Qagan, Lake
Cha-ling, Lake ..... Gyaring, Lake
Ch'ai-ta-mu, see Tsaidam Basin
Ch'ang Chiang, see Yangtze
Ch'ang-pai Mountains ..... Changbai Mountains
Ch'ang-pai Nature Reserve ..... Changbai Nature Reserve
Ch'ao, Lake ..... Chao, Lake
Ch'e-erh-ch'en, *river* ..... Qarqan
Ch'i-lien, Mount ..... Qilian, Mount
Ch'i-lin, Lake ..... Siling, Lake
Ch'iao-ko-li, Mount ..... Qogir, Mount
Chihli, Gulf of, see Po Hai
Ch'in, see Tsinling Mountains
Chin-sha, *river* ..... Jinsha
Ch'ing-hai, Lake, see Koko Nor
Ch'ing-tsang, see Tibet, Plateau of
Chu-mu-lang-ma, see Everest, Mount
Dzungarian (Chun-ko-erh) Basin ..... Junggar Basin
East China (Tung) Sea ..... Dong Sea
Everest (Chu-mu-lang-ma), Mount ..... Qomolangma, Mount
Fen, *river* ..... Fen
Feng-shui, Mount ..... Fengshui, Mount
Gobi Desert ..... Gobi Desert
Grand Canal (Ta Yün-ho) ..... Da Yunhe
Greater Khingan (Ta-hsing-an) Range ..... Da Hinggan Range
Hai-la-erh, *river* ..... Hailar
Hai-nan Island ..... Hainan Island
Han, *river* ..... Han
Han-ma Nature Reserve ..... Hanma Nature Reserve
Han-t'eng-ko-li, Mount ..... Hantengri, Mount
Hei-lung, see Amur
Heng, Mount ..... Heng, Mount
Heng-tuan Mountains ..... Hengduan Mountains
Himalayas (Hsi-ma-la-ya), *mountains* ..... Himalaya Mountains
Ho-lan Mountains ..... Helan Mountains
Hsi, *river* ..... Xi
Hsi-hsia-pang-ma, Mount ..... Xixabangma, Mount
Hsi-liao, *river* ..... Xiliao
Hsiao-hsing-an, see Lesser Khingan Range
Hsin-an-chiang Reservoir ..... Xin'anjiang Reservoir
Hsing-k'ai, see Khanka, Lake
Hsun, *river* ..... Xun
Hu-chuang Nature Reserve ..... Huzhong Nature Reserve
Hu-lun, Lake ..... Hulun, Lake
Hu-t'o, *river* ..... Hutuo
Hua, Mount ..... Hua, Mount
Huai, *river* ..... Huai
Huang, *river* ..... Huang
Huang, see Yellow Sea
Huang Ho, *river* ..... Huang He
Huang-kang-liang Mountain ..... Huanggang-liang Mountain
Huang Mountains ..... Huang Mountains
Huang-t'u, see Loess Plateau
Hung-shui, *river* ..... Hongshui
Hung-tse Lake ..... Hongze Lake
I-li, *river* ..... Ili
K'a-la-k'a-shih, *river* ..... Karakax
K'a-la-k'un-lun Range ..... Karakoram Range
Ka-shun, Lake ..... Gaxun, Lake
K'ai-tu, *river* ..... Kaidu
Kailas Range (Kang-ti-ssu Mountains) ..... Gangdise Mountains
Kan, *river* ..... Gan
Kang-jen-po-ch'i Mountain ..... Kangrinboqe Mountain

Kang-ti-ssu, see Kailas Range
Kerulen (K'o-lu-lun), *river* ..... Herlen
Khanka (Hsing-k'ai), Lake ..... Xingkai, Lake
K'o-k'o-hsi-li Mountains ..... Hoh Xil Mountains
K'o-li-ya, *river* ..... Keriya
K'o-lu-lun, see Kerulen
Koko (Ch'ing-hai) Nor, *lake* ..... Koko Nur (Qinghai)
K'u-la (Kula, Bhutan), Mount .... Gula, Mount
K'u-lu-k'o-t'a-ko, *mountains* ..... Kuruktag Mountains
Kuang-ming Mountain ..... Guangming Mountain
K'ung-ch'ueh, *river* ..... Konqi
Kung-k'a, Mount (conventional Minya Konka) ..... Gongga, Mount
Kung-ko-erh Mountain ..... Kongur Mountain
Kunlun (K'un-lun) Mountains .... Kunlun Mountains
La-lü, see Yalu
Lan-ts'ang, see Mekong
Lao-ha, *river* ..... Laoha
Lesser Khingan (Hsiao-hsing-an) Range ..... Xiao Hinggan Range
Li-hsien, see Black
Liao, *river* ..... Liao
Liu-p'an Mountains ..... Liupan Mountains
Loess (Huang-t'u) Plateau ..... Huangtu Plateau
Lop (Lo-pu) Nor, *lake* ..... Lop Nur
Lü-liang Mountains ..... Lüliang Mountains
Luan, *river* ..... Luan
Ma-ch'ing, Mount ..... Maqen, Mount
Ma-fa-mu Lake ..... Mapam Lake
Ma-tsung, Mount ..... Mazong, Mount
Ma'ch'üan, *river* ..... Maquan
Manchurian (Tung-pei) Plain ..... Dongbei Plain
Mao-wu-su, see Ordos Desert
Mekong (Lan-ts'ang), *river* ..... Lancang
Min, *river* ..... Min
Min Mountains ..... Min Mountains
Minya Konka, see Kung-k'a, Mount
Mu-shih, Mount ..... Muztag, Mount
Mu-shih-t'a-ko, Mount ..... Muztagata, Mount
Mu-tau, *river* ..... Mudan
Mu-tzu-t'a-ko, Mount ..... Muztag, Mount
Na-mu, Lake ..... Nam, Lake
Nan, see South China Sea
Nan Mountains ..... Nan Mountains
Nan-p'an, *river* ..... Nanpan
Nan-tu, *river* ..... Nandu
Nan-weng, *river* ..... Nanweng
Nen, *river* ..... Nen
Nien-ch'ing-t'ang-ku-la, Mount ... Nyainqentang-lha, Mount
Nien-ch'ing-t'ang-ku-la Mountains ..... Nyainqentanglha Mountains
No-min, *river* ..... Nuomin
Nu, see Salween
O-erh-ch'i-ssu, *river* ..... Ertix
O-erh-ku-na, see Argun
O-ling, Lake ..... Ngoring, Lake
O-mei, Mount ..... Emei, Mount
Ordos (Mao-wu-su) Desert ..... Mu Us Desert
Pamirs (P'a-mi-erh), *region* ..... Pamirs
Pa-yen-k'a-la Mountains ..... Bayan Har Mountains
Pai-shui River Nature Reserve ... Baishui River Nature Reserve
Pei, *river* ..... Bei
Pei-p'an, *river* ..... Beipan
Pei-pu, see Tonkin, Gulf of
Po Hai (conventional Chihli, Gulf of) ..... Bo Hai
Po-ko-ta, Mount ..... Bogda, Mount
Po-ssu-t'eng, Lake ..... Bosten, Lake
P'o-yang, Lake ..... Poyang, Lake
Red (Yüan), *river* ..... Yuan
Sai-li-mu, Lake ..... Sayram, Lake
Salween (Nu), *river* ..... Nu
Shantung (Shan-tung) Peninsula ..... Shandong Peninsula
Sheng-li, see Victory Peak
Shu-le, *river* ..... Shule
South China (Nan) Sea ..... Nan Sea
Ssu-ch'uan, *region* ..... Sichuan Basin
Su-ku, Lake ..... Sogo, Lake
Sung, Mount ..... Song, Mount
Sungari (Sung-hua), *river* ..... Songhua
Sungari (Sung-hua) Reservoir ... Songhua Reservoir
Ta-hsing-an, see Greater Khingan Range
Ta-hsüeh Mountains ..... Daxue Mountains
T'a-k'o-la-ma-kan, see Takla Makan Desert
Ta-kuo-k'uei, Mount ..... Daguokui, Mount
T'a-li-mu, see Tarim
T'a-li-mu, see Tarim Basin
Ta-liang Mountains ..... Daliang Mountains
Ta-lou Mountains ..... Dalou Mountains
Ta-pan Mountains ..... Daban Mountains
Ta-pieh Mountains ..... Dabie Mountains
Ta-tu, *river* ..... Dadu
Ta Yün-ho, see Grand Canal
T'ai, Lake ..... Tai, Lake
T'ai, Mount ..... Tai, Mount
T'ai-hang Mountains ..... Taihang Mountains
T'ai-pai, Mount ..... Taibai, Mount
Taiwan (T'ai-wan) Strait ..... Taiwan Strait
Takla Makan (T'a-k'o-la-ma-kan) Desert ..... Taklimaken Desert
Tan-chiang-k'ou Reservoir ..... Danjiangkou Reservoir
T'ang-ku-la Mountains ..... Tanggula Mountains
T'ao-erh, *river* ..... Tao'er
Tarim (T'a-li-mu), *river* ..... Tarim
Tarim (T'a-li-mu) Basin ..... Tarim Basin
Ti-erh-sung-hua, *river* ..... Di'er Songhua
Tibet (Ch'ing-tsang), Plateau of ..... Qing Zang Plateau
Tien (T'ien) Shan, *mountains* .... Tian Shan
Tien-ts'ang, Mount ..... Diancang, Mount
T'o-t'o, *river* ..... Tuotuo
Tonkin (Pei-pu), Gulf of ..... Beibu, Gulf of
Tsaidam (Ch'ai-ta-mu) Basin .... Qaidam Basin
Tsinling (Ch'in) Mountains ..... Qin Mountains
T'u-lu-fan, see Turfan Depression
Tumen (T'u-men), *river* ..... Tumen
Tung, see East China Sea
Tung-liao, *river* ..... Dongliao
Tung-pei, see Manchurian Plain
T'ung-t'ien, *river* ..... Tongtian
Tung-t'ing, Lake ..... Dongting, Lake
Turfan (T'u-lu-fan) Depression ... Turpan Depression
Tzu-ya, *river* ..... Ziya
Ussuri (Wu-su-li), *river* ..... Wusuli
Victory (Sheng-li) Peak ..... Shengli Peak
Wen-Ch'uan Wo-lung Nature Reserve ..... Wenchuan Wolong Nature Reserve
Wu, *river* ..... Wu
Wu-chih, Mount ..... Wuzhi, Mount
Wu-i Mountains ..... Wuyi Mountains
Wu-lan-ku, *river* ..... Ulungur
Wu-lan-su, Lake ..... Ulansuhai, Lake
Wu-liang Mountains ..... Wuliang Mountains
Wu-su-li, see Ussuri
Wu-t'ai, Mount ..... Wutai, Mount
Ya-la-ta-tse, Mount ..... Yagradagze, Mount
Ya-lu, *river* ..... Yalu
Ya-lu-tsang-pu, see Brahmaputra
Ya-lung, *river* ..... Yalong
Yalu (Ya-lü), *river* ..... Yalu
Yangtze (Ch'ang Chiang), *river* ..... Chang Jiang
Yeh-erh-ch'iang, *river* ..... Yarkant
Yellow (Huang) Sea ..... Huang Sea
Yin Mountains ..... Yin Mountains
Ying, *river* ..... Ying
Yü, *river* ..... Yu
Yü, Mount ..... Yu, Mount
Yu-i, Mount ..... Youyi, Mount
Yü-lung-hsüeh, Mount ..... Yulongxue, Mount
Yüan, see Red
Yüan, *river* ..... Yuan
Yung-ting, *river* ..... Yongding

shore islands. To the north, except along the Shantung and Liaotung peninsulas, the coast is sandy and flat.

The influence of relief on historic development

China's physical relief has dictated its development in many respects. The civilization of China originated in the southern part of the Loess Plateau, and from there it extended outward until it encountered the combined barriers of relief and climate. The long, protruding corridor, commonly known as the Kansu, or Hosi, Corridor, illustrates this fact. South of the corridor is the Plateau of Tibet, which was too high and too cold for the civilization of China to gain a foothold. North of the corridor is the Gobi Desert, which also formed a barrier. Consequently, Chinese civilization was forced to spread along the corridor, where melting snow and ice in the Ch'i-lien Mountains provided water for oasis farming. The westward extremities of the corridor became the meeting place of the ancient East and West.

Thus, for a long time the ancient political centre of China was located along the lower reaches of the Huang Ho (Yellow River). Because of topographical barriers, however, it was very difficult for the central government, except when represented by an unusually strong dynasty, to gain complete control over the entire country. In many instances the Szechwan Basin—an isolated region in southwest China, about twice the size of Scotland, that is well protected by high mountains and is self-sufficient in agricultural products—became an independent kingdom. A comparable situation often arose in the Tarim Basin. Linked to the rest of China only by the Kansu Corridor, this basin is even remoter than the Szechwan, and when the central government was unable to exert its influence, oasis states were established; only the three strong dynasties—the Han (206 BC–AD 220), the T'ang (AD 618–907), and the Ch'ing, or Manchu (1644–1911/12)—were capable of controlling the region.

Apart from the three altitudinal zones already men-

tioned, it is possible—on the basis of geologic structure, climatic conditions, and differences in geomorphologic development—to divide China into three major topographical regions: the eastern, northwestern, and southwestern zones. The eastern zone is a region shaped by the rivers, which have eroded landforms in some parts and have deposited alluvial plains in others; its climate is monsoonal (characterized by seasonal rain-bearing winds). The northwestern region is arid and eroded by the wind; it forms an inland drainage basin. The southwest is a cold, lofty, and mountainous region containing intermontane plateaus and inland lakes.

The three basic regions may be further subdivided into second-order geographic divisions. The eastern region contains 10 of these, the southwest contains two, and the northwest contains three. Below is a brief description of each division.

**The eastern region.** *The Manchurian Plain.* The Manchurian Plain (also known as the Northeast Plain and the Sung-liao Plain) is located in China's Northeast, the region formerly known as Manchuria. An undulating plain split into northern and southern halves by a low divide rising from 500 to 850 feet, it is drained in its northern part by the Sungari River and tributaries and in its southern part by the Liao River. Most of the area has an erosional rather than a depositional surface, but it is covered with a deep soil. The plain has an area of about 135,000 square miles. Its basic landscapes are forest-steppe, steppe, meadow-steppe, and cultivated land; its soils are rich and black, and it is a famous agricultural region. The river valleys are wide and flat with a series of terraces formed by deposits of silt. During the flood season the rivers inundate extensive areas.

*The Ch'ang-pai Mountains.* To the southeast of the Manchurian Plain is a series of ranges comprising the Ch'ang-pai, Chang-kuang-ts'ai, and Wan-ta mountains, which in Chinese are collectively known as the Ch'ang-pai Shan, or "Forever White Mountains"; broken by occasional open valleys, they reach altitudes mostly between 1,500 and 3,000 feet. In some parts the scenery is characterized by rugged peaks and precipitous cliffs. The highest peak is the volcanic cone of Mt. Pai-t'ou (9,003 feet), which has a beautiful crater lake at its snow-covered summit. As one of the major forest areas of China, the region is the source of many valuable furs and famous medicinal herbs. Cultivation is generally limited to the valley floors.

*The North China Plain.* Comparable in size to the Manchurian Plain, most of this plain lies at heights below 150 feet, and the relief is monotonously flat. It was formed by enormous sedimentary deposits brought down by the Huang Ho and Huai River from the Loess Plateau; the Quaternary deposits alone (*i.e.,* those from 10,000 to 2,500,000 years old) reach thicknesses of 2,500 to 3,000 feet. The river channels, which are higher than the surrounding locality, form local water divides, and the areas between the channels are depressions in which lakes and swamps are found. In particularly low and flat areas, the underground water table often fluctuates from five to six and a half feet, resulting in the formation of meadow swamp and, in some places, saline soils. A densely populated area that has long been under settlement, the North China Plain has the highest proportion of land under cultivation of any region in China.

Formation of the plain

*The Loess Plateau.* This plateau forms a unique region of loess-clad hills and barren mountains situated between the North China Plain and the deserts of the west. In the north, the Great Wall of China forms the boundary, while the southern limit is the Tsinling Mountains, in Shensi Province. The average altitude of the surface is between 4,000 and 5,000 feet, but individual ranges of bedrock are higher, reaching 9,285 feet in the Liu-p'an Mountains. Most of the plateau is covered with loess to thicknesses of 150 to 650 feet. In northern Shensi Province and eastern Kansu Province the loess may reach a thickness of 800 feet. The loess is particularly susceptible to erosion by water; ravines and gorges are found crisscrossing the plateau. It has been estimated that ravines cover approximately one-half of the entire region, with erosion reaching depths of 300 to 650 feet.

*The Shantung Hills.* These hills are basically composed of Archean (Early to Middle Precambrian) crystalline shales and granites and of Lower Paleozoic sedimentary rocks (*i.e.,* between 395,000,000 and 570,000,000 years old). Faults have played a major role in creating the present relief, and, as a result, many hills are horsts (blocks of the Earth's crust uplifted along faults), while the valleys have been formed by grabens (blocks of the Earth's crust that have been downthrown along faults). The Chiao-lai Plain divides this region into two parts. The eastern part is lower, lying at altitudes averaging less than 1,500 feet, with only certain peaks and ridges rising to 2,500 feet and (rarely) to 3,000 feet, and only one mountain, Mt. Lao, reaching a height of 3,707 feet. The western part is slightly higher; the highest peak is Mt. T'ai (5,026 feet), which is sacred as a symbol of the divine election of the ruling house of China. The Shantung Hills meet the sea along a bold and rocky shoreline.

*The Tsinling Mountains.* The Tsinling Mountains in Shensi Province are the greatest chain of mountains east of the Plateau of Tibet. The mountain chain consists of a high and rugged barrier extending from Kansu to Honan; geographers use a line between the chain and the Huai River to divide China proper into two parts—North and South. The altitude of the mountains varies from 3,000 to 10,000 feet. The western part is higher, with the highest peak, Mt. T'ai-pai, rising to 13,474 feet. The Tsinling Mountains consist of a series of parallel ridges, all running slightly south of east, separated by a maze of ramifying valleys whose canyon walls often rise sheer to a height of 1,000 feet above the valley streams.

Division between North and South China

*The Szechwan Basin.* This basin (also known as the Red Basin) is one of the most attractive geographical regions of China. The basin is surrounded by mountains, which are higher in the west and north. Protected against the penetration of cold northern winds, the basin is much warmer in the winter than are the more southerly plains of southeast China. Except for the Ch'eng-tu Plain, the region is very hilly. In the eastern half of the basin are numerous folds, forming a series of ridges and valleys that trend northeast to southwest. The lack of arable land has obliged farmers to cultivate the slopes of the hills, on which they have built terraces that frequently cover the slopes from top to bottom. The terracing has slowed down the process of erosion and has made it possible to cultivate additional areas by using the steeper slopes—some of which have grades up to 45° or more.

*The Southeast Mountains.* Southeastern China is bordered by a rocky shoreline backed by picturesque mountains. In general, there is a distinct structural and topographic trend from northeast to southwest. The higher peaks may reach altitudes of 5,000 to 6,500 feet. The rivers are short and fast-flowing and have cut steep-sided valleys. The chief areas of settlement are on narrow strips of coastal plain where rice is produced. Along the coast there are numerous islands, where the fishing industry is well developed.

*Plains of the Middle and Lower Yangtze.* East of I-ch'ang, in Hupeh Province, a series of plains of uneven width are found along the Yangtze River, or Ch'ang Chiang. The plains are particularly wide in the delta area and in places where the Yangtze receives its major tributaries—including large areas of lowlands around the lakes of Tung-t'ing, P'o-yang, T'ai, and Hung-tse, which are hydrologically linked with the Yangtze. The region is an alluvial plain, the accumulation of sediment laid down by the rivers throughout long ages. There are a few isolated hills, but in general the land is level, lying mostly below 150 feet. Rivers, canals, and lakes form a dense network of waterways. The surface of the plain has been converted into a system of flat terraces, which descend in steps along the slopes of the valleys.

*The Nan Mountains.* The Nan Mountains (in Chinese Nan Ling) are composed of many ranges of mountains running from northeast to southwest. These ranges form the watershed between the Yangtze to the north and the Pearl River (Chu Chiang) to the south. The main peaks along the watershed are above 5,000 feet, and some are more than 6,500 feet. But a large part of the land to the

south of the Nan Mountains is also hilly; flatland does not exceed 10 percent of the total area. The Pearl River Delta is the only extensive plain in this region and is also the richest part of South China. The coastline is rugged and irregular, and there are many promontories and protected bays, including that of Hong Kong. The principal river is the Hsi River, which rises in the highlands of eastern Yunnan and southern Kweichow.

**The southwest.** *The Yunnan–Kweichow highland region.* This region comprises the northern part of Yunnan and the western part of Kweichow; its edge is highly dissected. Yunnan is more distinctly a plateau and contains larger areas of rolling uplands than Kweichow, but both parts are distinguished by canyon-like valleys and precipitous mountains. The highest elevations lie in the west, where Mt. Tieh-chi'ang rises to 12,080 feet. In the valleys of the major rivers, elevations drop to 1,300–1,600 feet. Particularly sharp differences in elevation and the greatest ruggedness of relief occur in the western part of the region, in the gorges of the large rivers. In the eastern part, karst processes (creating sinks, ravines, and underground streams in the limestone landscape) have developed very strongly. Scattered throughout the highlands are small lake basins, separated by mountains.

*The Plateau of Tibet.* This great massif occupies about one-fourth of the whole country. A large part of the plateau lies at elevations above 13,000–15,000 feet. The border ranges of the plateau are even higher, with individual peaks rising to heights of 23,000–26,000 feet. The interior slopes of these border mountains, as a rule, are gentle, while the exterior slopes are very steep. In its eastern and southern periphery, great rivers—such as the Yangtze, Huang, Mekong, Salween, Indus, and Brahmaputra—find their sources. Only in the low valleys, chiefly along the Brahmaputra Valley, are there centres of human settlement.

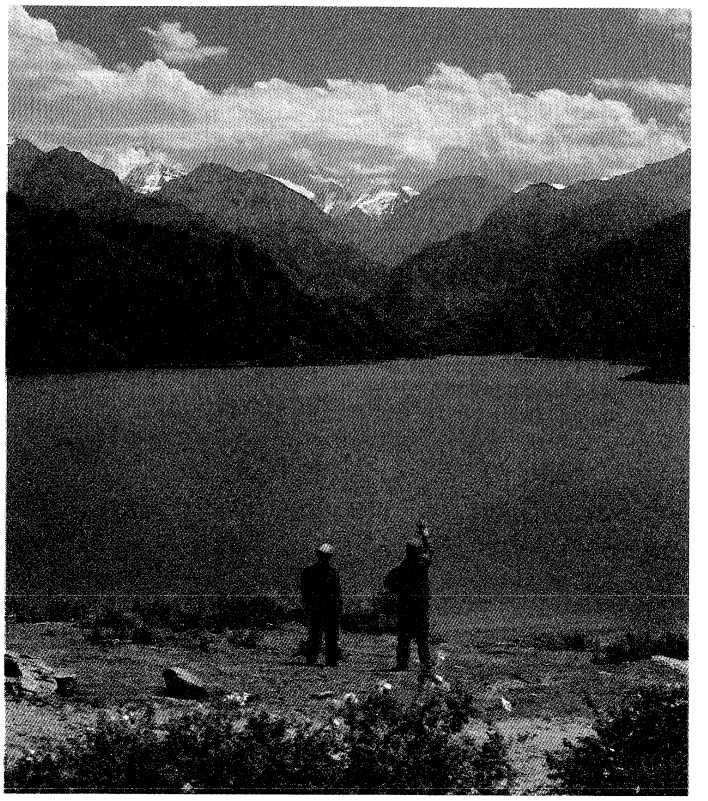
The Tsaidam Basin

The Tsaidam Basin is the largest, as well as the lowest, basin in the Plateau of Tibet. The broad northwestern part of the basin lies at elevations from approximately 8,800 to 10,000 feet; the narrow southeastern part lies between about 8,500 and 8,800 feet. Within the basin there is a predominance of gravel, sandy and clay deserts, semideserts, and salt wastes.

**The northwest.** *The Tarim Basin.* North of the Plateau of Tibet and at the much lower level of about 3,000 feet lies the Tarim Basin. It is hemmed in by great mountain ranges: the Tien Shan ("Celestial Mountains") on the north, the Pamirs on the west, and the Kunlun Mountains on the south. From these heights glacier-fed streams descend, only to lose themselves in the loose sands and gravels of the Takla Makan Desert, which occupies the centre of the basin. The Takla Makan is one of the most barren of the world's deserts; only a few of the largest rivers, such as the Tarim and Ho-t'ien (Khotan), cross the desert, but even their flow is not constant, and they have water throughout their entire courses only during the flood period. The area of the basin is about 215,000 square miles, and its elevations are from 2,500 to 4,600 feet above sea level. Its surface slants to the southeast, where the Lop Nor (a salt-encrusted lake bed) is situated.

*The Dzungarian Basin.* North of the Tarim Basin is another large depression, the Dzungarian. It is enclosed by the Tien Shan on the south, while to the northeast it is cut off from the Mongolian People's Republic by the Altai Mountains. The surface of the basin is flat, with a gentle slope to the southwest. The larger portion of the land lies at elevations between 1,000 and 1,500 feet, and in the lowest part the elevation drops to 620 feet. In general the main part of the basin is covered by a broad desert with barchans (crescent-shaped sand dunes that move); only in certain parts are dunes retained by vegetation.

*The Tien Shan.* The Chinese part of the Tien Shan consists of a complex system of ranges and depressions in which two major groups of ranges—the northern and southern—may be distinguished. They are separated by a strip of intermontane depressions that is broken up by the interior ranges. Ancient metamorphic rock (formed under heat and pressure) constitutes the larger portion of the ranges in the interior zone; Paleozoic (from 225,000,000 to 570,000,000 years old) sedimentary and igneous sedimentary beds form its northern and southern chains, while Mesozoic (from 65,000,000 to 225,000,000 years old) sandstones and conglomerates fill the intermontane depressions in the interior zone and constitute the foothill ridges. The height of the main chains of the eastern Tien Shan is between 13,000 and 15,000 feet, with individual peaks exceeding 16,000 feet; the interior chains reach 14,500 feet. In the western part, where there is adequate precipitation, large glaciers are formed, reaching a length of more than 20 miles. Large rivers with much water, such as the T'e-k'o-ssu River and its tributaries, begin their courses there. With predominantly alpine meadow steppe, this area is one of the best grazing lands of China.



Tien Lake at the foot of the Po-ko-ta Mountains in the eastern Tien Shan, Uighur Autonomous Region of Sinkiang.
K. Scholz—Shostal Assoc.

## DRAINAGE

China has more than 50,000 rivers with individual drainage areas of more than 40 square miles. Of the total annual runoff (amount of water they carry away) about 95 percent drains directly into the sea (more than 80 percent into the Pacific Ocean, 12 percent into the Indian Ocean, and less than 1 percent into the Arctic Ocean) and 5 percent disappears inland.

The principal rivers

The three principal rivers of China, all of which flow generally from west to east, draining into the China Sea, are the Huang, the Yangtze, and the Hsi. The Huang Ho, which rises in the Kunlun Mountains, is the northernmost of the three; it drains into the Po Hai (Gulf of Chihli), north of the Shantung Peninsula. The Yangtze, the longest river in the country, rises in the Tibetan Highlands and flows across central China, draining into the East China Sea north of Shanghai. The Hsi River, the southernmost of the three, rises in the Yunnan–Kweichow highlands and empties into the South China Sea via the Pearl River Delta, at Canton.

The distribution of surface water in China is extremely uneven. Only a small part of the country has enough water year-round. Much of the country has abundant runoff but only during the rainy summer, when enormous surpluses of water are received. From the southeast to the northwest, the surface water decreases as mountainous

features increase. A vast area of the northwest lacks water throughout the year. North China (north of the Tsinling Mountains–Huai River line), with its flat relief and long history of agriculture, contains almost two-thirds of China's cultivated land; paradoxically, because of scanty and erratic precipitation, the average annual runoff in the North accounts for only about one-sixth of the total for the country as a whole.

The mountains of the southeast and the mountainous Hai-nan Island have the most abundant surface water. Over the year they receive more than 60 inches (1,500 millimetres) of precipitation (in some places even more than 80 inches), of which almost two-thirds constitutes the runoff, so that a dense drainage network has developed. The amount of runoff is highest in the southeast, exceeding 40 inches. It gradually diminishes toward the west and north, so that in the true deserts of the northwest it is usually less than 0.4 of an inch. The arid climate of the northwest is reflected in the landscape of the dry steppes, which is characterized by richer grasses in the east, while in the west the landscape gradually changes to bare deserts.

In the lower reaches of the Yangtze, the Pearl River Delta, and the Ch'eng-tu Plain a dense network of waterways has been developed. In the North China Plain and the Manchurian Plain most of the rivers have a linear flow, and tributaries are few and unconnected. In the inland drainage area there are very few rivers because of scanty precipitation. Extensive areas such as the Tarim Basin and northeastern Kansu Province are often completely devoid of runoff. In these regions the rivers depend upon melted snow and ice; in consequence, they are mostly small and are found only in mountains and mountain foothills. As they drain farther and farther away from the mountains, most of them eventually disappear in the desert, while some form inland lakes. Because the northern part of the Plateau of Tibet is a cold desert, the rate of evaporation is slow, so that a denser network of rivers has developed; most of these, however, run into glaciated depressions, forming numerous lakes.

## SOILS

Because of its vast and diverse climatic conditions, China has a wide variety of soils. Indeed, all the soils of the Eurasian continent, except the soils of the tundra and the highly leached podzolic-gley soils of the northern taiga, are found in China. As a result of the climatic differences between the drier and cooler North and the wetter and hotter South, soils may be grouped into two classifications. Generally speaking, the soils north of the Tsinling Mountains–Huai River line are pedocals (calcareous) and are neutral to alkaline in reaction; those south of this line are pedalfers (leached noncalcareous soils), which are neutral to acid.

Apart from the great plateaus and high mountains to the southwest, marked soil zones are formed in China according to differences in climate, vegetation, and distance from the sea. The east and southeast coastal region is covered by the forest zone associated with a humid and semihumid climate, while the north and northwest inland regions belong mostly to the steppe zone, as well as to the semidesert and desert zone associated with a semiarid and arid climate. Between these two broad soil zones lies a transitional zone—the forest-steppe zone, where forest soils merge gradually with steppe soils.

Between the pedocals of the North and the pedalfers of the South lie the neutral soils. The floodplain of the Yangtze below the Three Gorges, where the river cuts through the Wu Mountains to empty onto the Hupeh Plain, is overlain with a thick cover of noncalcareous alluvium. These soils, sometimes classified as paddy (rice-growing) soils, for the most part are exceedingly fertile and of good texture. The paddy soil is a unique type of cultivated soil; it has been formed over a long period of time under the specific conditions of rice cultivation.

Along the coast of North China are belts of saline and alkaline soil. They are associated with a combination of poor drainage and aridity, where the rainfall is insufficient either to dissolve or to carry away the salts in solution.

The adverse effects of nature on the soil have been further intensified by centuries of concentrated cultivation, which has resulted in an almost universal deficiency of nitrogen and of organic matter. The shortage of organic matter is due primarily to the habitual use by farmers of crop stalks and leaves for livestock feed and fuel. The animal manure and night soil used for fertilizer contain too small an amount of organic matter to compensate for the loss of nutrients in the soil. The soils are also often deficient in phosphorus and potassium, but these deficiencies are neither so widespread nor so severe as that of nitrogen.

At one time, half of the territory of China may have been covered by forests, but now less than one-tenth of the country is forested. Extensive forests in central and southern China were cleared for farmlands, resulting in the inevitable erosion of soils from the hillsides and their deposition in the valleys. Farmers have constructed level terraces, supported by walls, in order to hold back water for rice fields, thus effectively controlling erosion. Wherever elaborate terraces have been built, soil erosion is virtually absent, and stepped terraces have become one of the characteristic features of the rural landscape.

Soil erosion

Excessive grazing and destruction of the grass cover also have resulted in soil removal. With the valuable crumb structure broken down and porosity lost, the topsoil is easily washed away through erosion in the rainy season; if the region happens to be a dry one, the wind produces the same effect. The Loess Plateau, constantly buffeted by rain and wind, is especially vulnerable to soil erosion, which results in a distinctive landscape. Steep-sided gullies, some of which may be several hundred feet deep, cut the plateau into fantastic relief. The damage done by heavy rain in summer is caused not only by the loss of topsoil but also by the frequent flooding of silt-laden rivers.

(C.-S.Ch./K.G.L./Ed.)

## CLIMATE

**The air masses.** China lies in Asia, the world's largest continent, and faces the Pacific, the world's largest ocean. Between the Pacific and China's part of Asia, there is a seasonal movement of air masses. The polar continental air mass, originating in Siberia, dominates a large part of China during the winter, while the tropical Pacific air mass exerts its influence during the summer. The sharply varied climatic conditions prevailing in summer and in winter are a direct result of the interaction of these two air masses, which are entirely different in nature.

The Siberian air mass, which is quite stable, is extremely cold and dry and often has marked layers of temperature inversion. After crossing the Mongolian Plateau, the air mass spreads southward and begins to invade North China, where it undergoes a series of rapid changes; its temperature rises slightly, and its stability decreases. During the day it may be quite warm, but at night or in shaded places the cold is often unbearable. In general, the diurnal (daily) range of temperature is more than 18° F (10° C); in extreme cases, it may exceed 45° F (25° C). Because North China is affected by this air mass most of the time, it is dry, with clear weather and an abundance of sunshine during the winter months.

The prevailing winter wind blows from November through March; it changes direction as it moves to the south. In northern and northeastern China its direction is from the northwest, in eastern China it comes from the north, and on the southeastern coasts it is from the northeast. The height of the winter wind belt usually does not exceed 13,000 feet. As it moves to the south, the height decreases; in Nanking it is about 6,500 feet, and in South China it is less than 5,000 feet. The Tsinling Mountains become an effective barrier to the advance of the cold waves to the south, particularly in the western section, where the average altitude of the mountains is mainly between 6,500 and 9,000 feet.

In China the tropical Pacific air mass is the chief source of the summer rainfall. When it predominates, it may cover the eastern half of China and penetrate deep into the border areas of the Mongolian Plateau and onto the eastern edge of the Plateau of Tibet. In summer the Siberian air mass retreats to the western end of Mongolia, although it occasionally extends southward and sometimes

Effects of the tropical Pacific air mass

may reach the Huai River Valley, which constitutes a battleground between the tropical Pacific and Siberian air masses in summer.

The movement of the two air masses is of immense significance to the climate of central and North China. In summer, when the tropical air mass predominates, the frontal zone between the two shifts northward; as a result, North China receives heavier rainfall. When the southeastern monsoon slackens, however, the frontal zone moves southward, and central China receives more rainfall, as a result of which floods may occur. The activity of the tropical Pacific air mass in winter is confined to the southeast coastal areas; during this season, therefore, drizzle frequently occurs in the hilly areas south of the Nan Mountains, and fog often appears in the morning.

Besides these two air masses, three other air masses also play their parts in the making of the climate of China. These are the equatorial continental air mass (a highly unstable southwest monsoon), the polar maritime air mass, and the equatorial maritime air mass. Furthermore, because China is a vast country with complex topography, the interaction between the air masses and relief produces many different types of climate.

The temperature pattern

**Temperature.** Temperatures generally decrease from south to north. The mean annual temperature is above 68° F (20° C) in the Pearl River Valley. It decreases to between 59° and 68° F (15° and 20° C) in the middle and lower reaches of the Yangtze, to about 50° F (10° C) in North China and the southern part of Sinkiang, and to 41° F (5° C) in the southern area of the Northeast, the northern part of Sinkiang, and places near the Great Wall. It drops below 32° F (0° C) in the northern part of Heilungkiang. The annual range of temperature between the extreme south and north is about 86° F (48° C). With few exceptions, January is the coldest month, and July is the hottest.

South of the Tsinling Mountains–Huai River line, the mean January temperature increases progressively, rising from 32° F to 72° F (22° C) on the southern coast of Hainan Island. Snow rarely falls, and the rivers do not freeze. North of this line, the temperature drops from 32° F to −18° F (−28° C) in the northern part of Heilungkiang.

In April the mean temperature is above 32° F for the whole of China with the exception of the extreme north of Heilungkiang. During this time the mean temperature for the Manchurian Plain is between 36° and 46° F (2° and 8° C), and for the extensive plain between Peking and Shanghai it is between 54° and 59° F (12° and 15° C). South of the Nan Mountains the mean temperature is well over 68° F (20° C). Along the coast of southern Kwangtung, willows start to bud in late January, but in Peking the budding of willows comes as late as early April.

In summer the range of temperature between North and South China is quite small. In July the difference in temperature between Canton and Peking is only about 5° F (3° C), and the isotherms in July are roughly parallel to the coastline. In July the isotherm of 82° F (28° C) marks an extensive area. The hottest places in China are to be found along the valleys of the Middle and Lower Yangtze. The mean July temperature of Nan-ch'ang and Ch'ang-sha is well over 84° F (29° C), and in many years it is above 86° F (30° C).

In North China autumn is generally cooler than spring; the mean October temperature in Peking is about 55° F (13° C), and in April the mean temperature is about 57° F (14° C). In South China the reverse is true. The mean October temperature in Canton is 75° F (24° C); in April the mean temperature is only about 70° F (21° C).

The middle and lower reaches of the Huang Ho are the areas where China's civilization and agriculture first developed. There the seasonal rhythm is well marked, and the duration of each season is evenly spaced. In other parts of China, however, there are variations among different regions in the duration, as well as in the starting and closing dates, of each season. In northern Heilungkiang summer is nonexistent, while in southern Kwangtung there is no winter. In K'un-ming, on the Yunnan plateau, the climate is mild throughout the year, and summer and winter are very brief.

In general, south of the Tsinling Mountains–Huai River line the mean daily temperature seldom falls below freezing, so that farming can be practiced all through the year. In the Yangtze Valley two crops a year are usually grown, but north of the Great Wall only one crop a year is grown.

**Rainfall.** Rainfall in China generally decreases from the southeast to the northwest. The annual precipitation of certain places along the southeastern coast amounts to more than 80 inches. The Yangtze Valley receives about 40 inches. Farther north, in the Huai River Valley the annual rainfall decreases to 30 inches. In the lower reaches of the Huang Ho only 20 inches of precipitation are received annually. The Northeast obtains much more precipitation than the North China Plain, where along the Southeast Mountains as much as 40 inches are received.

The southeast monsoon loses much of its moisture by the time it reaches the northern part of the Loess Plateau, where the annual precipitation is reduced to between 12 and 20 inches. Northwest of a line linking the Greater Khingan, Yin, Lang, Ch'i-lien, and A-erh-chin ranges the annual precipitation is less than 10 inches. This is because these regions are far from the sea, and the southern monsoon is prevented from reaching them by high mountains; only grasslands, therefore, are found there. In western Inner Mongolia, the Kansu Corridor, and the Tarim Basin, the annual precipitation is even less than four inches. These are areas of true desert, where sometimes not a single drop of rain is received for several years.

The Dzungarian Basin and the I-li Valley of northern Sinkiang are open to the influences of the westerlies; precipitation there is heavier. The precipitation of the Plateau of Tibet, like that of China as a whole, decreases from southeast to northwest. The valleys in the southeastern part of the plateau receive more than 40 inches of precipitation, while the eastern edge of the plateau receives 20 inches annually. But in the enclosed basin in the north—the Tsaidam—the annual precipitation is only four to 10 inches.

The high variability of rainfall is another characteristic of the climate of China. Usually, the less the rainfall, the greater the variability; this fact has a close connection with the high frequency of drought and flood in China. Spring rain is of immense significance to farming in China; unfortunately, spring is the season with the highest variability. In South China the variability exceeds 40 percent, along the Yangtze it is about 45 percent, and in North China it is more than 50 percent. The variability of a very large area in North China exceeds 70 percent in some places; east of Peking, for example, the rainfall variability in spring may even exceed 80 percent, as it also does in the central parts of the Yunnan–Kweichow highlands.

Rain falls mostly in the summer months, when plants need water the most. This is an important asset for farmers, but the rainfall in summer is usually too intense. In July, when the frontal zone shifts northward, cyclones (circulation of winds around centres of low atmospheric pressure) are much more active in North China than in South China, and North China begins to receive heavier rainfall. More than half of the North China Plain records eight-tenths of an inch of daily rainfall. In some places, as much as an inch or more may fall daily. During this time areas south of the Yangtze are covered by the tropical Pacific air mass, so that the weather becomes comparatively stable, and the amount of rainfall usually decreases, while the average rainfall intensity is less than that of June; a large area receives less than six-tenths of an inch. The intensity of August rainfall is in general less than that of July.

In the southeastern coastal regions, around Fu-chou and Swatow, the maximum daily rainfall may even approach 12 inches. This fact is clearly related to the high frequency of typhoons (tropical cyclones) striking this part of the coast, usually during the period from May to November; July, August, and September are the three months when typhoons are the most frequent.

Typhoons

In May typhoons usually strike the coast south of Swatow. Later in June they shift northward, arriving between Swatow and Wen-chou, and after July they invade areas north of Wen-chou. August has the highest frequency

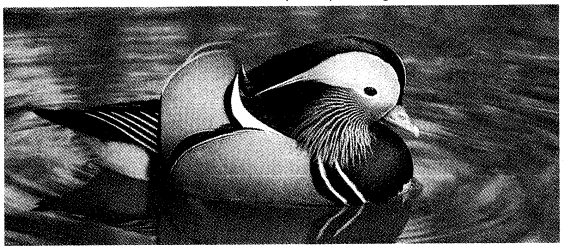
of typhoon invasions, when more than one-third of the typhoons reaching China arrive. After September the frequency of typhoons decreases, and the pattern again shifts southward. In October typhoons usually land south of Wen-chou; the late typhoons arriving in November and December strike south of Swatow.

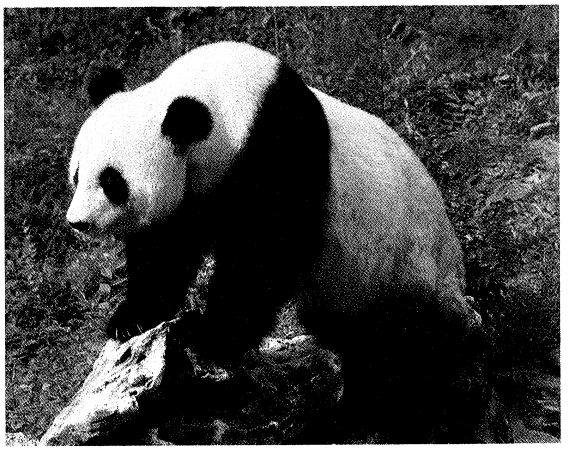
## PLANT AND ANIMAL LIFE

**Plant life.** Types of natural vegetation and their floristic composition are quite diverse. The total number of seed plants in China is about 30,000 species, representing 2,700 genera; more than 200 of these genera are restricted to China. There are about 2,500 species of forest trees, and, among flowering plants, 95 percent of the known woody group are to be found. Many of them are trees of economic importance, such as tung trees, camphor trees, lacquer trees, star anise (which yields an oil used as a flavouring additive), and privet. Contributing to the variety and intermixture of tropical and temperate plants in China are such factors as the lack of insurmountable topographic barriers, such as large stretches of desert, between the tropical, temperate, and semifrigid zones; wind systems that alternate in winter and summer; and the frequent occurrence of cyclones. If the vegetation of Heilungkiang Province in the North and Kwangtung Province in the South are compared, it is hardly possible to find a single common plant species, with the exception of certain weeds. In the taiga of the northern border of China or in the high mountains, however, there are many plant species also found inside the Arctic Circle, while in the Chinese tropics there are species that also grow below the Equator. From the ecological point of view, the tropical forests of South China, however, do not in general differ greatly from those of Indonesia and other Southeast Asian countries, while the desert and steppe vegetation of northwestern China is closely akin to that found in the Mongolian People's Republic or to that in the Kazakh Soviet Socialist Republic. Furthermore, the Chinese taiga of the frontier area near the Soviet Union in no way differs from that of Siberia.

Traveling in China, one may encounter practically all types of natural vegetation indigenous to the Northern Hemisphere, with the exception only of that of the polar tundra. There are mangrove swamps along the shores of the South China Sea; rain forests on Hai-nan Island and in the south of Yunnan; and deserts, steppes, meadows, and savannas, as well as regions where tropical and temperate coniferous, broad-leaved evergreen, and deciduous plants prevail.

(Top) © Gordon Langsbury—Bruce Coleman Inc.,
(bottom) © George Holton—Photo Researchers





(Top) The Mandarin duck of northern China. (Bottom) The endangered giant panda, native only to the bamboo forests of Tibet and Szechwan.

In accordance with the character of the vegetation, China may be divided—roughly along a diagonal from the southwest to the northeast—into two sharply different parts: the dry northwest and the humid southeast. The tropical area, adjoining the humid southeast, is geographically related more to Southeast Asia.

In the northwest, where desert-like conditions prevail, there are vast areas of very sparse drought-resistant vegetation; within these areas in the low-lying land and depressions are patches of salt-tolerant plants, notably in the Dzungarian, Tsaidam, and Gobi regions. Skirting the southern edge of the Gobi is a wide belt of grassland.

(C.-S.Ch.)

**Animal life.** Profusion of vegetation and a variety of relief have fostered the development of a great diversity of animal life and have permitted the survival of animals that elsewhere are extinct. Notable among such survivals are the great paddlefish of the Yangtze, the small species of alligator in eastern and central China, and the giant salamander (related to the Japanese giant salamander and the American hellbender) in western China. The diversity of animal life is perhaps greatest in the ranges and valleys of Tibet and Szechwan, to which region the giant panda is confined. The takin, or goat antelope, numerous species of pheasants, and a variety of laughing thrushes are to be found in all the Chinese mountains. China seems to be one of the chief centres of dispersal of the carp family and also of old-world catfishes.

Rare animal species

The regional affinities of Chinese animal life are complex. In the Northeast there are resemblances to the animal life of the Siberian forests. Animals from Central Asia inhabit suitable steppe areas in northern China. The life of the great mountain ranges is Palearctic (relating to a biogeographic region that includes Europe, Asia north of the Himalayas, northern Arabia, and Africa north of the Sahara) but with distinctively Chinese species or genera. To the southeast the lowlands and mountains alike permit direct access to the eastern region. This part of China presents a complete transition from temperate-zone Palearctic life to the wealth of tropical forms distinctive of southeastern Asia. Tropical types of reptiles, amphibians, birds, and mammals predominate in the southernmost Chinese provinces. (K.P.S.)

## SETTLEMENT PATTERNS

**Rural areas.** Except in the mountains and hills, an overwhelming majority of the rural settlements of China consist of sizable compact (nucleated) villages. The formation of such rural settlements is related not only to the increasing population and to a long historical background but also to water supply (the practice of drilling deep wells, for instance) and to defense (especially, in former days, against the attack of bandits). Many of the large villages have no urban atmosphere at all, even with populations of several thousand. Frequent markets may be held between the settlements to enable the peasants to barter their agricultural produce.

On the North China Plain, villages are fairly evenly distributed and are connected with one another by footpaths and cart tracks. Houses are built close together and are mostly made of sun-dried brick or pounded earth. Many of the market towns or even large villages are surrounded by walls. The number and length of the streets depend on the town's size and the nature of the terrain; some streets are merely narrow lanes.

Rural landscapes of central and southern China are dominated by rice fields. The Yangtze Delta has almost every type of human settlement, from the single farmstead to the fairly large market town. Villages to the south and east of T'ai Lake in Kiangsu Province are located one to two miles apart. Villages in central China, particularly on the Lower Yangtze, are larger than those of North China; many have a few shops that serve not only the villagers but also the dispersed residents nearby. In the centre of dozens of such villages is a market town, which collects

rural produce and distributes manufactured goods. Communication among the villages is mainly by boat, along the dense net of waterways. The most elegant structures in the landscape are the numerous stone bridges that span streams and canals. In the Ch'eng-tu Plain of the Szechwan Basin, a large part of the population lives in isolated farmsteads or scattered hamlets, surrounded by thickets of bamboo and broad-leaved trees.

Cave dwellings

Cave dwellings are another distinctive feature of the Chinese rural landscape. They are common on the Loess Plateau and particularly in northern Shensi, western Shansi, and southeastern Kansu, where the loess is thick and timber is scarce. The cave dwelling has the advantage of being naturally insulated, making it cooler in summer and warmer in winter.

**Urban areas.** Urbanization and industrialization are often closely related. Urbanization in China accelerated after 1953 as the government intensified its efforts to convert the country into an industrial power.

The dramatic change in the urban landscape is the result of the rapid development of modern manufacturing industries and of communications. Many new towns and cities have been built around manufacturing and mining centres. In the remoter areas of China the first appearance of railways and highways contributed to the rapid growth of some entirely new towns, such as Karmo, Shih-ho-tzu, and Sze-chuan-ho. For larger cities, Wu-lu-mu-ch'i (capital of the Uighur Autonomous Region of Sinkiang), Lan-chou (capital of Kansu), and Pao-t'ou (in Inner Mongolia Autonomous Region) provide examples of extremely rapid expansion. Lan-chou lies midway between southeast and northwest China. Pao-t'ou, formerly a bleak frontier town of traders, artisans, and immigrant farmers, has become one of the largest steel centres in China.

## The people

### ETHNIC AND LINGUISTIC GROUPS

China is a multinational country, with a population composed of a large number of ethnic and linguistic groups. Almost all its inhabitants are of Mongoloid stock: thus, the basic classification of the population is not so much ethnic as linguistic. The Han (Chinese), the largest group, outnumber the minority groups or minority nationalities in every province or autonomous region except Tibet and Sinkiang. The Han, therefore, form the great homogeneous mass of the Chinese people, sharing the same culture, the same traditions, and the same written language. Some 55 minority groups are spread over approximately 60 percent of the total area of the country. Where these minority groups are found in large numbers, they have been given some semblance of autonomy and self-government; autonomous regions of several types have been established on the basis of the geographical distribution of nationalities.

Han (Chinese)

The government takes great credit for its treatment of these minorities, including care for their economic well-being, the raising of their living standards, the provision of educational facilities, the promotion of their national languages and cultures, and the raising of their levels of literacy, as well as for the introduction of a written language where none existed previously. In this connection it may be noted that, of the 50-odd minority languages, only 20 had written forms before the coming of the Communists; and only relatively few written languages—for example, Mongolian, Tibetan, Uighur, Kazakh, Tai, and Korean—were in everyday use. Other written languages were used chiefly for religious purposes and by a limited number of persons. Educational institutions for national minorities are a feature of many large cities, notably Peking, Wu-han, Ch'eng-tu, and Lan-chou.

Four major language families are represented in China: the Sino-Tibetan, Altaic, Indo-European, and Austro-Asiatic. The Sino-Tibetan family, both numerically and in the extent of its distribution, is the most important; within this family, Han Chinese is the most widely spoken language. Although unified by their tradition, the written characters of their language, and many cultural traits, the Han speak several mutually unintelligible dialects and display marked regional differences. By far the most important Chinese tongue is the Mandarin, or *p'u-t'ung hua,* meaning "ordinary language" or "common language." There are three variants of Mandarin. The first of these is the northern variant, of which the Peking dialect, or Peking *hua,* is typical and which is spoken to the north of the Tsinling Mountains–Huai River line; as the most widespread Chinese tongue, it has officially been adopted as the basis for a national language. The second is the western variant, also known as the Ch'eng-tu or Upper Yangtze variant; this is spoken in the Szechwan Basin and in adjoining parts of southwest China. The third is the southern variant, also known as the Nanking or Lower Yangtze variant, which is spoken in northern Kiangsu and in southern and central Anhwei. Related to Mandarin are the Hunan, or Hsiang, dialect, spoken by people in central and southern Hunan, and the Kan dialect. The Hui-chou dialect, spoken in southern Anhwei, forms an enclave within the southern Mandarin area.

Less intelligible to Mandarin speakers are the dialects of the southeast coastal region, stretching from Shanghai to Canton. The most important of these is the Wu dialect, spoken in southern Kiangsu and in Chekiang. This is followed, to the south, by the Fu-chou, or Min, dialect of northern and central Fukien and by the Amoy-Swatow dialect of southern Fukien and easternmost Kwangtung. The Hakka dialect of southernmost Kiangsi and northeastern Kwangtung has a rather scattered pattern of distribution. Probably the best known of these southern dialects is Cantonese, which is spoken in central and western Kwangtung and in southern Kwangsi—a dialect area in which a large proportion of overseas Chinese originated.

In addition to the Han, the Manchu and the Hui (Chinese Muslims) also speak Mandarin and use Chinese characters. The Hui are descendants of Chinese who adopted Islām when it penetrated into China in the 7th century. They are intermingled with the Han throughout much of the country and are distinguished as Hui only in the area of their heaviest concentration, the Hui Autonomous Region of Ningsia. Other Hui communities are organized as autonomous prefectures (*tzu-chih-chou*) in Sinkiang and as autonomous counties (*tzu-chih-hsien*) in Tsinghai, Hopeh, Kweichow, and Yunnan. There has been a growing tendency for the Hui to move from their scattered settlements into the area of major concentration, possibly, as firm adherents of Islām, in order to facilitate intermarriage with other Muslims.

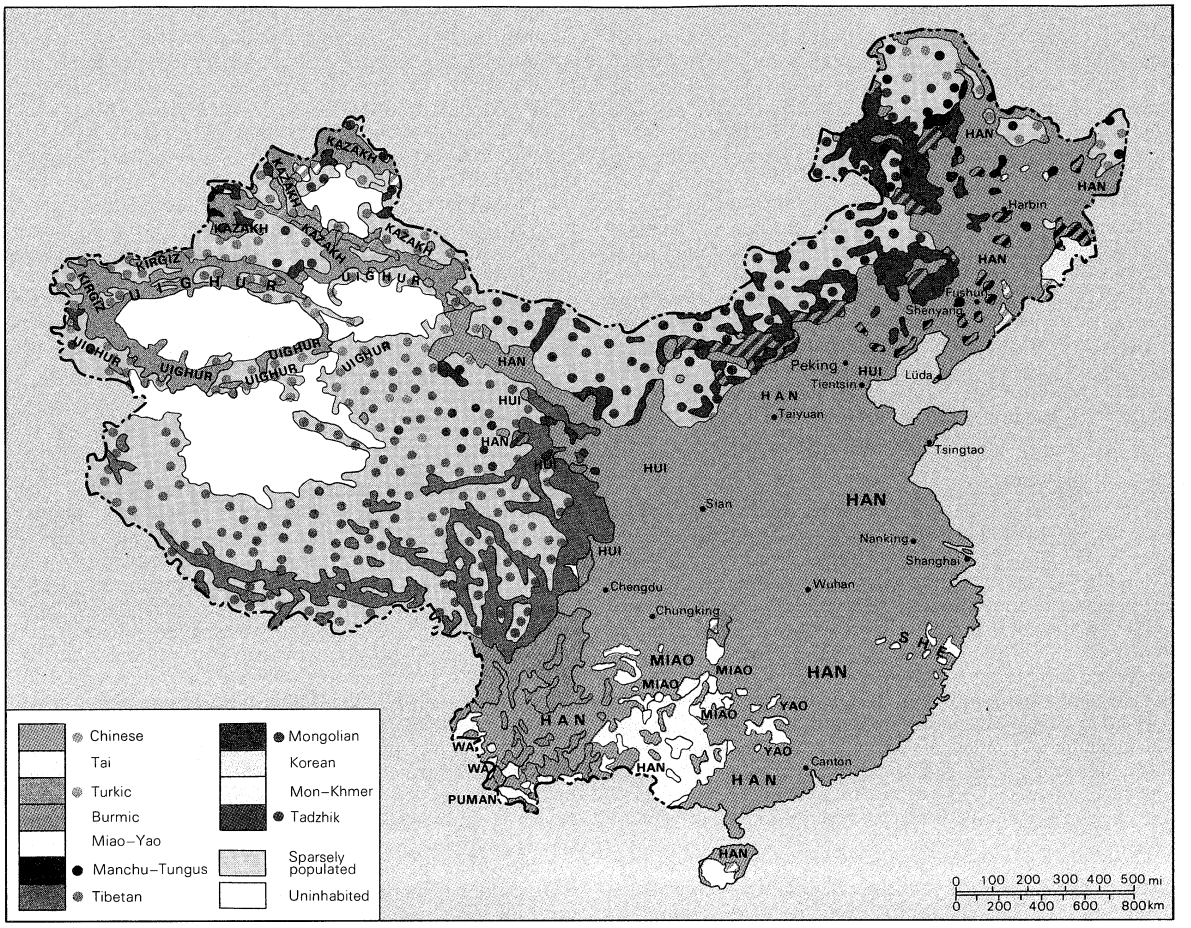
Manchu and Hui

The Manchu declare themselves to be descendants of the Manchu warriors who invaded China in the 17th century and founded the Ch'ing dynasty (1644–1911/12). Ancient Manchu is virtually a dead language, and the Manchu have been completely assimilated into Han Chinese culture. They are found mainly in North China and the Northeast, but they form no separate autonomous areas above the commune level. Some say the Koreans of the Northeast, who form an autonomous prefecture in eastern Kirin, cannot be assigned with certainty to any of the standard language classifications.

The Chuang-chia, or Chuang, are China's largest minority group. Most of them live in the Chuang Autonomous Region of Kwangsi. They are also represented in national autonomous areas in neighbouring Yunnan and Kwangtung. They depend mainly on the cultivation of rice for their livelihood. In religion they are animists, worshiping particularly the spirits of their ancestors. The Puyi (Chung-chia) group are concentrated in southern Kweichow, where they share an autonomous prefecture with the Miao group. The T'ung group are settled in small communities in Kwangsi and Kweichow; they share with the Miao group an autonomous prefecture set up in southeast Kweichow in 1956. The Tai group are concentrated in southern Yunnan and were established in two autonomous prefectures—one whose population is related most closely to the Tai of northern Thailand and another whose Tai are related to the Shan people of Burma. The Li of Hai-nan Island form a separate group of the Chinese-Tai language branch. They share with the Miao people a district in southern Hai-nan.

Chuang-chia

Tibetans are distributed over the entire Tsinghai-Tibetan




General ethnic composition of China.

plateau. Outside Tibet, Tibetan minorities constitute autonomous prefectures and autonomous counties. There are five Tibetan autonomous prefectures in Tsinghai, two in Szechwan, and one each in Yunnan and Kansu. The Tibetans still keep their tribal characteristics, but few of them are nomadic. Though essentially farmers, they also raise livestock and, as with other tribal peoples in the Chinese far west, also hunt to supplement their food supply. The major religion of Tibet has been Tibetan Buddhism since about the 17th century; before 1959 the social and political institutions of this region were still based largely on this faith. Many of the Yi (Lolo) were concentrated in two autonomous prefectures—one in southern Szechwan and another in northern Yunnan. They raise crops and sometimes keep flocks and herds.

The Miao-Yao branch, with their major concentration in Kweichow, are distributed throughout the central south and southwestern provinces and are found also in some small areas in east China. They are subdivided into many rather distinct groupings. Most of them have now lost their traditional tribal traits through the influence of the Han, and it is only their language that serves to distinguish them as tribal peoples. Two-thirds of the Miao are settled in Kweichow, where they share two autonomous prefectures with the T'ung and Puyi groups. The Yao people are concentrated in the Kwangsi–Kwangtung–Hunan border area.

In some areas of China, especially in the southwest, there are many different ethnic groups that are geographically intermixed. Because of language barriers and different economic structures, these peoples all maintain their own cultural traits and live in relative isolation from one another. In some places the Han are active in the towns and in the fertile river valleys, while the minority peoples depend for their livelihood on more primitive forms of agriculture or on grazing their livestock on hillsides and mountains. The vertical distribution of these peoples is in zones—usually the higher they live, the less complex their way of life. In former times they did not mix well with one another, but now, with highways penetrating deep into their settlements, they have better opportunities to communicate with other groups and are also enjoying better living conditions.

The Altaic language family

While the minorities of the Sino-Tibetan language family are thus concentrated in the south and southwest, the second major language family—the Altaic—is represented entirely by minorities in northwestern and northern China. The Altaic family falls into three branches: Turkic, Mongolian, and Manchu-Tungus. The Turkic language branch is by far the most numerous of the three Altaic branches. The Uighur, who are Muslims, form the largest Turkic minority. They are distributed over chains of oases in the Tarim Basin and in the Dzungarian Basin of Sinkiang. They mainly depend on irrigation agriculture for a livelihood. Other Turkic minorities in Sinkiang are splinter groups of nationalities living in Soviet Central Asia, including the Kazakh and Kirgiz. All these groups are adherents of Islām. The Kazakh and Kirgiz are pastoral nomadic peoples, still showing traces of tribal organization. The Kazakh live mainly in northwestern and northeastern Sinkiang as herders, retiring to their camps in the valleys when winter comes; they are established in the I-li-ha-sa-k'o (Ili Kazakh) Autonomous Prefecture. The Kirgiz are high-mountain pastoralists and are concentrated mainly in the westernmost part of Sinkiang.

The Mongolians, who are by nature a nomadic people, are the most widely dispersed of the minority nationalities of China. Most of them are inhabitants of the Inner Mongolia Autonomous Region. Small Mongolian and Mongolian-related groups of people are scattered throughout the vast area from Sinkiang through Tsinghai and Kansu and into the provinces of the Northeast (Kirin, Heilungkiang, and Liaoning). In addition to the Inner Mongolia Autonomous Region, the Mongolians are established in two autonomous prefectures in Sinkiang, a joint autonomous prefecture with Tibetans and Kazakh in Tsinghai, and

several autonomous counties in the western area of the Northeast. Some of them retain their tribal divisions and are pastoralists, but large numbers of Mongolians engage in sedentary agriculture, and some of them combine the growing of crops with herding. The tribes, who are dependent upon animal husbandry, travel each year around the pastureland—grazing sheep, goats, horses, cattle, and camels—and then return to their point of departure. A few take up hunting and fur trapping in order to supplement their income. The Mongolian language consists of several dialects, but in religion it is a unifying force; most Mongolians are believers in Tibetan Buddhism.

A few linguistic minorities in China belong to neither the Sino-Tibetan nor the Altaic language family. The Tadzhik of westernmost Sinkiang are related to the population of the Tadzhik Soviet Socialist Republic and belong to the Iranian branch of the Indo-European family. The Kawa people of the China–Burma border area belong to the Mon-Khmer branch of the Austro-Asiatic family.

### POPULATION GROWTH

Historical records show that, as long ago as 800 BC, in the early years of the Chou dynasty, China was already inhabited by about 13,700,000 people. Until the last years of the Hsi (Western) Han dynasty, about AD 2, comparatively accurate and complete registers of population were kept, and the total population in that year was given as 59,600,000. This first Chinese census was intended mainly as a preparatory step toward the levy of a poll tax. Many members of the population, aware that a census might work to their disadvantage, managed to avoid reporting; this explains why all subsequent population figures were unreliable until 1712. In that year the Emperor declared that an increased population would not be subject to tax; population figures thereafter gradually became more accurate.

The census of AD 2

During the later years of the Pei (Northern) Sung dynasty, in the early 12th century, when China was already in the heyday of its economic and cultural development, the total population began to exceed 100,000,000. Later, uninterrupted and large-scale invasions from the north reduced the country's population. When national unification returned with the advent of the Ming dynasty, the census was at first strictly conducted. The population of China, according to a registration compiled in 1381, was quite close to the one registered in AD 2.

From the 15th century onward, the population increased steadily; this increase was interrupted by wars and natural disasters in the mid-17th century and slowed by the internal strife and foreign invasions in the century that preceded the Communist takeover in 1949. During the 18th century China enjoyed a lengthy period of peace and prosperity, characterized by continual territorial expansion and an accelerating population increase. In 1762 China had a population of more than 200,000,000, and by 1834 the population had doubled. It should be noted that during this period there was no concomitant increase in the amount of cultivable land; and from this time on, land hunger became a growing problem.

After 1949 sanitation and medical care greatly improved, epidemics were brought under control, and the younger generation became much healthier. Public hygiene also improved, resulting in a death rate that declined faster than the birthrate and a rate of population growth that speeded up again. Population reached 1,000,000,000 in the early 1980s.

The continually growing population poses major problems for the government. Faced with difficulties in obtaining an adequate food supply and in combating the generally low standard of living, the authorities sponsored a drive for birth control in 1955–58. A second attempt at population control began in 1962, when advocacy of late marriages and the use of contraceptives became prominent parts of the program. The outbreak of the Cultural Revolution interrupted this second family-planning drive, but in 1970 a third and much stricter program was initiated. This program attempted to make late marriage and family limitation obligatory, and it culminated in 1979 in efforts to implement a policy of one child per family.

Drive for birth control

Other developments affected the rate of population growth more than the first two official family-planning campaigns. For example, although family planning had been rejected by Chinese Communist Party Chairman Mao Zedong (Mao Tse-tung) in 1958, the Great Leap Forward that he initiated in that year (see below *The economy*) caused a massive famine that resulted in more deaths than births and a reduction of population in 1960. By 1963 recovery from the famine produced the highest rate of population increase since 1949, at more than 3 percent, although the second birth-control campaign had already begun.

Since the initiation of the third family-planning program in 1970, however, state efforts have been much more effective. China's population growth rate is now unusually low for a developing country, although the huge size of its population still results in a large annual net population growth.

### POPULATION DISTRIBUTION

Because of complex natural conditions, the population of China is quite unevenly distributed. Population density varies strikingly, with the greatest contrast occurring between the eastern half of China and the lands of the west and the northwest. Exceptionally high population densities occur in the Yangtze Delta, in the Pearl River Delta, and on the Ch'eng-tu Plain of the western Szechwan Basin. Most of the high-density areas are coterminous with the alluvial plains on which intensive agriculture is centred.

In contrast, the isolated, extensive western and frontier regions, which are almost as large as the whole of Europe, are sparsely populated. Extensive uninhabited areas include the extremely high northern part of Tibet, the sandy wastes of the central Tarim and eastern Dzungarian basins in Sinkiang, and the barren desert and mountains east of Lop Nor.

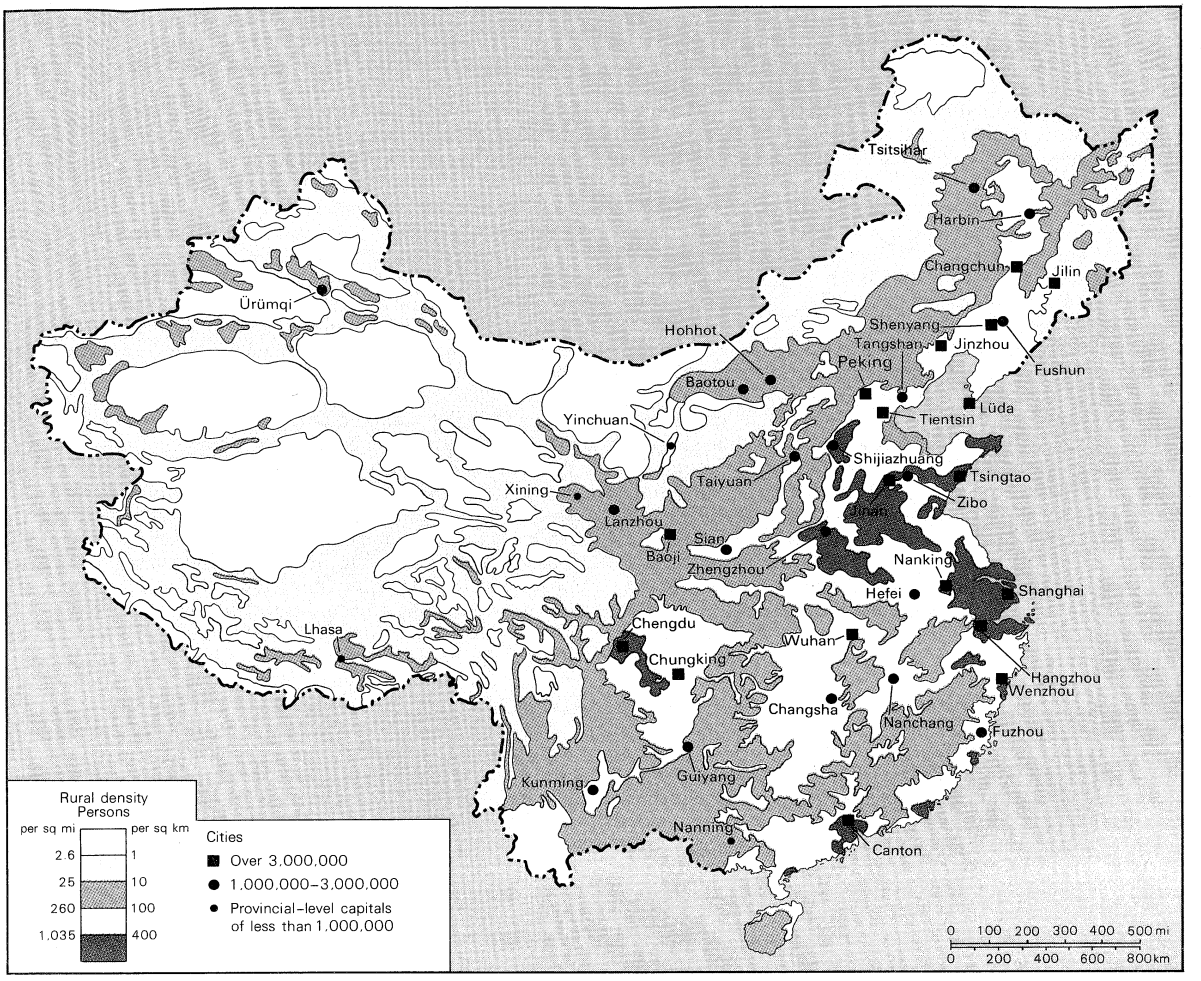
In the 1950s the government became increasingly aware of the importance of the frontier regions and initiated a drive for former members of the military and young intellectuals to settle there. Consequently, the population has increased, following the construction of new railways and highways that traverse the wasteland; a number of small mining and industrial towns have also sprung up.

### INTERNAL MIGRATION

Migrations have occurred often throughout the history of China. Sometimes they took place because a famine or political disturbance would cause the depopulation of an area already intensively cultivated, after which people in adjacent crowded regions would move in to occupy the deserted land. Sometime between 1640 and 1646 a peasant rebellion broke out in Szechwan, and there was a great loss of life. People from Hupeh and Shensi then entered Szechwan to fill the vacuum, and the movement continued until the 19th century. Again, during the middle of the 19th century, the Taiping Rebellion caused another large-scale disruption of population. Many people in the Lower Yangtze were massacred by the opposing armies, and the survivors suffered from starvation. After the defeat of the rebellion, people from Hupeh, Hunan, and Honan moved into the depopulated areas of Kiangsu, Anhwei, and Chekiang, where farmland was lying uncultivated for want of labour. Similar examples are provided by the Nien Rebellion in the Huai River region in the 1850s and '60s, the Muslim rebellions in Shensi and Kansu in the 1860s and '70s, and the great Shensi and Shansi famine of 1877–78.

In modern history the domestic movement of the Han to Manchuria (now known as the Northeast) is the most significant. Even before the establishment of the Ch'ing dynasty in 1644, Manchu soldiers launched raids into North China and captured Han labourers, who were then obliged to settle in Manchuria. In 1668 the area was closed to further Han migration by an Imperial decree, but this ban was never effectively enforced. By 1850, Han settlers had secured a position of dominance in their colonization of Manchuria. The ban was later partially lifted, partly because the Manchu rulers were harassed by disturbances among the teeming population of China proper and partly

Migration to Manchuria




Population density of China.
By courtesy of the Central Intelligence Agency

because the Russian Empire time and again tried to invade sparsely populated and thus weakly defended Manchuria. The ban was finally removed altogether in 1878, but settlement was encouraged only after 1900.

The influx of people into Manchuria was especially pronounced after 1923, and incoming farmers rapidly brought a vast area of virgin prairie under cultivation. About two-thirds of the immigrants entered Manchuria by sea, and one-third came overland. Because of the severity of the winter weather, migration in the early stage was highly seasonal, usually starting in February and continuing through the spring. After the autumn harvest a large proportion of the farmers returned south. As Manchuria developed into the principal industrial region of China, however, large urban centres arose, and the nature of the migration changed. No longer was the movement primarily one of agricultural resettlement; instead it became essentially a rural-to-urban movement of interregional magnitude.

Post-1949 migration

After 1949 the new government's efforts to foster planned migration into interior and border regions produced noticeable results. Although the total number of people involved in such migrations is not known, it has been estimated that by 1980 about 25 to 35 percent of the population of such regions and provinces as Inner Mongolia, Sinkiang, Heilungkiang, and Tsinghai consisted of recent migrants, and migration had raised the percentage of Han in Sinkiang from about 10 to 40 percent of the total. Efforts to control the growth of large cities led to the resettlement of 20,000,000 urbanites in the countryside after the failure of the Great Leap Forward and of 17,000,000 urban-educated youths in the decade after 1968. Within the next decade, however, the majority of these "rusticated youths" were allowed to return to the cities, and new migration from rural areas pushed urban population totals upward once again. (C.-S.Ch./K.G.L.)

## The economy

Despite China's size, the wealth of its resources, and the fact that more than 20 percent of the world's population lives within its borders, its role in the world economy traditionally has been relatively small. Since the late 1970s, however, when China decided to increase its interaction with the international economy, its role in world trade has steadily grown and its importance to the international economy has also increased apace. China's foreign trade has since grown faster than its gross national product (GNP). The government's decision to permit China to be used by Western firms as an export platform may eventually make the country a competitive threat to its neighbours South Korea, Singapore, and Malaysia.

The Chinese economy thus has been in a state of transition since the late 1970s as the country has moved away from a Soviet-type economic system. Agriculture has been decollectivized, the small nonagricultural private sector has grown rapidly, and government priorities have shifted toward light, rather than heavy, industry. Nevertheless, key bottlenecks continue to constrain growth. Available energy is sufficient to run less than 80 percent of installed industrial capacity, the transport system is inadequate to move sufficient quantities of such critical items as coal, and the communications system cannot meet the needs of a centrally planned economy of China's size and complexity.

China's underdeveloped transport system—combined with important differences in the availability of natural and human resources and in industrial infrastructure—has produced significant variations in the regional economies of China. The three wealthiest regions are along the southeast coast, centred on the Pearl River Delta; along the east coast, centred on the Lower Yangtze River; and near the

Po Hai (Gulf of Chihli), in the Peking–Tientsin–Liaoning region. It is the rapid development of these areas that is expected to have the most significant effect on the Asian regional economy as a whole, and Chinese government policy is designed to remove the obstacles to accelerated growth in these wealthier regions.

China is the world's largest producer of rice and is among the principal sources of wheat, corn (maize), tobacco, soybeans, peanuts (groundnuts), and cotton. The country is one of the world's largest producers of a number of industrial and mineral products—including cotton cloth, tungsten, and antimony—and is an important producer of cotton yarn, coal, crude oil, and a number of other products. Its mineral resources are probably among the richest in the world but are only partially developed.

Technological level

Although China has acquired some highly sophisticated production facilities through trade with Western countries—and also has built a number of advanced engineering plants capable of manufacturing an increasing range of sophisticated equipment, including nuclear weapons and Earth satellites—most of its industrial output still comes from relatively backward and ill-equipped factories. The technological level and quality standards of its industry as a whole are still fairly low.

Other major problems concern the labour force and the pricing system. There is large-scale underemployment in both urban and rural areas, and the fear of the disruptive effects of major, explicit unemployment is strong. The prices of key commodities, especially of industrial raw materials and major industrial products, are determined by the state. In most cases, basic price ratios were set in the 1950s and are often irrational in terms of current production capabilities and demands. China's increasing contact with the international economy and its growing efforts to use market forces to govern the domestic allocation of goods have exacerbated this problem. Over the years, large subsidies were built into the price structure, and these subsidies grew substantially in the late 1970s and '80s.

## RESOURCES

**Mineral resources.** China is well endowed with mineral resources, the most important of which is coal. Although deposits are widely scattered (some coal is found in every province), most of the total is located in the northern part of the country. The province of Shansi, in fact, is thought to contain about half of the total; other important coal-bearing provinces include Heilungkiang, Liaoning, Kirin, Hopeh, and Shantung. Apart from these northern provinces, significant quantities of coal are present in Szechwan, and there are some deposits of importance in Kwangtung, Kwangsi, Yunnan, and Kweichow. A large part of the country's reserves consists of good bituminous coal, but there are also large deposits of lignite. Anthracite is present in several places (especially Liaoning, Kweichow, and Honan), but overall it is not very significant.

In order to ensure a more even distribution of coal supplies and to reduce the strain on the less than adequate transport network, the authorities have pressed for the development of a large number of small, locally run mines throughout the country. This campaign was energetically pursued after the 1960s, with the result that thousands of small pits have been established, and they produce more than half the country's coal. This output, however, is typically expensive and is used for local consumption.

Oil and natural gas

China's onshore oil resources are located in the Northeast and in Sinkiang, Kansu, Tsinghai, Szechwan, Shantung, and Honan provinces. Shale oil is found in a number of places, especially at Fu-shun in Liaoning, where the deposits overlie the coal reserves, as well as in Kwangtung. Light oil of high quality has been found in the Pearl River estuary of the South China Sea, the Tsaidam Basin in Tsinghai, and the Tarim Basin in Sinkiang. China contracted with Western oil companies to jointly explore and develop oil deposits in the China Sea, the Yellow Sea, the Gulf of Tonkin, and the Po Hai. The country consumes most of its oil output but does export some crude oil and oil products.

The extent of China's natural gas reserves is unknown, as relatively little exploration for natural gas has been done. Szechwan Province accounts for almost half of the known natural gas reserves and production. Most of the rest of China's natural gas is associated gas produced in the Northeast's major oil fields, especially Ta-ch'ing. Other gas deposits have been found in the Tsaidam Basin, Hopeh, Kiangsu, Shanghai, and Chekiang, and offshore to the southwest of Hai-nan Island.

Iron ore is found in most provinces, and there are reserves on Hai-nan Island. Kansu, Kweichow, southern Szechwan, and Kwangtung provinces have rich deposits. The largest mined reserves are located north of the Yangtze River and supply neighbouring iron and steel enterprises. With the exception of nickel, chromium, and cobalt, China is well supplied with ferroalloys and manganese. Reserves of tungsten are also known to be fairly large. Copper resources are moderate, and high-quality ore is present only in a few deposits. Discoveries have been reported from the Hui Autonomous Region of Ningsia. Lead and zinc are available, and bauxite resources are thought to be plentiful. China's antimony reserves are the largest in the world. Tin resources are plentiful, and there are fairly rich deposits of gold. There are important deposits of phosphate rock in a number of areas. Pyrites occur in several places; Liaoning, Hopeh, Shantung, and Shansi have the most important deposits. China also has large resources of fluorite (fluorspar), gypsum, asbestos, and cement.

China also produces a fairly wide range of nonmetallic minerals. One of the most important of these is salt, which is derived from coastal evaporation sites in Kiangsu, Hopeh, Shantung, and Liaoning, as well as from extensive salt fields in Szechwan, Ningsia, and the Tsaidam Basin.

**Hydroelectric resources.** In view of China's extensive river network and mountainous terrain, there is ample potential for the production of hydroelectric power. Most of the total hydroelectric capacity is in the southwest, where coal supplies are poor but demand for energy is rapidly growing. The potential in the Northeast is fairly small, but it was there that the first hydroelectric stations were built (by the Japanese). As a result of considerable seasonal fluctuations in rainfall, the flow of rivers tends to drop during the winter, forcing many power stations to operate at less than normal capacity, while in the summer, on the other hand, floods often interfere with production.

Hydroelectric stations

Thus, while China has rich overall energy potential, most remains to be developed. In addition, the geographical distribution of energy places most of these resources far from their major industrial users. Basically the Northeast is rich in coal and oil, the central part of North China has abundant coal, and the southwest has great hydroelectric potential. But the industrialized regions around Canton and the Lower Yangtze region around Shanghai have too little energy, while there is very little industry located near major energy resource areas other than in the southern part of the Northeast.

## AGRICULTURE

**Farming and forestry.** As a result of topographic and climatic features, the area suitable for cultivation is relatively small: only about 10 percent of China's total land area. Of this, slightly more than half is unirrigated, and the remainder is divided roughly equally between paddy fields and irrigated areas. Nevertheless, about 80 percent of the population lives in the countryside, and until the 1980s a high percentage of them made their living directly from farming. Since then, many have been encouraged to leave the fields and pursue other activities, such as handicrafts, commerce, and transport; and by the mid-1980s farming accounted for less than half of the value of rural output.

The quality of the soil varies. Environmental problems such as floods, drought, and erosion pose serious threats in many parts of the country. The wholesale destruction of forests gave way to an energetic reforestation program that proved inadequate, and forest resources are still fairly meagre. The principal forests are found in the Tsinling Mountains and the central mountains and on the Szechwan–Yunnan plateau. Because they are inaccessible, the Tsinling forests are not worked extensively, and much of the country's timber comes from Heilungkiang, Kirin, Szechwan, and Yunnan.



Harvesting rice in a paddy field in the Ch'eng-tu Plain, Szechwan Province.
Peter Carmichael—Aspect Picture Library, London

Western China, comprising Tibet, Sinkiang, and Tsinghai, has little agricultural significance except for areas of oasis farming and cattle raising. Rice, China's most important crop, is dominant in the southern provinces, many of which yield two harvests a year. In the North wheat is of the greatest importance, while in central China wheat and rice vie with each other for the top place. Millet and kaoliang (a variety of grain sorghum) are grown mainly in the Northeast and some central provinces, which—together with some northern areas—also provide considerable quantities of barley. Most of the soybean crop is derived from the North and the Northeast; corn (maize) is grown in the centre and the North, while tea comes mainly from the hilly areas of the southeast. Cotton is grown extensively in the central provinces, but it is also found to a lesser extent in the southeast and in the North. Tobacco comes from the centre and parts of the South. Other important crops are potatoes, sugar beets, and oilseeds.

Farm machinery

There is still a lack of farm machinery; for the most part the Chinese peasant depends on simple, nonmechanized farming implements. Good progress has been made in increasing water conservancy, and about half the cultivated land is under irrigation.

**Livestock and fishing.** Animal husbandry constitutes the second most important component of agricultural production. China is the world's leading producer of pigs, chickens, and eggs, and it also has sizable herds of sheep and cattle. Since the mid-1970s, greater emphasis has been placed on increasing the livestock output.

China has a long tradition of ocean and freshwater fishing and of aquaculture. Pond raising has always been important and has been increasingly emphasized to supplement coastal and inland fisheries threatened by overfishing and to provide such valuable export commodities as prawns.

## INDUSTRY

The development of industry has been given considerable attention since the advent of the Communist regime. Overall industrial output has grown at a rate of more than 10 percent per year, and China's industrial work force probably exceeds the combined total for all other developing countries. Industry has surpassed all other sectors in economic growth and degree of modernization.

Among the various industrial branches the metallurgical and machine-building industries have received the highest priority. These two branches alone now account for about one-third of the total gross value of industrial output. In these, as in most other areas of industry, however, innovation has generally suffered at the hands of a system that has rewarded increases in gross output rather than improvements in variety and quality. China, therefore, still imports significant quantities of specialized steels. Much of the country's steel output comes from a small number of producing centres, the largest being An-shan in Liaoning.

The principal preoccupation of authorities in the chemical industry is to expand the output of chemical fertilizers, plastics, and synthetic fibres. The growth of this industry has placed China among the world's leading producers of nitrogenous fertilizers. In the consumer goods sector the main emphasis is on textiles and clothing, which also form an important part of China's exports. Textiles, a rapidly growing proportion of which consists of synthetics, account for about 15 percent of the gross industrial output. The industry tends to be scattered throughout the country, but there are a number of important textile centres, including Shanghai, Canton, and Harbin.

Energy production has increased rapidly, but it still falls considerably short of demand. This is partly due to artificial energy prices that have been held so low that industries have had few incentives to conserve. Coal provides about 70 percent of China's energy consumption. Petroleum production, which began growing rapidly from an extremely low base in the early 1960s, has basically remained at the same level since the late 1970s. There are large petroleum reserves in the inaccessible northwest and potentially significant offshore petroleum deposits, but about half of the country's oil production still comes from the major Ta-ch'ing oil field in the Northeast. China has much, and mostly untapped, hydroelectric power potential and natural gas reserves of unknown extent. The government has made plans to develop nuclear power plants in the Shanghai and Canton regions.

Energy resources

Overall, the distribution of industry remains very uneven, despite serious efforts from the mid-1950s to the late 1970s to build up industry in the interior at the cost of the major cities on the east coast. While percentage growth of industry in the interior provinces generally greatly exceeded that of the coastal areas, the far larger initial industrial base of the latter meant that a few coastal regions continued to dominate China's industrial economy. Thus, Shanghai alone produces almost 10 percent of China's gross value

Greenhill—Black Star



Rolling mill at an iron and steel works in An-shan, Liaoning Province.

of industrial output, and the east coast accounts for about 60 percent of the national industrial output.

### FINANCE

As in most other Communist countries, financial institutions are owned by the state. The principal instruments of fiscal and financial control are the People's Bank of China and the Ministry of Finance, both subject to the authority of the State Council. The People's Bank, which replaced the Central Bank of China in 1950 and gradually took over private banks, fulfills many of the functions of Western central and commercial banks. It issues the currency, controls circulation, and plays an important role in disbursing budgetary expenditures. Furthermore, it handles the accounts, payments, and receipts of government organizations and other bodies, which enables it to exercise detailed supervision over their financial and general performance in the light of the state's economic plans.

The People's Bank is also responsible for foreign trade and other overseas transactions (including remittances by overseas Chinese), but these functions are exercised through the Bank of China, which maintains branch offices in a number of European and Asian countries.

Other important financial institutions include the People's Construction Bank of China, responsible for capitalizing a portion of overall investment and for providing capital funds for certain industrial and construction enterprises; the Industrial and Commercial Bank of China, which conducts ordinary commercial transactions and acts as a savings bank for the public; the Agricultural Bank of China, which serves the agricultural sector; and the China Investment Bank, which handles foreign investment. A number of foreign banks maintain offices in China's larger cities and the special economic zones.

China's economic reforms have greatly increased the economic role of the banking system. Whereas virtually all investment capital was previously provided on a grant basis in the state plan, policy has shifted to a loan basis through the various state financial institutions. More generally, increasing amounts of funds are made available through the banks for economic purposes. Enterprises and individuals can go to the banks to obtain loans outside of the state plan, and this has proved to be a major source of financing both for new firms and for the expansion and modernization of older enterprises.

Foreign capital

Foreign sources of capital are also becoming important for new construction. China has received loans from the World Bank and several United Nations programs, as well as from countries (particularly Japan) and, to a smaller extent, commercial banks. The Chinese government has floated bonds both at home and abroad.

### TRADE

Trade constitutes a fairly small portion of China's overall economy. It became important, however, in the country's attempt to modernize its economy, especially before 1985, when China consistently ran trade surpluses. The direction of China's foreign trade has undergone marked changes since the early 1950s. In 1950 more than 70 percent of the total was accounted for by trade with the non-Communist world, but by 1954—the year after the end of the Korean War—the situation was completely reversed, and the Communist share stood at about 74 percent. During the next few years, the Communist world lost some of its former importance, but it was only after the Sino-Soviet breach of 1960, which resulted in the cancellation of Soviet credits and the withdrawal of Soviet technicians, that the non-Communist world began to see a rapid improvement in its position. In 1965 China's trade with fellow socialist countries made up only some 30 percent of the total.

A significant part of China's trade with the developing countries has been financed through credits, grants, and other forms of assistance. At first, from 1953 to 1955, aid went mainly to North Korea and North Vietnam and some other Communist states; but from the mid-1950s large amounts—mainly grants and long-term, interest-free loans—were promised to politically uncommitted developing countries. The principal efforts were made in Asia—especially to Indonesia, Burma (now Myanmar), Pakistan, and Ceylon (now Sri Lanka)—but large loans were also granted in Africa (Ghana, Algeria, Tanzania) and in the Middle East (Egypt). After Mao Zedong's (Mao Tse-tung's) death in 1976, however, the Chinese scaled back their efforts.

By the 1980s China's foreign trade had come full cycle: trade with all Communist countries was less than 10 percent of China's total foreign trade, more than 70 percent was transacted with non-Communist developed countries, and the remainder was with non-Communist developing countries. In general, China has run a significant trade surplus with developing countries and a trade deficit with developed countries.

China's imports

The vast majority of China's imports consists of industrial supplies and capital goods. The remainder is made up of foodstuffs and consumer durables. The majority of each category of these goods comes from the non-Communist developed countries, primarily Japan and the United States. (Hong Kong, China's second largest trading partner after Japan, is considered here as a developing country.) Regionally, about half of China's imports come from countries in East and Southeast Asia; somewhat more than half of China's exports go to the same countries.

About 70 percent of China's exports consist of manufacturing goods, of which textiles and clothing are by far the most important. Agricultural products, most of which are sent to Hong Kong, account for another quarter of exports, with the remainder being made up of exports of extractive products, including both coal and oil.

### ADMINISTRATION OF THE ECONOMY

**The role of the government.** China is a socialist country, and the government plays a predominant role in the economy. In the industrial sector, for example, the state owns outright firms that produce more than 60 percent of the gross value of industrial output, and all but a small portion of the remainder is owned collectively. In the urban sector the government has set the prices for key commodities, determined the level and general distribution of investment funds, prescribed output targets for major enterprises and branches, allocated energy resources, set wage levels and employment targets, run the wholesale and retail networks, and controlled financial policy and the banking system. The foreign trade system became a government monopoly in the early 1950s. In the countryside, from the mid-1950s, the government prescribed cropping patterns, set the level of prices, and fixed output targets for all major crops.

By the 1980s much of the above system was in the process of changing, as the role of the central government in managing the economy was reduced and the role of both private initiative and market forces increased. Nevertheless, the government continued to dominate the urban economy, and its policies on such issues as agricultural procurement still exerted a major influence on performance in the rural sector.

The effective exercise of control over the economy requires an army of bureaucrats and a highly complicated chain of command, stretching from the top down to the level of individual enterprise. The Communist Party reserves the right to make broad decisions on economic priorities and policies, but the government apparatus headed by the State Council assumes the major burden of running the economy. The State Planning Commission and the Ministry of Finance also are concerned with the functioning of virtually the entire economy.

The entire planning process involves a great deal of consultation and negotiation. The great advantage of including a project in an annual plan is that the raw materials, labour, financial resources, and markets are guaranteed by directives that have the force of law. In fact, however, a great deal of economic activity goes on outside of the scope of the detailed plan, and the tendency has been for the plan to become narrower rather than broader in scope.

Types of economic activity

There are three types of economic activity in China—those stipulated by mandatory planning, those done according to indicative planning (in which central planning of economic outcomes is indirectly implemented), and those governed by market forces. The second and third

categories have grown at the expense of the first, but goods of national importance and almost all large-scale construction come under the mandatory planning system. The market economy generally involves small-scale or highly perishable items that circulate within local market areas only. Almost every year brings additional changes in the lists of goods that fall under each of the three categories.

Operational supervision over economic projects has devolved primarily to provincial, municipal, and county governments. In addition, enterprises themselves are gaining increased independence in a range of activity. Overall, therefore, the Chinese industrial system contains a complex mixture of relationships. In general, the State Council exercises relatively tight control over resources deemed to be of core importance for the performance of the entire economy. Less key aspects of the system are devolved to lower levels for detailed decisions and management. In all spheres, moreover, the need to coordinate units that are in different bureaucratic hierarchies produces a great deal of informal bargaining and the building of consensus.

Although the state controlled agriculture in the 1950s and '60s, rapid changes were made in the system from the late 1970s. The major vehicles for dictating state priorities—the people's communes and their subordinate teams and brigades—have been either abolished or vastly weakened. Peasant incentives have been raised both by price increases for state-purchased agricultural products and by permission to sell excess production on a free market. Greater freedom is permitted in the choice of what crops to grow, and peasants are allowed to contract for land that they will work, rather than simply working most of the land collectively. The system of procurement quotas (fixed in the form of contracts) is being phased out, although the state can still buy farm products and control surpluses in order to affect market conditions.

From the 1950s to the '80s the central government's revenues derived chiefly from the profits of the state enterprises, which were remitted to the state. Some government revenues also came from taxes, of which the most important was the general industrial and commercial tax. The trend, however, has been for remitted profits of the state enterprises to be replaced with taxes on those profits. Initially, this tax system was adjusted so as to allow for differences in the capitalization and pricing situations of various firms, but eventually more uniform tax schedules were to be enforced.

**Trade unions.** Chinese trade unions are organized on a broad industrial basis. Membership is open to those who rely on wages for the whole or a large part of their income—a qualification that excludes most agricultural workers. In theory, membership is not compulsory, but in view of the unions' role in the distribution of social benefits, the economic pressure to join is considerable. The lowest unit is the enterprise union committee. Individual trade unions also operate at the provincial level, and there are trade union councils that coordinate all union activities within a particular area and operate at county, municipal, and provincial levels. At the top of the movement is the All-China Federation of Trade Unions, which discharges its functions through a number of regional federations.

Authority of the unions

In theory, the appropriate trade union organizations are consulted on the level of wages as well as on wage differentials, but in practice their role in these and similar matters is insignificant. They do not engage in collective bargaining—not at all surprising, since their principal duties include assisting the party and promoting production. In fulfilling these tasks, they have a role in enforcing labour discipline. From the point of view of the membership, the most important activities concern the social and welfare services. Thus, it is the unions that look after industrial safety; organize social and cultural activities; provide services such as clinics, rest and holiday homes, hostels, libraries, and clubs; and administer old-age pensions, workers' insurance, disability benefits, and other welfare schemes.

**Economic policies.** In the 1950s and '60s a number of far-reaching changes occurred in China's economic policies and priorities. During the First Five-Year Plan period (1953–57), emphasis was placed on rapid industrial development, partly at the expense of other sectors of the economy. The bulk of the state's investment was channeled into the industrial sector, while agriculture, which occupied more than 80 percent of the economically active population, was forced to rely on its own meagre capital resources for a substantial part of its fund requirements. Within industry, iron and steel, electric power, coal, heavy engineering, building materials, and basic chemicals were given first priority; in accordance with Soviet practice, the aim was to construct large, sophisticated, and highly capital-intensive plants.

This program could not be financed out of domestic resources, and a large number of the new plants were built with Soviet technical and financial assistance. The policy led to a rapid growth in heavy industry, but a few months after the introduction of the Second Five-Year Plan in 1958—which was to be on the same lines as its predecessor—the policy of the Great Leap Forward was announced. In agriculture, this involved the formation of communes, the abolition of private plots, and the increasing of output through greater cooperation and greater physical effort. In industry, the construction of large plants was to continue; but it was to be supplemented by a huge small-industry drive, making use of a large number of small, simple, locally built and run plants. The Chinese peasant, however, was not ready for the communes, and a spectacular drop in agricultural production ensued. Meanwhile, the indiscriminate backyard production drive failed to achieve the desired effects and yielded large quantities of expensively produced, substandard goods. These difficulties were aggravated by the withdrawal of Soviet aid and technicians, who made a point of taking their blueprints with them. In consequence, by late 1960 the country faced an economic crisis of the first order.

Great Leap Forward

The response of the authorities was a complete about-face in policy. Private plots were restored, the size of the communes was reduced, and greater independence was given to the production team. There was also a mass transfer of the unemployed from industry to the countryside, and industrial investment was temporarily slashed in order to free resources for farm production. This policy, which led to an immediate improvement in the agricultural situation, was maintained until 1963, when it again became possible to redirect some resources to the capital goods industry. As a result, industrial production and construction gathered some momentum, but care was taken to avoid the earlier mistake of sacrificing food production to iron and steel and similar industries. Then, in 1966 the "Great Proletarian Cultural Revolution" began. Unlike the Great Leap, the Cultural Revolution did not have an explicit economic philosophy. Nevertheless, industrial production was badly affected by the ensuing confusion and strife.

The Cultural Revolution left some difficult legacies for the Chinese economy. In industry, wages had been frozen and bonuses canceled. Combined with the policies of employing more workers than necessary to soak up unemployment and of never firing workers once hired, this action essentially eliminated incentives to work hard. In addition, technicians and many managers lost their authority and could not play an effective role in production in the wake of the movement. The entire urban system, moreover, provided less than minimal incentives to achieve efficiency in production. While overall output continued to grow, capital–output ratios declined. In agriculture, per capita output in 1977 was no higher than in 1957.

Post-Mao economic reforms

Post-Mao rural economic reform began with major price increases for agricultural products in 1979. By 1981 the emphasis had shifted to breaking up collectively tilled fields into land that was contracted out to private families to work. During this time, the size of private plots (land actually owned by individuals) was increased, and most restrictions on selling agricultural products in free markets were lifted. In 1984 much longer-term contracts for land were encouraged (generally 15 years or more), and the concentration of land through subleasing of parcels was made legal. In 1985 the government announced that it would dismantle the system of planned procurements with state-allocated production quotas in agriculture. Peasants who had stopped working the land were encouraged to

find private employment in the countryside or in small towns. They did not obtain permission to move to major cities, however.

The basic thrusts of urban economic reform have been toward integrating China more fully with the international economy; making enterprises responsible for their profits and losses; reducing the state's role in directing, as opposed to guiding, the allocation of resources; shifting investment somewhat away from the metallurgical and machine-building industries and toward light industry, while retaining an emphasis on resolving the energy, transportation, and communications bottlenecks; creating material incentives for individual effort and a consumer ethos to spur people to work harder; rationalizing the pricing structure; and putting individuals into jobs for which they have specialized training, skills, or talents. At the same time, the state has permitted a private sector to develop and has allowed it to compete with state firms in a number of service areas and, increasingly, in such larger-scale operations as construction.

A number of related measures were established to enhance the incentives for enterprise managers to increase the efficiency of their firms. Replacement of the profit-remission system with tax and contracting systems was designed to reward managers by permitting firms to retain a significant portion of increases in production. Managerial authority within firms has been strengthened, and bonuses have been restored and allowed to grow to substantial proportions. Managers have also been given enhanced authority to hire, fire, and promote workers. Reductions in mandatory planning have been accompanied by permission for enterprises to buy and sell surplus goods on essentially a free-market basis. In many cases the prices thus obtained are far higher than for goods produced to meet plan quotas. The state plan has also been used to redirect some resources into the light industrial sector. The state, for example, gives priority in energy consumption to some light industrial enterprises that produce high-quality goods.

The reduction in the scope of mandatory planning is based on the assumption that market forces can more efficiently allocate many resources. This assumption, in turn, requires a rational pricing system that takes into account any and all extant technologies and scarcities. Because extensive subsidies were built into the economic system, however, price reform became an extremely sensitive issue. The fear of inflation also served as a constraint on price reform. Nevertheless, the fact that products produced in excess of amounts targeted in the plan can be sold, in most cases, at essentially free-market prices has created a two-tiered price system that is designed to gradually wean the economy from the administratively fixed prices of an earlier era.

Efforts to create a freer labour market are also part of the overall stress on achieving greater efficiency. As with price reform, tampering with a system that keeps many citizens living more comfortably and securely than would an economically more rational system risks serious repercussions in relations with the public. Changes have proceeded slowly in this sensitive area.

Special economic zones

A decision was made in 1978 to permit direct foreign investment in several small "special economic zones" along the coast. The country lacked the legal infrastructure and knowledge of international practices to make this prospect attractive for many foreign businesses, however. In later years steps were taken to expand the number of areas that could accept foreign investment with a minimum of red tape, and related efforts were made to develop the legal and other infrastructures necessary to make this work well.

This additional effort resulted in making 14 coastal cities and three coastal regions "open" areas for foreign investment. All of these places provide favoured tax treatment and other advantages for the foreign investor. Laws on contracts, patents, and other matters of concern to foreign businesses were also passed in an effort to attract international capital to aid China's development. The largely bureaucratic nature of China's economy, however, poses inherent problems for foreign firms that want to operate in the Chinese environment, and thus the policies to attract foreign capital have had to evolve continually in the direction of presenting more incentives for the foreigner to invest in China.

Search for efficiency

The common threads of these reforms are the search for efficiency and an assumption that management of the economy by large governmental bureaucracies is unlikely to produce this result. The changes in China's economic thinking and strategy since 1978 have been so great—with the potential repercussions for important vested interests so strong—that actual practice inevitably has lagged considerably behind declaratory policy. Indeed, by the end of 1989 China's economic policy had again begun to place greater emphasis on centralized planning and on large state-run enterprises, signifying a marked slowdown of the reforms. (E.I.U./K.G.L.)

## TRANSPORTATION

Great emphasis has been placed on the development of transport, because it is closely related to the development of the national economy, the consolidation of the national defense system, and the strengthening of national unification. Nevertheless, China's domestic transport system continues to constitute a major constraint on economic growth and the efficient movement of goods and people. Railroads, mostly employing steam locomotives, provide the major means for freight haulage, but their capacity cannot meet demand, especially for the shipment of coal.

Transport-development policy

Since 1949 China's transport and communications policies, influenced by political, military, and economic considerations, have experienced changes of emphasis in different periods. Thus, from 1949 to 1952 the primary concern was to repair existing lines of communication, to give priority to military transport needs, and to strengthen political control. During the First Five-Year Plan new lines were built, while at the same time old lines were improved. During the Great Leap Forward much of the improvement of regional transportation became the responsibility of the general population, and many small railways were constructed. After 1963 emphasis was placed on developing transportation in rural, mountainous, and especially forested areas, in order to help promote agricultural production; simultaneously the development of international communications was energetically pursued, and the scope of ocean transport was broadened considerably.

Initially, as China's railways and highways were mostly concentrated in the coastal regions, communications to the interior were inconvenient. This situation has been partly rectified, as railways and highways have been built in the remote border areas of the northwest and southwest. All of China except for certain counties in Tibet is accessible by rail, road, water, or air.

**Railways.** Railway construction began in China in 1876. Because railways can conveniently carry a greater volume of goods over long distances than can road transport, they are of especial importance in China's transportation system. All trunk railways in China are under the administration of the Ministry of Railways. The central government operates a major rail network in the Northeast built on a base constructed by the Russians and Japanese during the decades before 1949 and an additional large system inside (that is, to the south or east of) the Great Wall. The framework for the railways inside the wall consists of several north–south and east–west lines.

The small railways

Apart from those operated by the central government, there is also a network of small, state-owned local railways that link mines, factories, farms, and forested areas. The construction of these smaller railways is encouraged by the central government, and technical assistance is provided by the state railway system when it is thought that the smaller railways can stimulate regional economic development.

Coal has long been the principal railway cargo. The rather uneven distribution of coalfields in China makes it necessary to transport coal over long distances, especially between the North and South. The increase in the production of oil and natural gas has made necessary the construction of both pipelines and additional railways.

Since the late 1950s, there has been a change in railway-construction policy. Prior to that time, most attention was

paid to the needs of the eastern half of China, where most of the coal network is found; but since then, more emphasis has been given to extending the rail system into the western provinces and improving the original railway system, including such measures as building bridges, laying double tracks, and using continuous welded rail. In addition, certain important rail links have been electrified.

Since 1960 hundreds of thousands of workers have been mobilized to construct major lines in the northwest and southwest. In the 1970s new lines were extended into previously unopened parts of the country. In the 1980s new regions in the northwest were linked to the national market and opened up for development. The best example was the line built from Lan-chou in Kansu Province westward into the oil fields of the Tsaidam Basin. These projects, which were coordinated on a national level, contrast to the pattern prevailing before World War II, when foreign-financed railroads were built in different places without any attempt at coordination or at standardization of the transport and communications system.

**Road networks.** The first modern highway in China was built in 1913 in Hunan Province. The highways of China may be divided into three categories: (1) state, provincial, or regional highways of political, economic, or military importance; (2) local highways of secondary importance, operated by counties or communes; (3) special-purpose highways, mostly managed by factories, mines, state farms, forestry units, or the military forces.

The three highways into Tibet

The most striking achievement in highway construction has been the road system built on the cold and high Tsinghai–Tibetan plateau. Workers, after overcoming various physical obstacles, within a few years built three of the highest and longest highways in the world, thus markedly changing the transport pattern in the western border regions of China and strengthening the national defense system. Of the three highways, one runs across Szechwan into Tibet; another runs from Tsinghai to Tibet; and the third runs from Sinkiang to Tibet.

A rural road-building program has aimed at the opening up of commercial routes to the villages to facilitate the transport of locally produced goods. The wide dispersion and seasonal and variable nature of agricultural production, as well as the large numbers of relatively small shipments involved, explain the preferability of trucks for shipping. Similarly, trucks best bring consumer goods, fertilizers, and farm machinery and equipment to rural areas.

When large-scale highway construction was in progress, China also began to develop its petroleum and motor vehicle industries. The first motor vehicle manufacturing plant began to operate during the First Five-Year Plan. By 1970, apart from Tibet and the Hui Autonomous Region of Ningsia, all the provinces and autonomous regions could make their own motor vehicles.

The basis of the local motor vehicle industry is generally simple, usually an extension of motor vehicle repair shops in which vehicles of various types are produced to serve the needs of the locality. Vehicles produced by the large state automotive factories are distributed by the central government to state enterprises and military units. Special vehicles may also be built by the state at these plants. During the 1980s many motor vehicles, especially for passenger transportation, were imported.

**Waterways.** The high cost of construction prevents railways from being built extensively, and rail transport conditions are often congested. Freight volume carried by highways is limited, and highways are not suitable for moving bulky goods. China's water transport potential is great, but it is still far from being fully developed. Nonetheless, there are more than 68,000 miles of inland waterways open to navigation in China, with many more used for the transport of timber and bamboo. The distribution of waterways is chiefly within central and South China, except for a few navigable streams in the Northeast.

The principal waterways

One of the first goals of the Communist government, upon taking power in 1949, was to establish a national network of waterways. A program of building and refurbishing harbour and port facilities and of dredging river channels was initiated. By 1961 some 15 principal waterways had been opened to navigation, of which the Yangtze, Pearl River, Huai River, Huang Ho (Yellow River), Han River, and Grand Canal had received the most attention. Water transport development continues to receive considerable emphasis. Dredging and improvement of inland waterways have proved an important aid to economic reconstruction, while capital and maintenance costs for water transport are much lower than for railway transport.

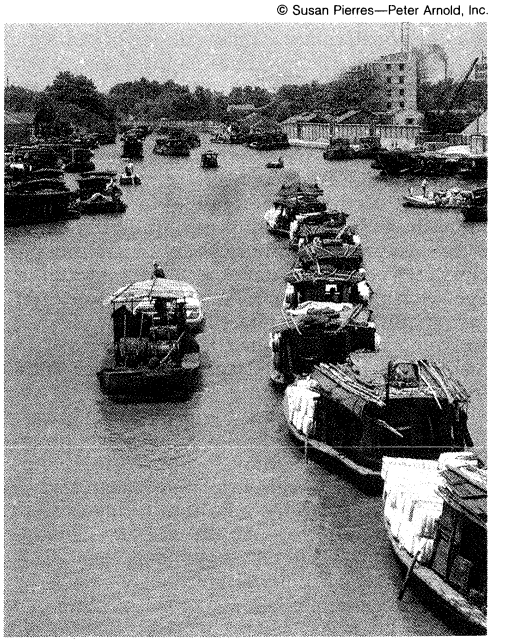
The Yangtze, the most important artery in China's waterway network, is also one of the most economically important rivers in the world. Together with its tributaries, it accounts for almost half of the nation's waterways, while the volume of the freight it carries represents about one-third of the total volume carried by river transport. Work undertaken in the mid-1950s to improve the middle course of the Yangtze allowed it to become navigable throughout the year from its mouth to I-pin in Szechwan. When the Yangtze is high in summer, it is navigable from its mouth to as far as Chungking for ships of up to 5,000 tons. By the early 1970s many cable-hauling stations had been established at rapids on the upper course of the Yangtze and of its major tributaries, such as the Wu River. Boats sailing against the current are hauled over the rapids with strong steel cables attached to fixed winches, thus augmenting their loading capacity, increasing speed, and saving time. Such improvements have permitted regular passenger and cargo services to be operated on the Yangtze.

The Hsi River is second in importance only to the Yangtze, being the major water transport artery in South China. Ships of 1,000 tons can sail up the Hsi River to Wu-chou, while shallow-water steamships and wooden boats can sail up the middle and upper courses of the Hsi, Pei, and Tung rivers and their tributaries. The Yangtze and the Hsi are not icebound in winter. Although the Sungari River, flowing across the Manchurian Plain, is navigable for half of its course, it is icebound from November through March; traffic is, however, very busy from April to October. The Amur, Sungari, and Ussuri rivers with their tributaries form a network of waterways totaling about 12,500 miles in length. In the past the Huang Ho was little navigated, especially on its middle and lower courses, but mechanized junks now navigate along the middle course in Honan.

Grand Canal

The Grand Canal, the only major Chinese waterway running from north to south, passes through the basins of the Hai, Huang, Huai, Yangtze, and Ch'ien-t'ang rivers in its 1,100-mile length from Peking to Hang-chou. One of the greatest engineering projects in China, equal in fame to the Great Wall, it is the world's longest artificial waterway; some of its sections follow the natural course of a river,

© Susan Pierres—Peter Arnold, Inc.



Cargo barges on the Grand Canal at Suchow, Kiangsu Province.

while other sections have been dug by hand. Work on the canal began as early as the 4th century BC and was completed by the end of the 13th century AD. It forms a north-to-south communications and transport link between the most densely populated areas in China. From the latter part of the 19th century, however, because of political corruption, mismanagement, and flooding from the Huang Ho, it gradually became silted up, and the higher section in Shantung became blocked. Since 1958 efforts have been made to reopen the Grand Canal to navigation, this time also by larger modern craft. The canal is important in the north–south transport of bulky goods, thus facilitating the nationwide distribution of coal and foodstuffs.

**Port facilities and shipping.** China's 8,700-mile-long coastline is indented by some 100 large and small bays and has some 20 deepwater harbours, most of which are ice-free throughout the year. Coastal shipping is divided into two principal navigation zones, the northern and southern marine districts. The northern district extends north from Amoy to the North Korean border, with Shanghai as its administrative centre. The southern district extends south from Amoy to the Vietnamese border, with Canton as the administrative centre. Most of the oceangoing routes begin from the ports of Lü-ta, Ch'in-huang-tao, T'ang-ku, Tsingtao, Shanghai, Huang-p'u, or Chan-chiang. Shanghai, the leading port of China since the early 19th century, maintains its position in all forms of water transport.

Northern and southern navigation zones

In 1961 China established an Ocean Shipping Company and subsequently signed ocean-shipping agreements with many countries; this laid the foundation for the development of ocean transport. Both before and after that year the Chinese government invested heavily in water transport construction. In addition to new port construction, older ports have been rebuilt and extended. A major effort has also been made to increase mechanization and containerization at major international ports.

**Aviation.** Aviation development is particularly suited to China, with its extensive territory and varied terrain. Chinese civil aviation has two major categories: air transport, which mainly handles passengers, cargoes, and mail, traveling on both scheduled and nonscheduled routes; and special-purpose aviation, which mainly serves industrial and agricultural production, national defense, and scientific and technological research.

The aims of civil aviation in China have been primarily to extend air routes; to strengthen the link between Peking and other important cities, as well as remote border and interior areas; to develop special-purpose flights serving the needs of agriculture, forestry, and geological prospecting; and to increase the number of large transport airplanes. With Peking, Shanghai, Canton, and Wu-lu-mu-ch'i as regional centres, an aviation-coordinating agency was formed to link major and local air routes. Air transport links between China and other countries were also arranged.

In the 1950s international aviation depended mainly on Soviet support; originally all principal international air routes passed through Moscow, where transit was made by Soviet planes. With the deterioration of the Sino-Soviet relationship in the late 1950s, China began to open direct air routes to other places as well. Thus, in addition to the original routes between China and the Soviet Union, North Korea, Mongolia, Vietnam, and Burma (Myanmar), air transport routes were opened to several of China's neighbouring countries, the United States, western Asia, Europe, and Africa. During the 1980s the number of air routes increased significantly.

Airport construction has increased greatly since Peking's first modern civilian airport was built in 1958. That airport was replaced in 1980 by the Capital Airport in Peking, and expanded international airports were completed at Shanghai and Canton. Flights to and from China are also available through several other cities. Airplanes, including various types of military aircraft, have long been made by China. Civil airliners for long-distance flights, however, are still mostly purchased abroad.

Chinese civil air efforts were carried out solely by the Civil Aviation Administration of China from 1949 until the mid-1980s. In an effort to improve efficiency and service, regional airlines were then introduced in competition with the airlines operated by the national administration. The Chinese Air Force controls a large number of airfields, and retired Air Force personnel have been the major source of civilian pilots.

**Posts and telecommunications.** Posts and telecommunications were established rapidly in the 1950s and '60s. By 1952 the principal posts and telecommunications network centred on Peking, and links to all large cities had been established. Great progress was made in improving the postal service under the First Five-Year Plan. Postal service was also developed in the rural areas. Besides extending rural postal routes, the problem of delivering mail to places below the county level was solved by enlisting the aid of the population. From 1954 onward a system of mail delivery by rural postal workers was tried in agricultural cooperatives, and in 1956 this system was extended throughout the country. By 1959 the national postal network was complete.

Postal service

From 1956 telecommunications routes were extended more rapidly. To increase the efficiency of the communication system, the same lines are used for both telegraphic and telephone service, while Teletype and television services also have been added. By 1963 telephone wire had been laid from Peking to the capitals of all provinces, autonomous regions, and large cities, while capitals of all provinces and autonomous regions were connected to the administrative seats of the counties and smaller municipalities and to larger market towns.

Immediately following 1949, telecommunications—by telegraph or telephone—mainly used wire; by the 1970s, however, radio telecommunications were increasingly used. Microwave and satellite transmissions have now become common. In 1956 the first automatic speed Teletype was installed on the Peking–Lhasa line. By 1964 such machines had been installed in most of China's major cities. Radio-television service has also been installed in major cities, and radio teleprinters are widely used. Overall, China's telecommunications services improved enormously during the 1980s. (C.-S.Ch./K.G.L.)

## Administration and social conditions

### GOVERNMENT AND PARTY

Despite its size, the People's Republic of China is organized along unitary rather than federal principles. Both the government and the Chinese Communist Party (CCP; Pinyin, Zhongguo Gongchan Dang; Wade–Giles romanization, Chung-kuo Kung-ch'an Tang), moreover, operate "from the top down," arrogating to the "Centre" all powers that are not explicitly delegated to lower levels. To run the country, the government and the CCP have established roughly parallel national bureaucracies extending from Peking down to local levels. These bureaucracies are assisted by various "mass organizations"—trade unions, a youth league, women's associations, writers' and other professional associations, and so forth—that encompass key sectors of the population. These organizations, with their extremely high memberships, have generally served as transmission lines for communicating and uniformly implementing policies affecting their members. No voluntary associations are permitted to function that are wholly independent of CCP and government leadership.

The CCP and government bureaucracies themselves are organized along territorial and functional lines. The territorial organization is based on a number of administrative divisions, with both a CCP committee and a "people's government" in charge of each. These territorial divisions include the national level in Peking (the Centre); 30 provincial-level units (three directly administered cities, five autonomous regions, and 22 provinces, excluding Taiwan); more than 150 prefectural bodies; more than 2,000 counties; and numerous cities, towns, and townships. Some larger cities are themselves divided into urban wards and counties. This territorial basis of organization is intended to coordinate and lend coherence to the myriad policies from the Centre that may affect any given locale.

Territorial administration

The functionally based political organization is led by, on the government side, ministries and commissions under

the State Council and, on the CCP side, Central Committee departments. These central-level functional bodies sit atop hierarchies of subordinate units that have responsibility for the sector or issue area under concern. Subordinate functional units typically are attached to each of the territorial bodies.

This complex structure is designed to coordinate national policy (such as that toward the metallurgical industry), assure some coordination of policy on a territorial basis, and enable the CCP to keep control over the government at all levels of the national hierarchy. One unintended result of this organizational approach is that China employs more than 10,000,000 officials—more officials than some countries have citizens.

There are tensions among these different goals, and thus a great deal of shifting has occurred since 1949. During the early and mid-1950s the government's functional ministries and commissions at the Centre were especially powerful. The Great Leap Forward, starting in 1958, shifted authority toward the provincial- and lower-level territorial CCP bodies. During the Cultural Revolution, starting in 1966, much of the political system became so disrupted that the People's Liberation Army (PLA) was called in and assumed control. When the PLA fell under a political cloud, the situation became remarkably fluid and confused for much of the 1970s.

Since then, the general thrust has been toward less detailed CCP supervision of the government and greater decentralization of government authority where possible. But the division of authority between CCP and government and between territorial and functional bodies has remained in a state of flux, as demonstrated by the trend again toward centralization at the end of the 1980s. The Chinese Communist political system still has not become institutionalized enough for the distribution of power among important bodies to be fixed and predictable.

Constitution of 1982

The fourth constitution of the People's Republic of China was adopted in 1982. It vests all national legislative power in the hands of the National People's Congress and its Standing Committee. The State Council and its Standing Committee, by contrast, are made responsible for executing rather than enacting the laws. This basic division of power is also specified for each of the territorial divisions—province, county, and so forth—with the proviso in each instance that the latitude available to authorities is limited to that specified by law.

All citizens over 18 years of age who have not been deprived of their political rights are permitted to vote, and direct popular suffrage is used to choose People's Congress members up to the county level. Above the counties, delegates at each level elect those who will serve at the People's Congress of the next higher level. Were this constitution an accurate reflection of the real workings of the system, the People's Congresses and their various committees would be critical organs in the Chinese political system. In reality, though, they are not.

Actual decision-making authority in China resides in the state's executive organs and in the CCP. At the national level the top government executive organ is the State Council, which is led by the premier. The constitution permits the appointment of vice-premiers, a secretary-general, and an unspecified number of councillors of state and heads of ministries and commissions. The premier, vice-premiers, state councillors, and secretary-general meet regularly as a kind of standing committee, in which the premier has the final decision-making power. This Standing Committee of the State Council exercises major day-to-day decision-making authority, and its decisions de facto have the force of law.

While it is not so stipulated in the constitution, each vice-premier and councillor assumes responsibility for the work of one or more given sectors or issues, such as education, energy policy, or foreign affairs. The leader concerned then remains in contact with the ministries and the commissions of the State Council that implement policy in that area. This division of responsibility permits a relatively small body such as the Standing Committee of the State Council (consisting of fewer than 20 people) to monitor and guide the work of a vast array of major bureaucratic entities. When necessary, of course, the Standing Committee may call directly upon additional expertise in its deliberations. The National People's Congress meets roughly annually and does only a little more than to ratify the decisions already made by the State Council.

Parallel to the State Council system is the central leadership of the CCP. The distribution of power among the various organs at the top of the CCP (the Standing Committee of the Politburo, the Politburo itself, and the Secretariat) has varied a great deal, and from 1966 until the late 1970s the Secretariat did not function at all. There is, in any case, a partial overlap of membership among these organs and between these top CCP bodies and the Standing Committee of the State Council. In addition, formally retired elder members of the party have often exercised decisive influence on CCP decision making.

According to the CCP constitution of 1982, the National Party Congress is the highest decision-making body. Since the Party Congress typically convenes only once in five years, the Central Committee is empowered to act when the Congress is not in session. Further, the Politburo can act in the name of the Central Committee when the latter is not in session, and the Standing Committee of the Politburo guides the work of the Politburo itself. The Secretariat is charged with the daily work of the Central Committee and the Politburo. The general secretary presides over the Secretariat and also is responsible for convening the meetings of the Politburo and its Standing Committee. The Secretariat works when necessary through several departments (the department for organization, for example, or the department for propaganda) under the Central Committee.

CCP chairmanship

Until 1982 the Chinese Communist Party had a chairmanship that was unique among ruling Communist parties. Mao Zedong (Mao Tse-tung) held this office until his death in 1976, and Hua Guofeng (Hua Kuo-feng) was chairman until his removal from office in June 1981. Hu Yaobang (Hu Yao-pang) then served as party chairman until the post was abolished in September 1982. The decision to redefine the position was part of the effort to reduce the chances of any one leader's again rising to a position above the party, as Mao had done. China's government still has a chairmanship, but the office has very limited power and is largely ceremonial.

The division of power among the leading CCP organs and between them and the State Council is constantly shifting. The Standing Committee of the Politburo and the Politburo as a whole have the authority to decide on any issue they wish to take up. The Secretariat has also at times played an extremely powerful and active role, meeting more frequently than either the Politburo or its Standing Committee and making many important decisions on its own authority. Similarly, the State Council has made many important decisions, but its power is always exercised at the pleasure of the CCP leadership.

Since the late 1970s China has taken a number of initiatives to move toward a more institutionalized system in which the office basically determines the power of its incumbent rather than vice versa, as has often been the case. Thus, for example, the CCP and state constitutions adopted in 1982 (and subsequently amended somewhat) for the first time stipulated a number of positions that confer membership status on the Standing Committee of the Politburo. These positions are the head of the Party Military Affairs Commission, the Secretary General of the CCP, the head of the Central Advisory Committee, and the head of the Central Discipline Commission. In addition, for the first time under the stipulations of the constitution, limits of two consecutive terms were placed on the government offices of premier, vice-premier, and state councillor. There were no similar constitutional restrictions on the tenure of incumbents to top CCP positions.

In theory, the Chinese Communist Party sets major policy directions and broadly supervises the implementation of policy to ensure that its will is not thwarted by the state and military bureaucracies. The CCP also assumes major responsibility for instilling proper values in the populace. The government, according to the theory, is responsible for carrying out CCP policy, making the necessary deci-

sions as matters arise. Of course, this clear division of labour quickly becomes blurred for a number of reasons. For example, only since the late 1970s has a concerted effort been made to appoint different people to the key executive positions in the CCP and the government. Prior to that time, the same individual would head both the CCP committee and the government body in charge of any given area. At the highest levels the premier of the government and the chairman of the party continue to sit on the CCP Politburo.

More fundamentally, it is often impossible to clearly separate policy formation and implementation in a huge, complex set of organizations charged with a multiplicity of tasks. The tendency has been for CCP cadres to become increasingly involved in day-to-day operations of the government, until some major initiative was taken by the top national leadership to reverse the trend. While the distinction between the CCP and the government is of considerable significance, therefore, the ruling structure in China can also be viewed from the functional point of view mentioned above. The careers of individual officials may shift among posts in both the CCP and the government, but for most officials all posts are held within one area of concern, such as economics, organization or personnel, security, propaganda, or culture.

### ADMINISTRATION

A hierarchy of organization and personnel has been embedded in virtually all CCP and government bodies. Even on the government side, all officials in these personnel departments are members of the CCP, and they follow rules and regulations that are not subject to control by the particular bodies of which they are formally a part. This system has been used to assure higher-level CCP control over the appointments to all key positions in the CCP, government, and other major organizations (enterprises, universities, and so forth).

For much of the period between 1958 and 1978, these personnel departments applied primarily political criteria in making appointments. They systematically discriminated against intellectuals, specialists, and those with any ties or prior experience abroad. From 1978 to 1989, however, official policy was largely the reverse, with ties abroad being valued because of China's stress on "opening the door" to the international community. A good education became an important asset in promoting careers, while a history of political activism counted for less or could even hinder upward mobility. A partial reversion to pre-1978 criteria was decreed in 1989.

Two important initiatives have been taken to reduce the scope of the personnel bureaucracies. First, during 1984 the leaders of various CCP and government bodies acquired far greater power to appoint their own staffs and to promote from among their staffs on their own initiative. The leaders themselves still must be appointed via the personnel system, but most others are no longer fully subject to those dictates. Second, a free labour market has been encouraged for intellectuals and individuals with specialized skills, a policy that could further reduce the power of the personnel bodies.

### ARMED FORCES

People's Liberation Army

The People's Liberation Army (PLA) is the unified organization of all Chinese land, sea, and air forces. The history of the PLA is officially traced to the Nan-ch'ang Uprising of Aug. 1, 1927, which is celebrated annually as PLA Day. The PLA is one of the world's largest military forces, with approximately 3,000,000 members. Military service is compulsory for all men who attain the age of 18; women may register for duty in the medical, veterinary, and other technical services. Demobilized servicemen are carried in a ready reserve, which is reinforced by a standby reserve of veterans and by the militia.

The PLA is formally under the command of the central Military Commission of the CCP; there is also an identical commission in the government, but it has no clear independent functions. The CCP commission is far more powerful than the Ministry of National Defense, which operates under the State Council, and it assures continuing CCP control over the armed forces. The political leadership has made a concerted effort to create a professional military force restricted to national defense and to the provision of assistance in domestic economic construction and emergency relief. This conception of the role of the PLA requires the promotion of specialized officers who can understand modern weaponry and handle combined arms operations. Troops around the country are stationed in seven military regions and more than 20 military districts. Despite the drive to modernize the PLA, limited military budgets and other constraints have caused the sophistication of conventional military armaments and of logistics and command-and-control systems to lag behind that of other major military powers.

Public Security forces

The role of the Public Security forces of China began to change in the late 1970s. The definition and designation of what poses a threat to security, for example, were narrowed, and there was a decline in the scope of activities of the security forces. The practice of political suppression, the victims of which once numbered in the tens of millions, was reduced, and in the late 1970s a large (but unknown) number of people were released from labour or other camps run by the Public Security forces. Also, during the 1980s the "open-door" policy toward the outside world led to the adoption of a more relaxed official attitude toward "protecting" Chinese from the "corrupting" influences of the West. Efforts by the Public Security forces to control and restrict the activities of foreigners in China diminished accordingly. By 1990, however, the trend was again toward a stricter policy and tighter controls.

Specific organizational and policy initiatives also have affected the role of the Public Security forces. The trend toward creating a body of codified law and toward establishing a legal system that operates according to that law has in itself reduced the arbitrary power that had been exercised by the Public Security system. (By the 1970s that system had effectively acquired the power to arrest, convict, sentence, and detain any individual without interference from any other "outside" body.) The Public Security Ministry also has relinquished administrative control over counterespionage and economic crimes, which was transferred to a newly organized Ministry of State Security.

### JUSTICE

Legal apparatus

The legal apparatus that existed before the changes made during the Cultural Revolution has been resurrected. The State Council again has a Ministry of Justice, and procuratorial organs and a court system have been reestablished. The legal framework for this system has been provided through the adoption of various laws and legal codes. For the first time, the law provides that there should be no discrimination among defendants based on their class origin. China has also reestablished a system of lawyers.

The actual functioning of this legal apparatus, however, continues to be adversely affected by a shortage of qualified personnel and by deeply ingrained perspectives that do not accord the law priority over the desires of political leaders. Thus, for example, when the top CCP leadership ordered a severe crackdown on criminal activity in 1983, thousands were arrested and executed without fully meeting the requirements of the newly passed law on criminal procedures. That law was subsequently amended to conform more closely with the actual practices adopted during the crackdown.

### HEALTH AND WELFARE

The Chinese government faces a mammoth task in trying to provide medical and welfare services adequate to meet the basic needs of the immense number of Chinese people whose per capita income and caloric intake are far below the world's average. The medical system, moreover, labours under the tension of whether to stress quality of care or to spread scarce medical resources as widely as possible. In addition, there has been repeated debate over the relative balance that should be struck between the use of Western and traditional Chinese medicine. While the Cultural Revolution pushed the balance toward widespread minimum care with great attention paid to traditional medicine, policy after the late 1970s moved in

the other direction on both issues. The Ministry of Public Health of the State Council oversees the health-services system, which includes a substantial rural collective sector but almost no private sector. All of the major medical facilities are run by the government.

Life expectancy

The health of the Chinese populace has improved considerably since 1949. Average life expectancy has gone up about three decades and now ranks nearly at the level of that in advanced industrial societies. Many communicable diseases, such as plague, smallpox, cholera, and typhus, have either been wiped out or brought under control. Chinese statistics also indicate that the incidences of malaria and schistosomiasis have declined by 90 percent and 80 percent, respectively, since 1949.

As evaluated on a per capita basis, China's health facilities remain unevenly distributed. Only slightly more than 50 percent of the country's medical and health personnel work in rural areas, where approximately 80 percent of the population resides. The doctors of Western medicine, who constitute about one-sixth of the total medical personnel, are even more concentrated in the cities. Similarly, close to two-thirds of the country's hospital beds are located in the cities.

China has a health insurance system that provides virtually free coverage for people employed in urban state enterprises and relatively inexpensive coverage for their families. The situation for workers in the rural areas or in urban employment outside of the state sector is far more varied. There are some cooperative health-care programs, but their voluntary nature produced a decline in membership from the late 1970s.

The severest limitation on the availability of health services, however, appears to be an absolute lack of resources, rather than discrimination in access based on the ability of individuals to pay. An extensive system of paramedical care has been fostered as the major medical resource available to most of the rural population, but the care thus provided has been found to be of quite uneven quality. The paramedical system feeds patients into the more sophisticated commune-level and county-level hospitals when they are available.

Changes in the leading causes of death reflect the longer life span and improved living conditions that have developed in China since 1949. By the end of the 1980s the major causes had become pulmonary and cerebrovascular diseases, cardiac disease, and malignant tumours. Severe environmental pollution has become a major health hazard in several parts of the country.

Because a large percentage of what in the West would be considered public welfare obligations is in China the responsibility of factories, offices, and rural collectives and families, the real level of "welfare" spending and the strengths and weaknesses of the welfare system are difficult to gauge. Little statistical information is available that could help clarify the situation. The state provides pensions for retirees from state enterprises and official service, but this includes only a small percentage of the total work force. The state's welfare resources are heavily concentrated in the urban areas, where they include subsidies for housing, medical care, education, and some foods. In the cities the level of subsidized services, though, depends largely on the nature and conditions of the unit in which a person works. Unemployed individuals are typically taken care of by their working relatives, and all Chinese citizens have a legal obligation to care for their elderly parents. There is a small number of old-age homes for the elderly who have no children or other relatives to support them.

In the rural areas much of the responsibility for welfare is left to the local collective units, the resources of which have declined since the late 1970s. China's rural collectives are supposed to provide for their poor, but the actual level of services varies greatly depending on both the financial standing of the locality and the inclinations of the villagers. The Chinese government does allocate emergency relief to areas that have suffered from natural disasters (including crop failures). Generally, the Ministry of Civil Affairs of the State Council assumes primary responsibility for administering the government's portion of China's welfare system.

## EDUCATION

The educational system in China is a major vehicle for both inculcating values in and teaching needed skills to its people. Traditional Chinese culture attached great importance to education as a means of enhancing a person's worth and career. In the early 1950s the Chinese Communists worked hard to increase the country's rate of literacy, an effort that won them considerable support from the population. By the end of that decade, however, the government could no longer provide jobs adequate to meet the expectations of those who had acquired some formal schooling. Other pressing priorities squeezed educational budgets, and, of course, the anti-intellectualism inherent in the more radical mass campaign periods affected the status and quality of the educational effort. These conflicting pressures made educational policy a sensitive barometer of larger political trends and priorities. The shift to rapid and pragmatic economic development as the overriding national goal in the late 1970s quickly affected China's educational system.

China's school system

The Chinese educational structure provides for six years of primary school, three years each of lower middle school and upper middle school, and four years in the standard university curriculum. All urban schools are financed by the state, while rural schools depend far more heavily on their own financial resources. Official policy stresses scholastic achievement, with particular emphasis on the natural sciences. A significant effort is made to enhance vocational training opportunities for students who do not attend a university. The quality of education that is available in the cities is generally far higher than that in the countryside, where relatively few students acquire even a secondary education.

The overall trend in Chinese education is toward fewer students and higher scholastic standards, resulting in a steeply hierarchical educational system. Only about one-third of the nation's primary school students gain access to some secondary education, while less than one-tenth of 1 percent ever attend a regular university. Only the best students are allowed to go beyond a primary school, and numerous secondary schools are closed because of a lack of students. For the overwhelming majority of students, admission to a university since 1977 has been based on competitive nationwide examinations, and attendance at a university is usually paid for by the government. In return, a university student has had to accept the job provided by the state upon graduation.

The system developed in the 1950s of setting up "key" urban schools that were given the best teachers, equipment, and students was reestablished in the late 1970s. The inherently elitist values of such a system put enormous pressure on secondary-school administrators to improve the rate at which their graduates passed tests for admission into universities.

Major universities

Six universities, all administered directly by the State Education Commission in Peking, are the flagships of the Chinese higher educational system. They are Peking University, the leading nontechnical university; Ch'ing-hua University, an institution that is oriented primarily toward engineering; People's University of China, the only major university founded after 1949; Nan-k'ai University in Tientsin, which is especially strong in the social sciences; Fu-tan University, a comprehensive institution in Shanghai; and Sun Yat-sen (Chung-shan) University in Canton, the principal university of South China. In addition, every province has a key provincial university, and there are hundreds of other technical and comprehensive higher educational institutions in locations around the country.

The damage done to China's human capital by the ravages of the Great Leap Forward and, especially, by the Cultural Revolution was so great that it could take years to fully make up the loss. After the 1970s, however, China's educational system increasingly trained individuals in technical skills so that they could fulfill the needs of the advanced, modern sector of the economy. The social sciences and humanities also receive more attention than in earlier years, but the base in those disciplines is relatively weak—many leaders still view them with suspicion—and the resources devoted to them are thin. (K.G.L.)

## Cultural life

Chinese culture is remarkable for its duration and diversity. Skeletal remains and stone implements date to the Paleolithic stage of cultural development, from the 29th to the 17th millennium BC. Decorated artifacts, primarily marked pottery vessels, have been found in dozens of Incipient Neolithic and Neolithic sites, dating from the 12th to the 2nd millennium BC.

Yang-shao and Lung-shan pottery

Chinese Neolithic pottery shapes and types are mostly classified into two families—the earlier Yang-shao ware from the central Chung-yüan region, characterized by geometric painted decorations, and the later Lung-shan ware, primarily from the Northeast but also found in the Chung-yüan area. Lung-shan ware is unpainted and is elevated from the ground on a circular foot or tripod legs.

The Bronze Age includes the first historically verified dynasty, the Shang (18th–12th century BC), and China's first written records. The Late Shang is well known from oracle bones recovered from the site of the last Shang capital near An-yang. The bones are turtle plastrons and ox scapulae with inscribed texts, used by the Shang kings in a highly regularized system of ritual divination and sacrifice aimed at securing the support of the ruler's deceased ancestors. Through their use, writing became linked to authority in a way that endured throughout premodern Chinese history. During the Shang and Chou (1111–255 BC) dynasties the art of bronze casting became highly developed. Finely cast and richly decorated pieces included cooking and serving vessels, bells, drums, weapons, and door fittings.

The written language

The written language is central to China's culture. Scholars have identified ideographic inscriptions on pottery dating to about 4000 BC, and written Chinese has developed continuously since the Late Shang period. Chinese culture is inextricably bound up with the writing system in three ways. First, writing is the medium for the preservation and dissemination of culture. Indeed, China's word for culture (*wen-hua*) means "to become literate." Second, command of the writing system distinguishes the Chinese and their culture, seen as the centre of the world, from all non-Chinese peoples, categorized by the Chinese as "barbarians." Third, culture and the writing system are inseparably linked to statecraft in that a command of writing and knowledge of the written tradition were necessary and requisite skills for holding office. Thus, from the Shang dynasty oracle bones to the products of the modern printing press, culture in the form of written works has been a key instrument in the development of political thought and a tool of governance.

The oldest art forms in China are music and dance. A 5,000-year-old pottery bowl from Tsinghai Province is painted with a ring of 15 dancers, adorned in headdresses and sashes and stepping in unison. Music played an important role in early Chinese ritual and statecraft. Bronze bells were instruments of investiture and reward. A bronze bell set from the ancient state of Tseng in Hupeh, interred *c.* 430 BC, contains 64 bells, each of which produces two distinct, tuned strike notes. More than 120 instruments were unearthed from the same tomb, including stringed zithers, mouth organs, flutes, drums, and stone chimes. Music and related rituals helped to provide a structure for activities in the courts of rulers at all levels in the feudal hierarchy.

The *Shih Ching*

The *Shih Ching* ("Classic of Poetry"), an anthology of poetry given definitive form in about 500 BC, is one of China's oldest classics and contains 305 folk songs and ritual psalms. Although the T'ang dynasty (AD 618–907) is called the Golden Age of Chinese poetry, having produced the poets Tu Fu and Li Po, there are poets of renown from every dynasty, and the writing of poetry was practiced by most well-educated Chinese for both personal and social reasons.

China's tradition of historical narrative is also unsurpassed in the world. Twenty-five dynastic histories preserve a unique record from the unverified Hsia dynasty (22nd–19th/18th century BC) to the Ch'ing (AD 1644–1911/12), and sprawling historical romances have been a mainstay in the reading of the educated since the spread of printing in the 11th and 12th centuries AD.

The May Fourth Movement (1917–21) attacked much of this great literary and cultural tradition, viewing it as a source of China's weakness. Students and faculty at Peking University abandoned the demanding literary language and created a new popular fiction, written in a more accessible colloquial language on themes from the lives of ordinary people. Literary culture continued to be a subject of intense debate. Mao Zedong, who composed poetry in both contemporary and traditional styles, dictated that art must serve politics in his talks at Yen-an in 1942. Throughout the following decades, writers received both admiration and ridicule. Indeed, the fate of most important writers was closely linked to the vicissitudes of national politics from the 1950s onward. Only in the mid-1980s did writers again begin to enjoy a period of relative freedom in which there was some official tolerance of "art for art's sake."

Painting and calligraphy, like poetry, were the domain of the elite, and most educated Chinese traditionally boasted of some competence in them. There are early anonymous and folk-oriented paintings on tomb and cave walls, and many works are known from the Han dynasty (206 BC–AD 220). Fine-art painters are known by name from as early as the 6th century AD from historical records and serially copied versions of their works. Chinese painting is predominantly of landscapes, done in black pine-soot ink on fine paper or silk, occasionally with the addition of faint colour washes. The most vigorous period for landscape painting spanned the years from the Sung (960–1279) to Ming (1368–1644) dynasties.

Calligraphy rivals painting as a fine art in China, and paintings are often captioned with artfully written poems. Calligraphy reveals the great fondness the Chinese have for their written characters, and it ranges in style from meticulously and laboriously scribed "seal" characters to flamboyant and unconstrained "grass" characters. Calligraphy, as painting, is prized for a number of abstract aesthetic qualities, described by such terms as "balance," "vitality," "energy," "bones," "wind," and "strength."

Painting has undergone numerous style changes in the 20th century. Before 1949 painters such as Ch'i Pai-shih (1863–1957) developed distinct new styles that internationalized traditional Chinese aesthetics. After 1949 pressure for socialist realism made painters shift their focus to such subjects as factory scenes, peasant villages, and convoys of tour buses. But with the liberalization of the arts that followed Mao's death in 1976, more traditional values reasserted themselves.

Chinese sculpture and carving

Sculpture and carving date back to the Chou dynasty (*c.* 1111–255 BC) or earlier. Tombs frequently contained burial dolls, said to have been made to replace live sacrificial victims, and many early jade carvings are related to burial practices and include body orifice stoppers and bangle bracelets. Of all the arts, sculpture received the greatest boost from the introduction of Buddhism to China during the Han dynasty and from the spread of Buddhism during the Six Dynasties (AD 220–589) and T'ang periods. Statues and carved reliefs of Buddhas and *bodhisattvas* were made by the thousands and, along with cave paintings at prominent sites like Tun-huang, Kansu Province, represent the pinnacle of Chinese religious art.

Chinese art and artifacts have found their way into various collections around the world. The most important collection of fine arts is in the National Palace Museum in Taipei, Taiwan, the bulk of the superb traditional palace collection having been ferried across the Taiwan Strait during the Nationalist retreat in 1948–49. Excellent collections of Chinese painting, calligraphy, and bronzes are found in a number of North American museums, most notably the William Rockhill Nelson Gallery and Atkins Museum of Fine Arts in Kansas City, Mo.; the Smithsonian Institution, Freer Gallery of Art in Washington, D.C.; the Fogg Art Museum at Harvard University; and the Museum of Fine Arts, Boston. Significant collections remain in major museums in Peking, Shanghai, Nanking, and Wu-han. Since the 1950s, new archaeological discoveries have filled China's provincial and local museums with fabulous treasures, and new facilities have been constructed for the study and display of these artifacts. Espe-

cially notable is the life-size terra-cotta army of the first Ch'in emperor, Shih huang-ti. The army, complete with soldiers, horses, and chariots, was discovered in an enormous underground vault near Sian, in Shensi Province. Many of its figures have been painstakingly removed and prepared for public display.

Theatre is the most important popular art in China. It originated in early religious dances, performed at festivals to exorcise demons, reenact important historical events, or prepare for harvest, hunting, or warfare. Urban storytelling and theatrical genres are well documented from the Sung dynasty but are known to have matured during the Yüan dynasty (1206–1368). Yüan dramas, or operas as they are more accurately called, consisted of virtuoso song and dance organized around plots on historical or contemporary themes. The operas were performed in special theatres, with elegant costumes and decorated stages. From Yüan drama later forms developed, including contemporary Cantonese and Peking operas, that feature song and dance, elaborate costumes and props, and displays of martial arts and acrobatics.

During the Cultural Revolution, an enormous number of cultural treasures of inestimable value were seriously damaged or destroyed and the practice of many arts and crafts was prohibited. Since the early 1980s, however, official repudiation of those policies has been complemented by vigorous efforts to renew China's remarkable cultural traditions. China's culture thus remains highly complex, encompassing ancient traditions and modern experiments, in what sometimes appears to be a rather tenuous mix.

(K.J.DeW.)

For statistical data on the land and people of China, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL.

# HISTORY

## Prehistory

### ARCHAEOLOGY IN CHINA

The practice of archaeology in China has been rooted in modern Chinese history. The intellectual and political reformers of the 1920s challenged the historicity of the legendary inventors of Chinese culture, such as Shen Nung, the divine farmer, and Huang Ti, the Yellow Emperor. At the same time, scientific study of the prehistoric period was being sponsored by Western archaeologists and paleoanthropologists. The establishment of the Academia Sinica (Chinese Academy of Sciences) in 1928 enabled Chinese scholars to study Chinese archaeology for themselves, but the eruption of the Sino-Japanese War in 1937 made excavation difficult. That some researchers found themselves, with their collections, in Taiwan after 1949 and that much archaeology practiced in the People's Republic of China was reported within a Marxist framework further demonstrate archaeology's links to politics. The waning of the Cultural Revolution in the 1970s meant the resumption of archaeological excavation and publication. A modernizing nation began to produce scholarship, increasingly informed by scientific analysis, in a quantity and quality commensurate with its size and its traditions of learning.

### EARLY MAN

The fossil record in China promises fundamental contributions to the understanding of human origins. There is considerable evidence of *Homo erectus* by the time of the Lower Paleolithic (the Paleolithic Period began *c.* 2,500,000 years ago and ended 10,000 years ago) at sites such as Lan-t'ien, Shensi; Ho-hsien, Anhwei; Yüan-mou, Yunnan; and, the most famous, that of so-called Peking man at Chou-k'ou-tien, Peking Municipality. The Lower Cave at the last site has yielded evidence of intermittent human use from *c.* 460,000 to 230,000 years ago. Many caves and other sites in Anhwei, Hupeh, Honan, Liaoning, Shantung, Shansi, and Shensi in North China and in Kweichow and Hupeh in the South suggest that *H. erectus* achieved wide distribution in China. Whether *H. erectus pekinensis* intentionally used fire and practiced ritual cannibalism are matters under debate.

Peking man

Significant *Homo sapiens* cranial and dental fragments have been found together with Middle Paleolithic artifacts. Such assemblages have been unearthed at Ting-ts'un, Shansi; Ch'ang-yang, Hupeh; Ta-li, Shensi; Hsu-chia-yao, Shansi; and Ma-pa, Kwangtung. Morphological characteristics such as the shovel-shaped incisor, broad nose, and mandibular torus link these remains to the modern Mongoloid race. Few archaeological sites have been identified in the south.

A number of widely distributed *H. erectus* sites dating from the upper Pleistocene manifest considerable regional and temporal diversity. Upper Paleolithic sites are numerous in North China. Thousands of stone artifacts, most of them small (called microliths), have been found, for example, at Hsiao-nan-hai, near An-yang, Honan; Shuo-hsien and Ch'in-shui, Shansi; and Yang-yüan, Hopeh; these findings suggest an extensive microlith culture in North China. Hematite, a common iron oxide ore used for colouring, was found scattered around skeletal remains in the Upper Cave at Chou-k'ou-tien (*c.* 10th millennium BC) and may represent the first sign of human ritual.

### NEOLITHIC PERIOD

The complex of developments in stone tool technology, food production and storage, and social organization that is often characterized as the "Neolithic Revolution" was in progress in China by at least the 6th millennium BC. Developments in the Chinese Neolithic were to establish some of the major cultural dimensions of the subsequent Bronze Age.

**Climate and environment.** Although the precise nature of the paleoenvironment is still in dispute, temperatures in Neolithic China were probably some 4° to 7° F (2° to 4° C) warmer than they are today. Rainfall, although more abundant, may have been declining in quantity. The Tsinling Mountains in northwest China separated the two phytogeographical zones of North and South China, while the absence of such a mountain barrier farther east encouraged a more uniform environment and the freer movement of Neolithic peoples about the North China

Adapted from A. Herrmann, *An Historical Atlas of China* (1966); Aldine Publishing Company




Principal sites of prehistoric and Shang China.

Plain. East China, particularly toward the south, may have been covered with thick vegetation, some deciduous forest, and scattered marsh. The Loess Plateau in the northwest is thought to have been drier and even semiarid, with some coniferous forest growing on the hills and with brush and open woodland in the valleys.

**Food production.** The primary Neolithic crops, domesticated by the 5th millennium BC, were drought-resistant millet (usually *Setaria italica*), grown on the eolian or alluvial loess soils of the northwest and the north, and glutenous rice (*Oryza sativa*), grown in the wetlands of the southeast. These staples were supplemented by a variety of fruits, nuts, legumes, vegetables, and aquatic plants. The main sources of animal protein were pigs, dogs, fish, and shellfish. By the Bronze Age millet, rice, soybeans, tea, mulberries, hemp, and lacquer had become characteristic Chinese crops. That most, if not all, of these plants were native to China indicates the degree to which Neolithic culture developed indigenously. The distinctive cereal, fruit, and vegetable complexes of the northern and southern zones in Neolithic and early historic times suggest, however, that at least two independent traditions of plant domestication may have been present.

Millet and rice crops

The stone tools used to clear and prepare the land reveal generally improving technology. There was increasing use of ground and polished edges and of perforation. Regional variations of shape included oval-shaped axes in central and northwest China, square- or trapezoid-shaped axes in the east, and axes with stepped shoulders in the southeast. By the Late Neolithic a decrease in the proportion of stone axes to adzes suggests the increasing dominance of permanent agriculture and a reduction in the opening up of new land. The burial in high-status graves of finely polished, perforated stone and jade tools such as axes and adzes with no sign of edge wear indicates the symbolic role such emblems of work had come to play by the 4th and 3rd millennia.

**Major cultures and sites.** There was not one Chinese Neolithic but a mosaic of regional cultures whose scope and significance are still being determined. Their location in the area defined today as China does not necessarily mean that all the Neolithic cultures were Chinese or even proto-Chinese. Their contributions to the Bronze Age civilization of the Shang, which may be taken as unmistakably Chinese in both cultural as well as geographical terms, need to be assessed in each case. In addition, the presence of a particular ceramic ware does not necessarily define a cultural horizon; and transitional phases, both chronological and geographical, cannot be discussed in detail in the following paragraphs.

*Incipient Neolithic.* Study of the historical reduction of the size of human teeth suggests that the first human beings to eat cooked food did so in South China. The southern sites of Hsien-jen-tung in Kiangsi and Tseng-p'i-yen in Kwangsi have yielded artifacts from the 10th to the 7th millennium BC that include low-fired, cord-marked sherds with some incised decoration and mostly chipped stone tools; these pots may have been used for cooking and storage. Pottery and stone tools from shell middens in South China also suggest Incipient Neolithic occupations. These early southern sites may have been related to the Neolithic Bac-Son culture in Vietnam; connections to the subsequent Neolithic cultures of northwest and North China have yet to be demonstrated.

*Sixth millennium BC.* Two major cultures can be identified in the northwest: Lao-kuan-t'ai, in eastern and southern Shensi and northwestern Honan, and Ta-ti-wan I—a development of Lao-kuan-t'ai culture—in eastern Kansu and western Shensi. In these cultures pots were low-fired, sand-tempered, and mainly red in colour, and bowls with three stubby feet or ring feet were common. The painted bands of this pottery may represent the start of the Painted Pottery culture.

In North China the people of P'ei-li-kang (north central Honan) made less use of cord marking and painted design on their pots than did those at Ta-ti-wan I; the variety of their stone tools, including sawtooth sickles, indicates the importance of agriculture. The Tz'u-shan potters (southern Hopeh) employed more cord-marked decoration and made a greater variety of forms, including basins, cups, serving stands, and pot supports. The discovery of two pottery models of silkworm chrysalides and 70 shuttle-like objects at a 6th-millennium-BC site at Nan-yang-chuang (southern Hopeh) suggests the early production of silk, the characteristic Chinese textile.

Tz'u-shan pottery

*Fifth millennium BC.* The lower stratum of the Pei-shou-ling culture is represented by finds along the Wei and Ching rivers; bowls, deep-bodied jugs, and three-footed vessels, mainly red in colour, were common. The lower stratum of the related Pan-p'o culture, also in the Wei River drainage area, was characterized by cord-marked red or red-brown ware, especially round and flat-bottomed bowls and pointed-bottomed amphorae. The Pan-p'o inhabitants lived in semisubterranean houses and were supported by a mixed economy of millet agriculture, hunting, and gathering. The importance of fishing is confirmed by designs of stylized fish painted on a few of the bowls and by numerous hooks and net sinkers.

Pei-shou-ling, Pan-p'o, and Hou-kang cultures

In the east by the start of the 5th millennium the Pei-hsin culture in central and southern Shantung and northern Kiangsu was characterized by fine clay or sand-tempered pots decorated with comb markings, incised and impressed designs, and narrow, appliquéd bands. Artifacts include many three-legged, deep-bodied tripods, goblet-like serving vessels, bowls, and pot supports. Hou-kang (lower stratum) remains have been found in southern Hopeh and central Honan. The vessels, some finished on a slow wheel, were mainly red coloured and had been fired at high heat. They include jars, tripods, and round-bottomed, flat-bottomed, and ring-footed bowls. No pointed amphorae have been found, and there were few painted designs. A characteristic red band under the rim of most gray-ware bowls was produced during the firing process.

Archaeologists have generally classified the lower strata of Pei-shou-ling, Pan-p'o, and Hou-kang cultures under the rubric of Painted Pottery (or, after a later site, Yang-shao) culture, but two cautions should be noted. First, a distinction may have existed between a more westerly, Wei Valley culture (early Pei-shou-ling and early Pan-p'o) that was rooted in the Lao-kuan-t'ai culture and a more easterly one (Pei-hsin, Hou-kang) that developed from the P'ei-li-kang and Tz'u-shan cultures. Second, since only 2 to 3 percent of the Pan-p'o pots were painted, the designation Painted Pottery culture seems premature.

In the region of the lower Yangtze River the Ho-mu-tu site in northern Chekiang has yielded caldrons, cups, bowls, and pot supports made of porous, charcoal-tempered black pottery. The site is remarkable for its wooden and bone farming tools, the bird designs carved on bone and ivory, the superior carpentry of its pile dwellings (a response to the damp environment), a wooden weaving shuttle, and the earliest lacquer ware and rice remains yet reported in the world (*c.* 5000 to 4750 BC). The Ch'ing-lien-kang culture, which succeeded that of Ho-mu-tu in Kiangsu, northern Chekiang, and southern Shantung, was characterized by ring-footed and flat-bottomed pots, *kuei* pouring vessels, tripods (common north of the Yangtze), and serving stands (common south of the Yangtze). Early fine-paste red ware gave way in the later period to fine-paste gray and black ware. Polished stone artifacts include axes and spades, some perforated, and jade ornaments. Another descendant of Ho-mu-tu culture was that of Ma-chia-pang, which had close ties with the Ch'ing-lien-kang culture in southern Kiangsu, northern Chekiang, and Shanghai. In southeastern China a cord-marked pottery horizon, represented by the site of Fu-kuo-tun on the island of Quemoy, existed by at least the early 5th millennium. The suggestion that some of these southeastern cultures belonged to an Austronesian complex remains to be fully explored.

Ho-mu-tu site

*Fourth and third millennia BC.* A true Painted Pottery culture developed in the northwest partly from the Wei Valley and Pan-p'o traditions of the 5th millennium. The Miao-ti-kou I horizon, dated from the first half of the 4th millennium, produced burnished bowls and basins of fine red pottery, some 15 percent of which were painted, generally in black, with dots, spirals, and sinuous lines. It was succeeded by a variety of Ma-chia-yao cultures (late

4th to early 3rd millennium) in eastern Kansu, eastern Tsinghai, and northern Szechwan. Thirty percent of Ma-chia-yao vessels were decorated on the upper two-thirds of the body with a variety of designs in black pigment; multiarmed radial spirals, painted with calligraphic ease, were the most prominent. Related designs involving sawtooth lines, gourd-shaped panels, spirals, and zoomorphic stick figures were painted on pots of the Pan-shan (mid-3rd millennium) and Ma-ch'ang (last half of 3rd millennium) cultures. Some two-thirds of the pots found in the Ma-ch'ang burial area at Liu-wan in Tsinghai, for example, were painted. In the North China Plain, Ta-ho culture sites contain a mixture of Miao-ti-kou and eastern, Ta-wen-k'ou vessel types (see below), indicating that a meeting of two major traditions was taking place in this area in the late 4th millennium.

In the northeast the Hung-shan culture (4th millennium and probably earlier) was centred in western Liaoning and eastern Inner Mongolia. It was characterized by small bowls (some with red tops), fine red-ware serving stands, painted pottery, and microliths. Numerous jade amulets in the form of birds, turtles, and coiled dragons reveal strong affiliations with the other jade-working cultures of the east coast, such as Liang-chu (see below).

In east China the Liu-lin and Hua-t'ing sites in northern Kiangsu (first half of 4th millennium) represent regional cultures that derived, in large part, from that of Ch'ing-lien-kang. Upper strata also show strong affinities with contemporary Ta-wen-k'ou sites in southern Shantung, northern Anhwei, and northern Kiangsu. Ta-wen-k'ou culture (mid-5th to at least mid-3rd millennium) is characterized by the emergence of wheel-made pots of various colours, some of them remarkably thin and delicate; vessels with ring feet and tall legs (such as tripods, serving stands, and goblets); carved, perforated, and polished tools; and ornaments in stone, jade, and bone. The people practiced skull deformation and tooth extraction. Mortuary customs involved ledges for displaying grave goods, coffin chambers, and the burial of animal teeth, pig heads, and pig jawbones.

Ta-wen-k'ou culture

In the middle and lower Yangtze River valley during the 4th and 3rd millennia the Ta-hsi and Ch'ü-chia-ling cultures shared a significant number of traits, including rice production, ring-footed vessels, goblets with sharply angled profiles, ceramic whorls, and black pottery with designs painted in red after firing. Characteristic Ch'ü-chia-ling ceramic objects not generally found in Ta-hsi sites include eggshell-thin goblets and bowls painted with black or orange designs; double-waisted bowls; tall, ring-footed goblets and serving stands; and many styles of tripods. Admirably executed and painted clay whorls suggest a thriving textile industry. The chronological distribution of ceramic features suggests a transmission from Ta-hsi to Ch'ü-chia-ling, but the precise relationship between the two cultures has been much debated.

The Ma-chia-pang culture in the T'ai Lake basin was succeeded during the 4th millennium by that of Sung-tse. The pots, increasingly wheel-made, were predominantly clay-tempered gray ware. Tripods with a variety of leg shapes, serving stands, *kuei* pitchers with handles, and goblets with petal-shaped feet were characteristic. Ring feet were used, silhouettes became more angular, and triangular and circular perforations were cut to form openwork designs on the short-stemmed serving stands. A variety of jade ornaments, a feature of Ch'ing-lien-kang culture, has been excavated from Sung-tse burial sites.

Sites of the Liang-chu culture (from the last half of the 4th to the last half of the 3rd millennium) have generally been found in the same area. The pots were mainly wheel-made, clay-tempered gray ware with a black skin and were produced by reduction firing; oxidized red ware was less prevalent. Some of the serving stand and tripod shapes had evolved from Ma-chia-pang prototypes, while other vessel forms included long-necked *kuei* pitchers. The walls of some vessels were black throughout, eggshell-thin, and burnished, resembling those found in Late Neolithic sites in Shantung (see below). Extravagant numbers of highly worked jade *pi* disks and *ts'ung* tubes were placed in certain burials, such as one at Ssu-tun (southern Kiangsu) that contained 57 of them. Liang-chu farmers had developed a characteristic triangular shale plow for cultivating the wet soils of the region. Fragments of woven silk from *c.* 3000 BC have been found at Ch'ien-shang-yang (northern Chekiang). Along the southeast coast and on Taiwan the Ta-p'en-k'eng corded-ware culture emerged during the 4th and 3rd millennia. This culture, with a fuller inventory of pot and tool types than had previously been seen in the area, developed in part from that of Fu-kuo-tun but may also have been influenced by cultures to the west and north, including Ch'ing-lien-kang, Liang-chu, and Liu-lin. The pots were characterized by incised line patterns on neck and rim; low, perforated foot rims; and some painted decoration.

*Regional cultures of the Late Neolithic.* By the 3rd millennium BC the regional cultures in the areas discussed above showed increased signs of interaction and even convergence. That they are frequently referred to as varieties of the Lung-shan culture (*c.* 2500–2000 BC) of east central Shantung—characterized by its lustrous, eggshell-thin black ware—suggests the degree to which these cultures are thought to have experienced eastern influence. That influence, diverse in origin and of varying intensity, entered the North China Plain from sites such as Ta-tun-tzu and Ta-wen-k'ou to the east and also moved up the Han River from the Ch'ü-chia-ling area to the south. A variety of eastern features are evident in the ceramic objects of the period, including use of the fast wheel, unpainted surfaces, sharply angled profiles, and eccentric shapes. There was a greater production of gray and black, rather than red, ware; componential construction was emphasized, in which legs, spouts, and handles were appended to the basic form (which might itself have been built sectionally). Greater elevation was achieved by means of ring feet and tall legs. Ceramic objects included three-legged tripods, steamer cooking vessels, *kuei* pouring pitchers, serving stands, fitted lids, cups and goblets, and asymmetrical *pei hu* vases for carrying water that were flattened on one side to lie against a person's body. In stone and jade objects, eastern influence is evidenced by perforated stone tools and ornaments such as *pi* disks and *ts'ung* tubes used in burials. Other burial customs involved ledges to display the goods buried with the deceased and large wooden coffin chambers. In handicrafts, an emphasis was placed on precise mensuration in working clay, stone, and wood. Although the first, primitive versions of the eastern ceramic types may have been made, on occasion, in the North China Plain, in virtually every case these types were elaborated in the east and given more precise functional definition, greater structural strength, and greater aesthetic coherence. It was evidently the mixing in the 3rd and 2nd millennia of these eastern elements with the strong and extensive traditions native to the North China Plain—represented by such Late Neolithic sites as Ko-la-wang-ts'un (near Cheng-chou), Wang-wan (near Lo-yang), Miao-ti-kou (in central and western Honan), and T'ao-ssu and Teng-hsia-feng (in southwest Shansi)—that stimulated the rise of early Bronze Age culture in the North China Plain and not in the east.

Lung-shan culture

**Religious beliefs and social organization.** The inhabitants of Neolithic China were, by the 5th millennium if not earlier, remarkably assiduous in the attention they paid to the disposition and commemoration of their dead. There was a consistency of orientation and posture, with the dead of the northwest given a westerly orientation and those of the east an easterly one. The dead were segregated, frequently in what appear to be kinship groupings (*e.g.*, at Yuan-chün-miao, Shensi). There were graveside ritual offerings of liquids, pig skulls, and pig jaws (*e.g.*, Pan-p'o and Ta-wen-k'ou), and the demanding practice of collective secondary burial, in which the bones of up to 70 or 80 corpses were stripped of their flesh and reburied together, was extensively practiced as early as the first half of the 5th millennium (*e.g.*, Yuan-chün-miao). Evidence of scapulimantic divination from the end of the 4th millennium (Fu-ho-kou-men, Liaoning) implies the existence of ritual specialists. There was a lavish expenditure of energy by the 3rd millennium on tomb ramps and coffin chambers (*e.g.*, Liu-wan [in eastern Tsinghai] and

Neolithic burial

Ta-wen-k'ou) and on the burial of redundant quantities of expensive grave goods (*e.g.*, Ta-fan-chuang in Shantung, Fu-ch'üan-shan in Shanghai, and Liu-wan), presumably for use by the dead in some afterlife.

Although there is no firm archaeological evidence of a shift from matriliny to patriliny, the goods buried in graves indicate during the course of the 4th and 3rd millennia an increase in general wealth, the gradual emergence of private or lineage property, increasing social differentiation and gender distinction of work roles, and a reduction in the relative wealth of women. The occasional practice of human sacrifice or accompanying-in-death from scattered 4th- and 3rd-millennium sites (*e.g.*, Miao-ti-kou I, Chang-ling-shan in Kiangsu, Ch'in-wei-chia in Kansu, and Liu-wan) suggests that ties of dependency and obligation were conceived as continuing beyond death and that women were likely to be in the dependent position. Early forms of ancestor worship, together with all that they imply for social organization and obligation among the living, were deeply rooted and extensively developed by the Late Neolithic Period. Such religious belief and practice undoubtedly served to validate and encourage the decline of the more egalitarian societies of earlier periods.

## The first historical dynasty: the Shang

**The advent of bronze casting.** The 3rd and 2nd millennia witnessed the appearance of increasing warfare, complex urban settlements, intense status differentiation, and administrative and religious hierarchies that legitimated and controlled the massive mobilization of labour for dynastic work or warfare. The casting of bronze has left the most evident archaeological traces of these momentous changes, but its introduction must be seen as part of a far larger shift in the nature of society as a whole, representing an intensification of the social and religious practices of the Neolithic.

A Chalcolithic Age stretching back to the mid-5th millennium may be dimly perceived. A growing number of 3rd-millennium sites, primarily in the northwest but also in Honan and Shantung, have yielded primitive knives, awls, and drills made of copper and bronze. Stylistic evidence, such as the sharp angles, flat bottoms, and strap handles of certain Ch'i-chia clay pots (in Kansu; *c.* 2250–1900 BC), has led some scholars to posit an early sheet- or wrought-metal tradition possibly introduced from the west by migrating Indo-European peoples, but no wrought-metal objects have been found.

Use of copper and bronze

The construction and baking of the clay cores and sectional piece molds employed in Chinese bronze casting of the 2nd millennium indicate that early metalworking in China rapidly adapted to, if it did not develop indigenously from, the sophisticated, high-heat ceramic technology of the Late Neolithic potters, who were already using ceramic molds and cores to produce forms such as the hollow legs of the *li* caldron. Chinese bronze casting represents, as the continuity in vessel shapes suggests, an aesthetic and technological extension of that ceramic tradition rather than its replacement. The bronze casters' preference for vessels elevated on ring feet or legs further suggests aesthetic links to the east rather than the northwest.

The number, complexity, and size—the Ssu Mu Wu tetrapod weighed 1,925 pounds (875 kilograms)—of the Late Shang ritual vessels reveal high technological competence married to large-scale, labour-intensive metal production. Bronze casting of this scale and character—which placed large groups of ore miners, fuel gatherers, ceramists, and foundry workers under the prescriptive control of the model designers and labour coordinators—must be understood as a manifestation, both technological and social, of the high value that Shang culture placed upon hierarchy, social discipline, and central direction in all walks of life. The prestige of owning these metal objects must have derived in part from the political control over others that their production implied.

The Hsia dynasty

Chinese legends of the 1st millennium BC describe the labours of Yü, the Chinese "Noah" who drained away the floods to render China habitable and established the first Chinese dynasty, called Hsia. Seventeen Hsia kings are listed in the *Shih-chi*, a comprehensive history written during the 1st century BC, and much ingenuity has been devoted to identifying certain Late Neolithic fortified sites—such as Wang-ch'eng-kang ("the mound of the royal city") in north central Honan and Teng-hsia-feng in Hsia *hsien* (thus the site of Hsia-hsü, "the ruins of Hsia"?) in southern Shansi—as early Hsia capitals. T'ao-ssu, also in southern Shansi, has been identified as Hsia for the "royal" nature of five large male burials found there lavishly provided with grave goods. Although they fall within the region traditionally assigned to the Hsia, particular archaeological sites will be hard to identify dynastically unless written records are found. The possibility that Hsia and Shang were partly contemporary, as cultures if not as dynasties, further complicates site identifications. A related approach has been to identify as Hsia an archaeological horizon that lies developmentally between Late Neolithic and Shang strata.

**The Shang dynasty.** The first dynasty to leave historical records is thought to have ruled from the mid-16th to mid-11th century BC. (Some scholars date the Shang dynasty from the mid-18th to the late 12th century BC.) One must, however, distinguish Shang as an archaeological term from Shang as a dynastic one. Erh-li-t'ou in north central Honan, for example, was initially classified archaeologically as Early Shang; its developmental sequence from *c.* 2400 to 1450 BC documents the vessel types and burial customs that link Early Shang culture to the Late Neolithic cultures of the east. In dynastic terms, however, Erh-li-t'ou periods I and II (*c.* 1900 BC?) are now thought by many to represent a pre-Shang (and thus, perhaps, Hsia) horizon. In this view, the two palace foundations, the elite burials, the ceremonial jade blades and sceptres, the bronze axes and dagger axes, and the simple ritual bronzes—said to be the earliest yet found in China—of Erh-li-t'ou III (*c.* 1700–1600 BC?) signal the advent of the dynastic Shang.

Middle Shang sites

The archaeological classification of Middle Shang is represented by the remains found at Erh-li-kang (*c.* 1600 BC) near Cheng-chou, some 50 miles (80 kilometres) to the east of Erh-li-t'ou. The massive rammed-earth fortification, 118 feet (36 metres) wide at its base and enclosing an area of 1.2 square miles (3.2 square kilometres), would have taken 10,000 men more than 12 years to build. Also found were ritual bronzes, including four monumental tetrapods (the largest weighing 190 pounds; palace foundations; workshops for bronze casting, pot making, and bone working; burials; and two inscribed fragments of oracle bones. Another rammed-earth fortification, enclosing about 0.7 square mile and also dated to the Erh-li-kang period, has been found at Yen-shih, about three miles east of the Erh-li-t'ou III palace foundations. While these walls and palaces have been variously identified by modern scholars—the identification now favoured is of Cheng-chou as Po, the capital of the Shang dynasty during the reign of T'ang, the dynasty's founder—their dynastic affiliations are yet to be firmly established. The presence of two large, relatively close contemporary fortifications at Cheng-chou and Yen-shih, however, indicates the strategic importance of the area and impressive powers of labour mobilization.

P'an-lung-ch'eng in Hupeh, 280 miles south of Cheng-chou, is an example of Middle Shang expansion into the northwest, northeast, and south. A city wall, palace foundations, burials with human sacrifices, bronze workshops, and mortuary bronzes of the Erh-li-kang type form a complex that duplicates on a smaller scale Cheng-chou. A transitional period spanning the gap between the Upper Erh-li-kang phase of Middle Shang and the Yin-hsü phase of Late Shang indicates a widespread network of Shang cultural sites that were linked by uniform bronze-casting styles and mortuary practices. A relatively homogeneous culture united the Bronze Age elite through much of China around the 14th century BC.

Ruins of Yin

The Late Shang period is best represented by a cluster of sites focused on the village of Hsiao-t'un, west of An-yang in northern Honan. Known to history as Yin-hsü, "the Ruins of Yin" (Yin was the name used by the succeeding Chou dynasty for the Shang), it was a seat of royal power for the last nine Shang kings, from Wu-ting to Ti-hsin.

According to the "short chronology" used here, which is based upon modern studies of lunar eclipse records and reinterpretations of Chou annals, these kings would have reigned *c.* 1200–1045 BC. (One version of the traditional "long chronology," based primarily upon a 1st-century-BC source, would place the last 12 Shang kings, from P'an-keng onward, at Yin-hsü from 1398 to 1112 BC.) Sophisticated bronze, ceramic, stone, and bone industries were housed in a network of settlements surrounding the unwalled cult centre at Hsiao-t'un, which had rammed-earth temple-palace foundations. And Hsiao-t'un itself lay at the centre of a larger network of Late Shang sites—such as Hsing-t'ai to the north and Hsin-hsiang to the south—in southern Hopeh and northern Honan.

*Royal burials.* The royal cemetery lay less than two miles northwest of Hsiao-t'un, at Hsi-pei-kang. The hierarchy of burials at this and other cemeteries in the area reflected the social organization of the living. Large pit tombs, some nearly 42 feet deep, were furnished with four ramps and massive grave chambers for the kings. Retainers who accompanied their lords in death lay in or near the larger tombs; members of the lesser elite and commoners were buried in pits that ranged from medium size to shallow; those of still lower status were thrown into refuse pits and disused wells; and human and animal victims of the royal mortuary cult were placed in sacrificial pits. Only a few undisturbed elite burials have been unearthed, the most notable being that of Fu Hao, a consort of Wu-ting. That her relatively small grave contained 468 bronze objects, 775 jades, and more than 6,880 cowries suggests how great the wealth placed in the far larger royal tombs must have been.

*The chariot.* The light chariot, with 18 to 26 spokes per wheel, first appeared, according to the archaeological and inscriptional record, around 1200 BC. Glistening with bronze, it was initially a prestigious command car used primarily in hunting. The 16 chariot burials found at Hsiao-t'un raise the possibility of some form of Indo-European contact with China, and there is little doubt that the chariot, which probably originated in the Caucasus, entered China via Central Asia and the northern steppe. Animal-headed knives, always associated with chariot burials, are further evidence of a northern connection.

*Art.* Late Shang culture is also defined by the size, elaborate shapes, and evolved decor of the ritual bronzes, many of which were used in wine offerings to the ancestors and some of which were inscribed with ancestral dedications such as "Made for Father Ting." Their surfaces were ornamented with zoomorphic and theriomorphic elements set against intricate backgrounds of geometric meanders, spirals, and quills. Some of the animal forms—which include tigers, birds, snakes, dragons, cicadas, and water buffalo—have been thought to represent shamanistic familiars or emblems that ward away evil. The exact meaning of the iconography, however, may never be known. That the predominant *t'ao-t'ieh* monster mask—with bulging eyes, fangs, horns, and claws—may have been anticipated by designs carved on jade *ts'ung* tubes and axes from Liang-chu culture sites in the Yangtze Delta and from the Late Neolithic in Shantung suggests that its origins were ancient. But the degree to which pure form or intrinsic meaning took priority, in either Neolithic or Shang times, is hard to assess.

The *t'ao-t'ieh* mask

*Late Shang divination and religion.* Although certain complex symbols painted on Late Neolithic pots from Shantung suggest that primitive writing was emerging in the east in the 3rd millennium, the Shang divination inscriptions that appear at Hsiao-t'un form the earliest body of Chinese writing yet known. In Late Shang divination as practiced during the reign of Wu-ting (*c.* 1200–1180 BC), cattle scapulae or turtle plastrons, in a refinement of Neolithic practice, were first planed and bored with hollow depressions to which an intense heat source was then applied. The resulting T-shaped stress cracks were interpreted as lucky or unlucky. After the prognostication had been made, the day, the name of the presiding diviner (some 120 are known), the subject of the charge, the prognostication, and the result might be carved into the surface of the bone. Among the topics divined were sacrifices, campaigns, hunts, the good fortune of the 10-day week or of the night or day, weather, harvests, sickness, childbearing, dreams, settlement building, the issuing of orders, tribute, divine assistance, and prayers to various spirits. Some evolution in divinatory practice and theology evidently occurred. By the reigns of the last two Shang kings, Ti-i and Ti-hsin (*c.* 1100 to 1045 BC), the scope and form of Shang divination had become considerably simplified: prognostications were uniformly optimistic, and divination topics were limited mainly to the sacrificial schedule, the coming 10 days, the coming night, and hunting.

Early writing

*State and society.* The ritual schedule records 29 royal ancestors over a span of 17 generations who, from at least Wu-ting to Ti-hsin, were each known as *wang* (king). Presiding over a stable politico-religious hierarchy of ritual specialists, officers, artisans, retainers, and servile peasants, they ruled with varying degrees of intensity over the North China Plain and parts of Shantung, Shansi, and Shensi, mobilizing armies of at least several thousand men as the occasion arose.

The worship of royal ancestors was central to the maintenance of the dynasty. The ancestors were designated by 10 "stem" names (*chia, i, ping, ting,* etc.) that were often prefixed by kin titles, such as "father" and "grandfather," or by status appellations, such as "great" or "small." The same stems were used to name the 10 days (or suns) of the week, and ancestors received cult on their name days according to a fixed schedule, particularly after the reforms of Tsu-chia. For example, Ta-i ("Great I," the sacrificial name of T'ang, the dynasty founder) was worshiped on *i* days, Wu-ting on *ting* days. The Shang dynastic group, whose lineage name was Tsu (according to later sources), appears to have been divided into 10 units corresponding to the 10 stems. Succession to the kingship alternated on a generational basis between two major groupings of *chia* and *i* kings on the one hand and *ting* kings on the other. The attention paid in the sacrificial system to the consorts of "great lineage" kings—who were themselves both sons (possibly nephews) and fathers (possibly uncles) of kings—indicates that women may have played a key role in the marriage alliances that ensured such circulation of power.

The goodwill of the ancestors, and of certain river and mountain powers, was sought through prayer and offerings of grain, millet wine, and animal and human sacrifice. The highest power of all, with whom the ancestors mediated for the living king, was the relatively remote deity Ti, or Shang Ti, "the Lord on High." Ti controlled victory in battle, harvest, the fate of the capital, and the weather, but, on the evidence of the oracle bone inscriptions, he received no cult. This suggests that Ti's command was too inscrutable to be divined or influenced; he was, in all likelihood, an impartial figure of last theological resort, needed to account for inexplicable events.

The deity Ti

Although Marxist historians have categorized the Shang as a slave society, it would be more accurate to describe it as a dependent society. The king ruled a patrimonial state in which royal authority, treated as an extension of patriarchal control, was embedded in kinship and kinship-like ties. Despite the existence of such formal titles as "the many horse" or "the many archers," administration was apparently based primarily on kinship alliances, generational status, and personal charisma. The intensity with which ancestors were worshiped suggests the strength of the kinship system among the living; the ritualized ties of filiation and dependency that bound a son to his father, both before and after death, are likely to have had profound political implications for society as a whole. This was not a world in which concepts such as freedom and slavery would have been readily comprehensible. Everybody, from king to peasant, was bound by ties of obligation—to former kings, to ancestors, to superiors, and to dependents. The routine sacrificial offering of human beings, usually prisoners from the Ch'iang tribe, as if they were sacrificial animals, and the rarer practice of accompanying-in-death, in which 40 or more retainers, often of high status, were buried with a dead king, suggest the degree to which ties of affection, obligation, or servitude were thought to be stronger than life itself. If slavery existed, it was psychological and ideological, not legal. The

political ability to create and exploit ties of dependency originally based on kinship was one of the characteristic strengths of early Chinese civilization.

Such ties were fundamentally personal in nature. The king referred to himself as *yü i jen,* "I, the one man," and he was, like many early monarchs, peripatetic. Only by traveling through his domains could he ensure political and economic support. These considerations, coupled with the probability that the position of king circulated between social or ritual units, suggest that, lacking a national bureaucracy or effective means of control over distance, the dynasty was relatively weak. The Tzu should, above all, be regarded as a politically dominant lineage that may have displaced the Ssu lineage of the Hsia and that was in turn to be displaced by the Chi lineage of the Chou. But the choices that the Shang made—involving ancestor worship, the politico-religious nature of the state, patrimonial administration, the mantic role of the ruler, and a pervasive sense of social obligation—were not displaced. These choices endured and were to define, restrict, and enhance the institutions and political culture of the full-fledged dynasties yet to come. (D.N.K.)

## The Chou and Ch'in dynasties

### THE HISTORY OF THE CHOU (1111–255 BC)

The Chou ancestors

The origin of the Chou royal house is lost in the mists of time. Although the traditional historical system of the Chinese contains a Chou genealogy, no dates can be assigned to the ancestors. The first ancestor was Hou Chi, literally translated as "Lord of Millet." He appears to have been a cultural hero and agricultural deity rather than a tribal chief. The earliest plausible Chou ancestor was Tan Fu, the grandfather of Wen-wang. Prior to and during the time of Tan Fu, the Chou people seem to have migrated to avoid pressure from strong neighbours, possibly nomadic people to the north. Under the leadership of Tan Fu they settled in the valley of the Wei River in the present province of Shensi. The fertility of the loess soil there apparently made a great impression on these people, who had already been engaged in farming when they entered their new homeland. A walled city was built, and a new nation was formed. Archaeological remains, including ruins of courtyards surrounded by walls and halls on platforms, confirm literary evidence of a Chou state.

**Chou and Shang.** The name Chou appears often in the oracle bone inscriptions of the Shang kingdom, sometimes as a friendly tributary neighbour and at other times as a hostile one. This pattern is confirmed by records found at the Chou archaeological site. Marriages were occasionally made between the two ruling houses. The Chou also borrowed such arts as bronze casting from their more cultivated neighbour. The Chou royal house, however, had already conceived the idea of replacing Shang as the master of China—a conquest that took three generations. Although the conquest was actually carried out by his sons, Wen-wang should be credited with molding the Chou kingdom into the most formidable power west of the Shang. Wen-wang extended the Chou sphere of influence to the north of the Shang kingdom and also made incursions to the south, thus paving the way for the final conquest by Wu-wang.

Wen-wang and Wu-wang

In Chinese historical tradition Wen-wang was depicted as intelligent and benevolent, a man of virtue who won popularity among his contemporaries and expanded the realm of the Chou. His son, Wu-wang, though not as colourful as his father, was always regarded as the conqueror. In fact, Wu, his posthumous name, means "Martial." But the literary records indicate that the Chou actually controlled two-thirds of all China at the time of Wen-wang, who continued to recognize the cultural and political superiority of the Shang out of feudal loyalty. There is not enough evidence either to establish or to deny this. A careful historian, however, tends to take the Chou subjugation to the Shang as a recognition of Shang strength. It was not until the reign of the last Shang ruler, Chou, that the kingdom exhausted its strength by engaging in large-scale campaigning against nomads in the north and against a group of native tribes in the east. At this time Wu-wang organized the first probing expedition and reached the neighbourhood of the Shang capital. A full-scale invasion soon followed. Along with forces of the Chou, the army was made up of the Chiang, southern neighbours of the Chou, and of eight allied tribes from the west. The Shang dispatched a large army to meet the invaders. The pro-Chou records say that, after the Shang vanguard defected to join the Chou, the entire army collapsed, and Wu-wang entered the capital without resistance. Yet Mencius, the 4th-century-BC thinker, cast doubt upon the reliability of this account by pointing out that a victory without enemy resistance should not have been accompanied by the heavy casualties mentioned in the classical document. One may speculate that the Shang vanguard consisted of remnants of the eastern tribes suppressed by the Shang ruler Chou during his last expedition and that their sudden defection caught the Shang defenders by surprise, making them easy prey for the invading enemy. The decisive battle took place in 1111 BC (as tabulated by Tung Tso-pin, although it is commonly dated at 1122, and other dates have also been suggested). Wu-wang died shortly after the conquest, leaving a huge territory to be consolidated. This was accomplished by one of his brothers, Chou Kung, who served as regent during the reign of Wu's son, Ch'eng-wang.

The defeated Shang could not be ruled out as a potential force, even though their ruler, Chou, had immolated himself. Many groups of hostile "barbarians" were still outside the sphere of Chou power. The Chou leaders had to yield to reality by establishing a rather weak control over the conquered territory. The son of Chou was allowed to organize a subservient state under the close watch of two other brothers of Wu-wang, who were garrisoned in the immediate vicinity. Other leaders of the Chou and their allies were assigned lands surrounding the old Shang domain. But no sooner had Chou Kung assumed the role of regent than a large-scale rebellion broke out. His two brothers, entrusted with overseeing the activities of the son of Chou, joined the Shang prince, and it took Chou Kung three full years to reconquer the Shang domain, subjugate the eastern tribes, and reestablish the suzerainty of the Chou court.

These three years of extensive campaigning consolidated the rule of the Chou over all of China. An eastern capital was constructed on the middle reach of the Huang Ho (Yellow River) as a stronghold to support the feudal lords in the east. Several states established by Chou kinsmen and relatives were transferred further east and northeast as the vanguard of expansion, including one established by the son of Chou Kung. The total number of such feudal states mentioned in historical records and later accounts varies from 20 to 70; the figures in later records would naturally be higher, since enfeoffment might take place more than once. Each of these states included fortified cities. They were strung out along the valley of the Huang Ho between the old capital and the new eastern capital, reaching as far as the valleys of the Huai and Han rivers in the south and extending eastward to the Shantung Peninsula and the coastal area north of it. All these colonies mutually supported each other and were buttressed by the strength of the eastern capital, where the conquered Shang troops were kept, together with several divisions of the Chou legions. Ancient bronze inscriptions make frequent mention of mobilizing the military units at the eastern capital at times when the Chou feudal states needed assistance.

The Chou states

**The Chou feudal system.** The feudal states were not contiguous but, rather, were scattered at strategic locations surrounded by potentially dangerous and hostile nations. The fortified city of the feudal lord was often the only area that he controlled directly; the state and the city were therefore identical, both being *kuo,* a combination of city wall and weapons. Satellite cities were established at convenient distances from the main city in order to expand the territory under control. Each feudal state consisted of an alliance of the Chou, the Shang, and the local population. A Chinese nation was formed on the foundation of Chou feudalism.

The scattered feudal states gradually acquired something like territorial solidity as the neighbouring populations established closer ties with them, either by marriage or

by accepting vassal status; the gaps between the fortified cities were thus filled by political control and by cultural assimilation. This created a dilemma for the Chou central court: the evolution of the feudal network buttressed the structure of the Chou order, but the strong local ties and parochial interests of the feudal lords tended to pull them away from the centre. Each of these opposing forces became at one time or another strong enough to affect the history of the Chou order.

For about two centuries Chou China enjoyed stability and peace. There were wars against the non-Chou peoples of the interior and against the nomads along the northern frontier, but there was little dispute among the Chinese states themselves. The southern expansion was successful, and the northern expansion worked to keep the nomads away from the Chinese areas. The changing strength of the feudal order can be seen from two occurrences at the Chou court. In 841 BC the nobles jointly expelled Li-wang, a tyrant, and replaced him with a collective leadership headed by the two most influential nobles until the crown prince was enthroned. In 771 BC the Chou royal line was again broken when Yu-wang was killed by invading barbarians. The nobles apparently were split at this time, because two courts ensued, headed by two princes, each of whom had the support of part of the nobility. One of the pretenders, P'ing-wang, survived the other, but the royal order had lost prestige and influence. The cohesion of the feudal system had weakened. Thereafter, it entered a new phase traditionally known as the Ch'un-ch'iu ("Spring and Autumn") period (770–476 BC).

The Ch'un-ch'iu period

In the Ch'un-ch'iu period there was a gradual dilution of the familial relationship among the nobles. A characteristic of the Chou feudal system was that the extended family and the political structure were identical. The line of lordship was regarded as the line of elder brothers, who, therefore, enjoyed not only political superiority but also seniority in the family line. The head of the family not only was the political chief but also had the unique privilege of offering sacrifice to and worshiping the ancestors, who would bestow their blessings and guarantee the continuity of the mandate of Heaven. After the weakening of the position of the Chou king in the feudal structure, he was not able to maintain the position of being the head of a big family in any more than a normal sense. The feudal structure and familial ties fell apart, continuing in several of the Ch'un-ch'iu states for various lengths of time with various degrees of modification. Over the next two centuries the feudal-familial system gradually declined and disappeared.

Social ranks

In the first half of the Ch'un-ch'iu period the feudal system was a stratified society, divided into ranks as follows: the ruler of a state; the feudal lords who served at the ruler's court as ministers; the *shih* (roughly translated as "gentlemen") who served at the households of the feudal lords as stewards, sheriffs, or simply warriors; and, finally, the commoners and slaves. The state ruler and the ministers were clearly a superior class, and the commoners and slaves were an inferior class; the class of *shih* was an intermediate one in which the younger sons of the ministers, the sons of *shih,* and selected commoners all mingled to serve as functionaries and officials. The state rulers were, in theory, divided into five grades; in reality, the importance of a ruler was determined by the strength of his state. The ministerial feudal lords, however, often had two or three grades among themselves, as determined by the lord-vassal relationship. In general, each state was ruled by a group of hereditary feudal lords who might or might not be of the same surname as the state ruler. The system was not stable in the Ch'un-ch'iu period, and everywhere there were changes.

The first important change occurred with the advent of interstate leadership. For several decades after 722 BC the records show chiefly battles and diplomatic maneuvers among the states on the central plain and in the middle and lower reaches of the Huang Ho valley. These states, however, were too small to hold the leadership and too constricted in the already crowded plain to have potentiality for further development. The leadership was soon taken over by states on the peripheral areas.

The first to achieve this leadership was Huan Kung (reigned 685–643 BC), the ruler of the state of Ch'i on the Shantung Peninsula. He successfully rallied around him many other Chinese states to resist the pressure of non-Chinese powers in the north and south. While formally respecting the suzerainty of the Chou monarchy, Huan Kung adopted a new title of "overlord" (*pa*). He convened interstate meetings, settled disputes among states, and led campaigns to protect his followers from the intimidation of non-Chinese powers.

Rivalry among the Chou states

After his death the state of Ch'i failed to maintain its leading status. The leadership, after a number of years, passed to Wen Kung of Chin (reigned 636–628 BC), the ruler of the mountainous state north of the Huang Ho. Under Wen Kung and his capable successors, the overlordship was institutionalized until it took the place of the Chou monarchy. Interstate meetings were held at first during emergencies caused by challenges from the rising southern state of Ch'u. States answering the call of the overlord were expected to contribute and maintain a certain number of war chariots. Gradually the meetings became regular, and the voluntary contribution was transformed into a compulsory tribute to the court of the overlord. The new system of states under the leadership of an overlord developed not only in North China under Chin but also in the South under Ch'u. Two other states, Ch'in and Ch'i, while not commanding the strength of the formidable Chin and Ch'u, each absorbed weaker neighbours into a system of satellite states. A balance of power emerged among the four states of Ch'i, Ch'in, Chin, and Ch'u. The balance was occasionally tipped when two of them went to war, but it was subsequently restored by the transference of some small states from one camp to another.

Breakdown of the system

A further change began in the 5th century BC when the states of Wu and Yüeh far to the south suddenly challenged Ch'u for hegemony over the southern part of China, at a time when the strong state of Chin was much weakened by an internecine struggle among powerful magnates that subsequently split Chin into three contending powers. Wu got so far as to claim overlordship over North China in an interstate meeting held in 482 BC after defeating Ch'u. But Wu's hegemony was short-lived; it collapsed after being attacked by Yüeh. Although Yüeh held the nominal overlordship for a brief period, Chin, Ch'in, and Ch'i were weakened by internal disturbances and declined, and a series of defeats paralyzed Ch'u; thus the balance-of-power system was rendered unworkable.

A half century of disorder followed. Small states fell prey to big ones, while in the big states usurpers replaced the old rulers. When the chaos ended, there were seven major powers and half a dozen minor ones. Among the seven major powers, Chao, Han, and Wei had formerly been parts of Chin; the Ch'i ruling house had changed hands; and Ch'in was undergoing succession problems. The only "old" state was Ch'u. Even Ch'u, a southern state, had become almost completely assimilated to the northern culture (except in art, literature, and folklore). The minor powers had also changed: some had retained only small portions of their old territories, some had new ruling houses, and some were new states that had emerged from non-Chinese tribes. The long period of power struggle that followed is known as the Chan-kuo ("Warring States") period (475–221 BC).

## SOCIAL, POLITICAL, AND CULTURAL CHANGES

The years from the 8th century BC to 221 BC witnessed the painful birth of a unified China. It was a period of bloody wars and also of far-reaching changes in politics, society, and intellectual outlook.

**The decline of feudalism.** The most obvious change in political institutions was the replacement of the old feudal structure by systems of incipient bureaucracy under monarchy. The decline of feudalism took its course in the Ch'un-ch'iu period, and the rise of the new order may be seen in the Chan-kuo period. The Chou feudalism suffered from a continual dilution of authority. As a state expanded, its nobility acquired vassals, and these in turn acquired their own vassals. The longer this went on, the more diluted became the family tie, and the more

dependent the ruler became on the combined strength of the vassals. At a certain point the vassals might acquire an advantageous position, and the most dominant figures among them might eclipse the king. The Chou royal house perhaps reached the turning point earlier than the other feudal states. The result was the shrinkage of the Chou royal domain and royal influence when P'ing-wang moved his court to the east. The ruling houses of other states suffered the same fate. Within a century after the Chou court moved to the east, the ruling houses in most of the feudal states had changed. Sometimes a dominating branch replaced the major lineage; sometimes a powerful minister formed a strong vassaldom and usurped the authority of the legitimate ruler. Bloody court intrigues and power struggles eliminated many established houses. The new power centres were reluctant to see the process continue and therefore refused to allow further segmentation and subinfeudation. Thus, the feudal system withered and finally collapsed.

**Urbanization and assimilation.** At the same time a process of urbanization was occurring. Minor fortified cities were built, radiating out from each of the major centres, and other towns radiated from the minor cities. From these cities and towns, orders were issued, and to them the resources of the countryside were sent. The central plain along the Huang Ho was the first to be saturated by clusters of cities. This is probably the reason why the central states soon reached the maximum of their influence in the interstate power struggle: unlike the states in peripheral areas, they had no room to expand.

Assimilation of non-Chou peoples

The period of urbanization was also a period of assimilation. The non-Chou population caught in the reach of feudal cities could not but feel the magnetic attraction of the civilization represented by the Chou people and Chou feudalism. The bronze inscriptions of the Hsi (Western) Chou (*c.* 1111–771 BC) refer to the disturbances of the "barbarians," who could be found practically everywhere. Those were the non-Chou groups scattered in the open spaces. The barbarians in inland China were forced to integrate with one or another of the contenders in the interstate conflicts. Their lands were annexed; their populations, moved or absorbed. The strength of the large states owed much to their success at incorporating these non-Chinese groups. By the time of the unification of China in the 3rd century BC, there was virtually no significant concentration of non-Chinese groups north of the Yangtze River valley and south of the steppe. Bronze pieces attributable to non-Chou chiefs in the late Ch'un-ch'iu period show no significant difference in writing system and style from those of the Chinese states.

Chou civilization was not assimilated so easily in the south, where the markedly different Ch'u culture flourished. For some centuries Ch'u was the archenemy of the Chinese states, yet the nobles of the Ch'u acquired enough of the northern culture to enable their envoy to the courts of the north to cite the same verses and observe the same manners. The Ch'u literature that has survived is the fruit of these two distinctive heritages.

To the north were the nomadic peoples of the steppe. As long as they remained divided, they constituted no threat, but under strong leaders, able to forge a nomadic empire challenging the dominance of the Chinese, there were confrontations. The "punitive" action into the north during the reign of Hsüan-wang (827–781 BC) does not seem to have been very large in scope; both sides apparently had little ambition for territorial aggrandizement. Cultural exchange in the northern frontier region was far less than the assimilation that occurred in the south along the Yangtze valley, and it was mainly concerned with techniques of cavalry warfare.

**The rise of monarchy.** As states grew in both population and area, there were internal political changes. The most basic change was in the pattern of the delegation of power. Under feudalism, authority had been delegated by the lord to the vassal. The new state rulers sought ways of maintaining and organizing their power.

In the state of Chin the influence of kinsmen of the ruling house had been trimmed even before Wen Kung established his overlordship. Wen Kung reorganized the government, installing his most capable followers in the key posts. He set up a hierarchical structure that corresponded to the channels of military command. Appointments to these key positions came to be based on a combination of merit and seniority, thus establishing a type of bureaucracy that was to become traditional in Chinese government.

Establishment of central controls

The Ch'u government was perhaps the oldest true monarchy among all the Ch'un-ch'iu states. The authority of the king was absolute. Ch'u was the only major state in which the ruling house survived the chaotic years of the Chan-kuo period.

Local administration went through a slow evolution. The prefecture system developed in both Chin and Ch'u was one innovation. In Chin there were several dozens of prefects across the state, having limited authority and limited tenure. The Chin prefect was no more than a functionary, in contrast to the feudal practice. In Ch'u similar local administrative units grew up. New lands taken by conquest were organized into prefectures governed by ranking officials who were evidently appointed by the king. The prefecture system of Chin and Ch'u was to become the principal form of local administration in the Chan-kuo period.

By that time practically all the major states had chancellors. A chancellor acted as the leader of the court, which was composed of numerous officials. Whereas in the feudal state the officials had been military officers, the more functionally differentiated court of the Chan-kuo period usually had a separate corps of civil service personnel. Local administration was entrusted to prefects, who served limited terms. Prefects were often required to submit annual reports to the court so that the ruler could judge their performance. Regional supervisors were sometimes dispatched to check the work of the prefects, a system developed by the later Chinese Imperial government into the "censor" system. Fiefs of substantial size were given to very few people, usually only to close relatives of the ruler. There was little opportunity for anyone to challenge the sovereignty of the state. The majority of government employees were not relatives of the ruler, and some of them might not even be citizens of the state. Officials were paid in grain or perhaps in a combination of cash and grain. Archives kept by scribes were on wooden blocks and bamboo strips. These features indicate the emergence of some form of bureaucracy.

The new pattern was the result of the efforts of many reformers in different states. Both practical men and theoreticians helped to form the new structure, which, though still crude, was the forerunner of the large and complex bureaucracy of later Chinese dynasties.

Military technique also underwent great changes in the Chan-kuo period. In the feudal era, war had been a profession of the nobles. Lengthy training was needed to learn the technique of driving and shooting from a chariot drawn by horses. There was also an elaborate code of behaviour in combat. The nature of war had changed by the late Ch'un-ch'iu period. The nobility had given way to professional warriors and mercenaries. In some states special titles of nobility were created for successful warriors, regardless of their origin. Foot soldiers were replacing war chariots as the main force on the battlefield. The expansion of the major states into mountainous areas and the rise of the southern powers in an area of swamps, lakes, and rivers increased the importance of the infantry.

New methods of warfare

Battles were fought mostly by hordes of foot soldiers, most of them commoners, aided by cavalry units; war chariots apparently served only auxiliary roles, probably as mobile commanding platforms or perhaps as carriers. All the Chan-kuo powers seem to have had conscription systems to recruit able-bodied male citizens. The organization, training, and command of the infantry required experts of a special type. The Chan-kuo period produced professional commanders who conducted battles involving several thousand men, with lines extending hundreds of miles. A few treatises on the principles of warfare still survive, including *Ping-fa* (*The Art of War*), by Sun-tzu. Cavalry warfare developed among the northern states, including Ch'in, Chao, and Yen. The Ch'in cavalrymen were generally drawn from the northern and northwest-

ern border areas, where there were constant contacts with the steppe peoples. The rise of Yen from a rather obscure state to a major power probably owed much to its successful adoption of cavalry tactics, as well as to its northern expansion.

Agriculture

**Economic development.** Important changes occurred in agriculture. Millet had once been the major cereal food in the north, but gradually wheat grew in importance. Rice, imported from the south, was extended to the dry soil of the north. The soybean in a number of varieties proved to be one of the most important crops. Chinese farmers gradually developed a kind of intensive agriculture. Soil was improved with the use of organic fertilizers. The fallow system was replaced by planting in carefully regulated rows. The importance of plowing and seeding at the proper time was stressed (especially in the fine-grained loess soil of North China). Frequent weeding was done throughout the growing season. Farmers also knew the value of rotating crops to preserve the fertility of the soil; soybeans were often part of the rotation. Although iron had been used to cast implements in the 5th century BC (probably even as early as the 8th century BC), those discovered by archaeologists are of rather inferior quality.

As population pressure forced the extension of cropland, irrigation became necessary. In the late Ch'un-ch'iu period and in the Chan-kuo period, irrigation works were constructed in many states. These projects were built to drain swampy areas, to leach out alkaline soil and replace it with fertile topsoil, and in the south and in the Szechwan Basin, to carry water into the rice paddies. The irrigation systems unearthed by modern archaeologists indicate that these were small-scale works carried out for the most part by state or local authorities.

Trade

Another significant change in the economic sphere was the growth of trade among regions. Coins excavated in scattered spots show, by their great variety, that active trade had been extended to all parts of Chou China. Great commercial centres had arisen, and the new cities brought a demand for luxuries. The literary records as well as the archaeological evidence show that wealthy persons had possessions made of bronze and gold, silver inlays, lacquer, silk, ceramics, and precious stones. The advancement of ferrous metallurgy led to the earliest recorded blast furnace and the earliest steel. The Chinese had been casting bronze for more than 1,000 years; turning to iron, they became very skillful at making weapons and tools. The Han historian Ssu-ma Ch'ien (writing about 100 BC) told of men making fortunes in the iron industry.

The Chan-kuo period witnessed the demise of the old feudal regimes and their replacement by centralized monarchies. The feudal nobility fell victim to power struggles within the states and to conquest by stronger states. During the Ch'un-ch'iu period these parallel processes drastically reduced the numbers of the nobility.

Emergence of a new intellectual elite

In the late Ch'un-ch'iu period there arose a new elite class, composed of the former *shih* class and the descendants of the old nobility. The members of this class were distinguished by being educated, either in the literary tradition or in the military arts. The *shih* provided the administrators, teachers, and intellectual leaders of the new society. The philosophers Confucius (551–479 BC), Mencius (*c.* 371–289 BC), Mo-tzu (5th century BC), and Hsün-tzu (*c.* 298–*c.* 230 BC) were members of the *shih* class, as were also a large proportion of high-ranking officials and leaders of prominence. The interstate competition that drove rulers to select the most capable and meritorious persons to serve in their courts resulted in an unprecedented degree of social mobility.

The populace, most of whom were farmers, also underwent changes in status. In feudal times the peasants had been subjects of their lords. They owned no property, being at most permitted to till a piece of the lord's land for their own needs. The ancient texts tell of the "well-field" system, under which eight families were assigned 100 *mou* (15 acres, or six hectares) each of land to live on while collectively cultivating another 100 *mou* as the lord's reservation. As farming became more intensive, there was a transition to individual ownership. This can be seen in the growth of the practice of taxing farmers according to the amount of land that they owned. By the time of the Chan-kuo period, the land tax had become a common practice. By paying taxes, the tiller of the field acquired the privilege of using the land as his own possession, which perhaps was the first step toward private ownership. As states expanded and as new lands were given to cultivation, an increasing number of "free" farmers were to be found tilling land that had never been part of a lord's manor. With the collapse of the feudal structure, farmers in general gradually ceased to be subjects of a master and became subjects of a state.

A similar transformation occurred among the merchants and craftsmen, who gradually passed from being household retainers of a lord to the status of independent subjects. Thus, the feudal society was completely reshaped in the two centuries preceding the Ch'in unification.

**Cultural change.** These great political and socioeconomic changes were accompanied by intellectual ferment, as the people tried to adjust themselves to a rapidly changing world. Ideas about the proper relationships between members of society were naturally questioned when the old feudal order was shaken; and, in this period, the great teacher Confucius elaborated the social concepts that were to become normative for later Chinese civilization. In place of rigid feudal obligations, he posited an order based on more universal human relationships (such as that between father and son) and taught that ability and moral excellence, rather than birth, were what fitted a man for leadership.

The great thinkers who followed Confucius, whether or not they agreed with his views, were conditioned by his basic assumptions. Mo-tzu, originally a Confucian, based his system on a concept of universal love that was largely an extension of the Confucian idea of humanity; the "worthy man" Mo-tzu recommended as the ideal leader was a development of Confucius' notion of excellence, combining virtue and ability. Even the individualist thinkers known as Taoists, who did not follow Confucius, formulated their teachings as a rebuttal to his system.

Confucius and other pre-Ch'in thinkers viewed the traditional political institutions of China as bankrupt and tried to devise a rationale for something to replace them. Some, such as Confucius, put their main emphasis on the quality of the ruling elite group; others, such as Shang Yang (died 338 BC) and Han-fei-tzu (died 233 BC), regarded a well-organized governing mechanism as the only way to an orderly society. The development of the new centralized monarchical state after the middle of the Ch'un-ch'iu period is not only the embodiment of the ideas of these various thinkers but also the working premise in the context of which they elaborated their theories. The high degree of social and political consciousness that characterized most of the pre-Ch'in philosophical schools set the pattern for the close association of the intellectual with government and society in later China.

The burgeoning commercial life of the period also had its influence in other spheres, especially in the prevalence of contractual relationships. Thus, a minister would roam from one court to another, "selling" his knowledge and service to the most accommodating prince, and the quality of his service was determined by the treatment he received. This kind of contractual relationship remained common in China until the tide of commercialism was ended by the restriction of commercial activity under the Han emperor Wu-ti in the 2nd century BC.

In the Ch'un-ch'iu period the local cultures of China were blended into one common civilization. Through contacts and interchanges, the gods and legends of one region became identified and assimilated with those of other regions. Local differences remained, but, from this time on, the general Chinese pantheon took the form of a congregation of gods with specific functions, a celestial projection of the unified Chinese empire with its bureaucratic society.

Bold challenges to tradition have been rare in Chinese history, and the questioning and innovating spirit of the Ch'un-ch'iu period was to have no parallel until the ferment of the 20th century, after two millennia had elapsed under the domination of Confucian orthodoxy.

## THE CH'IN EMPIRE (221–206 BC)

The early Ch'in

**The Ch'in state.** The history of the Ch'in dynasty may be traced back to the 8th century BC. When the Chou royal house was reestablished at the eastern capital in 770 BC, the Ch'in ruling house, according to the Ch'in historical record, was entrusted with the mission of maintaining order in the previous capital. This may be an exaggeration of the importance of the Ch'in ruling house, and the Ch'in may have been only one of the ruling families of the old nations who recognized Chou suzerainty and went to serve the Chou court. The record is not clear. In the old annals Ch'in did not appear as a significant power until the time of Mu Kung (reigned 659–621 BC), who made Ch'in the main power in the western part of China. Although Ch'in attempted to obtain a foothold in the central heartland along the Huang Ho, it was blocked by the territories of Chin. After a number of failures to enter the eastern bloc of powers, Ch'in had to limit its activities to conquering, absorbing, and incorporating the non-Chinese tribes and states scattered within and west of the big loop of the Huang Ho. Ch'in's success in this was duly recognized by other powers of the Ch'un-ch'iu period, so that the two superpowers Ch'u and Chin had to grant Ch'in, along with Ch'i, the status of overlord in its own region. The eastern powers, however, regarded Ch'in as a "barbarian" state because of the non-Chinese elements it contained.

Ch'in played only a supporting role in the Ch'un-ch'iu power struggle; its location made it immune to the cutthroat competition of the states in the central plain. Ch'in, in fact, was the only major power that did not suffer battle within its own territory. Moreover, being a newly emerged state, Ch'in did not have the burden of a long-established feudal system; this allowed it more freedom to develop its own pattern of government. As a result of being "underdeveloped," it offered opportunity for eastern-educated persons; with the infusion of such talent, it was able to compete very well with the eastern powers, yet without the overexpanded ministerial apparatus that embarrassed other rulers. This may be one reason why Ch'in was one of the very few ruling houses that survived the great turmoil of the late Ch'un-ch'iu period.

A period of silence followed. Even the Ch'in historical record that was adopted by the historian Ssu-ma Ch'ien yields almost no information for a period of some 90 years in the 5th century BC. The evidence suggests that Ch'in underwent a period of consolidation and assimilation during the years of silence. When it reemerged as an important power, its culture appeared to be simpler and more martial, perhaps because of the non-Chinese tribes it had absorbed.

**Struggle for power.** Until the 5th century BC China was dominated by the central-plain power Wei, a successor to Chin, and by the eastern power Ch'i, a wealthy state with a new ruling house. Ch'in remained a secondary power until after the great reforms of Hsiao Kung (361–338 BC) and Shang Yang (Wei Yang).

Shang Yang's reforms

Shang Yang, a frustrated bureaucrat in the court of Wei, went westward seeking a chance to try out his ideas. In the court of Ch'in he established a rare partnership with the ruler Hsiao Kung in the creation of the best organized state of their time. Shang Yang first took strong measures to establish the authority of law and royal decree. The law was to be enforced impartially, without regard to status or position. He convinced Hsiao Kung that the rank of nobility and the privileges attached to it should be awarded only to those who rendered good service to the state, especially for valour in battle. This deprived the existing nobility of their titles and privileges, arousing much antagonism in the court.

One of his most influential reforms was the standardization of local administration. It was a step toward creating a unified state by combining various localities into counties, which were then organized into prefectures under direct supervision of the court. This system was expanded to all China after unification in 221 BC.

Another measure taken by Shang Yang was the encouragement of production, especially in agriculture. Farmers were given incentive to reclaim wasteland; game and fishing reserves were also opened to cultivation. A shortage of labour was met by recruiting able-bodied men from neighbouring states, especially from Han, Chao, and Wei. This policy of drawing workers to Ch'in had two consequences: an increase of production in Ch'in and a loss of manpower in the neighbouring states. In order to increase incentives, the Ch'in government levied a double tax on any male citizen who was not the master of a household. The result was a breakdown of the extended-family system, since younger children were forced to move out and establish their own households. The nuclear family became the prevalent form in Ch'in thereafter. As late as the 2nd century BC, Han scholars were still attacking the Ch'in family structure as failing to observe the principle of filial piety, a cardinal virtue in the Confucian moral code. Shang Yang also standardized the system of weights and measures, a reform of some importance for the development of trade and commerce.

Under the joint labours of Hsiao Kung and Shang Yang, Ch'in grew wealthy and powerful. After Hsiao Kung's death, Shang Yang was put to death by enemies at the Ch'in court. Tablets of the Ch'in law substantiate the survival of Shang Yang's policies after his death.

What remained of the Chou royal court still survived, ruling over a fragmentary domain: poor, weak, and totally at the mercy of the contending powers. It was commonly felt that China ought to be unified politically, although the powers disagreed as to how it was to be done and on who was to be the universal king. Hui-wang, son of Hsiao Kung, claimed the royal title in 325 BC. The adoption of the royal title by Ch'in, of course, was a challenge to Ch'i and Wei. Ch'in pursued a strategy of dividing its rivals and individually defeating them. Ch'in appealed to the self-interest of other powers in order to keep them from intervening in a military action it was taking against one of its neighbours. It befriended the more distant states while gradually absorbing the territories of those close to it.

Ch'in strategy

Within half a century, Ch'in had acquired undisputed predominance over the other contending powers. It continued maneuvering in order to prevent the others from uniting against it. A common topic of debate in the courts of the other states was whether to establish friendly relations with Ch'in or to join with other states in order to resist Ch'in's expansion. The Ch'in strategists were ruthless: all means, including lies, espionage, bribery, and assassination, were pressed into the service of their state.

For a time the eastern power Ch'i had seemed the most likely to win. It defeated Wei, crushed Yen in 314 BC, and annexed Sung in 286 BC. But Ch'i was overturned by an allied force of five states, including Ch'in. Chao, the power with extensive territory in the northern frontier, succeeded Ch'i as the most formidable contender against Ch'in. In 260 BC a decisive battle between Ch'in and Chao destroyed Chao's military strength, although Ch'in was not able to complete its conquest of Chao for several decades.

**The empire.** When Ch'in succeeded in unifying China in 221 BC, its king claimed the title of first sovereign emperor, Shih huang-ti. He was a strong and energetic ruler, and, although he appointed a number of capable aides, the emperor remained the final authority and the sole source of power.

Shih huang-ti made a number of important reforms. He abolished the feudal system completely and extended the administration system of prefectures and counties, with officials appointed by the central government sent into all of China. Circuit inspectors were dispatched to oversee the local magistrates. China was divided into some 40 prefectures. The empire of Shih huang-ti was to become the traditional territory of China. In later eras China sometimes held other territories, but the Ch'in boundaries were always considered to embrace the indivisible area of China proper. In order to control this vast area, Shih huang-ti constructed a network of highways for the movement of his troops. Several hundred thousand workers were conscripted to connect and strengthen the existing walls along the northern border. The result was a complex of fortified walls, garrison stations, and signal towers extending from the Po Hai (Gulf of Chihli) westward across the pastureland of what is today Inner Mongolia and through the fertile loop of the Huang Ho to the edge of Tibet.

The Great Wall

This defense line, known as the Great Wall, marked the frontier where the nomads of the great steppe and the Chinese farmers on the loess soil confronted each other. Yet the Emperor failed in another great project, the digging of a canal across the mountains in the south to link the southern coastal areas with the main body of China. Shih huang-ti, with his capable chancellor Li Ssu, also unified and simplified the writing system and codified the law.

All China felt the burden of these 11 or 12 years of change. Millions of men were dragooned to the huge construction jobs, many dying on the long journey to their destination. Rich and influential men in the provinces were compelled to move to the capital. Weapons were confiscated. Hundreds of intellectuals were massacred for daring to criticize the Emperor's policies. Books dealing with subjects other than law, horticulture, and herbal medicine were kept out of public circulation because the Emperor considered such knowledge to be dangerous and unsettling. These things have contributed to make Shih huang-ti appear the archtyrant of Chinese history.

Some of the accusations leveled against him by historians are perhaps exaggerated, such as the burning of books and the indiscriminate massacre of intellectuals. Shih huang-ti himself claimed in the stone inscriptions of his time that he had corrected the misconduct of a corrupted age and given the people peace and order. Indeed, his political philosophy did not deviate much from that already developed by the great thinkers of the Chan-kuo period and adopted later by the Han emperors, who have been generally regarded as benevolent rulers. Shih huang-ti was afraid of death. He did everything possible to achieve immortality. Deities were propitiated, and messengers were dispatched to look for an elixir of life. He died in 210/209 BC while on a tour of the empire. Excavation of his tomb, near modern Sian (ancient Ch'ang-an), revealed more than 6,000 life-sized statues of soldiers still on guard.

Fall of the Ch'in dynasty

His death led to the fall of his dynasty. The legitimate heir was compelled to commit suicide when his younger brother usurped the throne. Capable and loyal servants, including Li Ssu and Gen. Meng T'ien, were put to death. Ehr-shih ti, the second emperor, reigned only four years. Rebellion broke out in the Yangtze River area when a small group of conscripts led by a peasant killed their escort officers and claimed sovereignty for the former state of Ch'u. The uprising spread rapidly as old ruling elements of the six states rose to claim their former titles. Escaped conscripts and soldiers who had been hiding everywhere emerged in large numbers to attack the Imperial armies. The second emperor was killed by a powerful eunuch minister, and in 206 BC a rebel leader accepted the surrender of the last Ch'in prince. (C.-y.H.)

## The Han dynasty

The Han dynasty was founded by Liu Pang (best known by his temple name, Kao-tsu), who assumed the title of emperor in 202 BC. Eleven members of the Liu family followed in his place as effective emperors until AD 9. In that year the dynastic line was challenged by Wang Mang, who established his own regime under the title of Hsin. In AD 25 the authority of the Han dynasty was reaffirmed by Liu Hsiu (posthumous name Kuang-wu ti), who reigned as Han emperor until 57/58. Thirteen of his descendants maintained the dynastic succession until 220, when the rule of a single empire was replaced by that of three separate kingdoms. While the whole period from 206 or 202 BC to AD 220 is generally described as that of the Han dynasty, the terms Hsi (Western; also called Former) Han and Tung (Eastern; also called Later) Han are used to denote the two subperiods. During the first period, from 206 BC to AD 25, the capital city was situated at Ch'ang-an, in the west; in the second period, from AD 25 to 220, it lay farther east at Lo-yang.

The four centuries in question may be treated as a single historical period by virtue of dynastic continuity; for, apart from the short interval of 9–25, Imperial authority was unquestionably vested in successive members of the same family. The period, however, was one of considerable changes in Imperial, political, and social development. Organs of government were established, tried, modified, or replaced, and new social distinctions were brought into being. Chinese prestige among other peoples varied with the political stability and military strength of the Han house, and the extent of territory that was subject to the jurisdiction of Han officials varied with the success of Han arms. At the same time the example of the palace, the activities of government, and the growing luxuries of city life gave rise to new standards of cultural and technological achievement.

Ch'in and Han dynasties compared

China's first Imperial dynasty, that of Ch'in, had lasted barely 15 years before its dissolution in the face of rebellion and civil war. By contrast, Han formed the first long-lasting regime that could successfully claim to be the sole authority entitled to wield administrative power. The Han forms of government, however, were derived in the first instance from the Ch'in dynasty; and these, in turn, incorporated a number of features of the government that had been practiced by earlier kingdoms. The Han Empire left as a heritage a practical example of Imperial government and an ideal of dynastic authority to which its successors have always aspired. But the Han period has been credited with more success than is its due; it has been represented as a period of 400 years of effective dynastic rule, punctuated by a short period of usurpation by a pretender to power, and it has been assumed that Imperial unity and effective administration advanced steadily with each decade. In fact, there were only a few short periods marked by dynastic strength, stable government, and intensive administration. Several reigns were characterized by palace intrigue and corrupt influences at court, and on a number of occasions the future of the dynasty was seriously endangered by outbreaks of violence, seizure of political power, or a crisis in the Imperial succession.

### DYNASTIC AUTHORITY AND THE SUCCESSION OF EMPERORS

**Hsi (Western) Han.** Since at least as early as the Shang dynasty the Chinese had been accustomed to acknowledging the temporal and spiritual authority of a single leader and its transmission within a family, at first from brother to brother and later from father to son. Some of the early kings had been military commanders, and they may have organized the corporate work of the community, such as the manufacture of bronze tools and vessels. In addition, they acted as religious leaders, appointing scribes or priests to consult the oracles and thus to assist in making major decisions covering communal activities, such as warfare or hunting expeditions. In succeeding centuries the growing sophistication of Chinese culture was accompanied by demands for more intensive political organization and for more regular administration; as kings came to delegate tasks to more officials, so was their own authority enhanced and the obedience that they commanded the more widely acknowledged. Under the kingdoms of Chou an association was deliberately fostered between the authority of the king and the dispensation exercised over the universe by Heaven, with the result that the kings of Chou and, later, the emperors of Chinese dynasties were regarded as being the Sons of Heaven.

Events immediately preceding the foundation of the Han dynasty

From 403 BC onward seven kingdoms other than Chou constituted the ruling authorities in different parts of China, each of which was led by its own king or duke. In theory, the king of Chou, whose territory was by now greatly reduced, was recognized as possessing superior powers and moral overlordship over the other kingdoms, but practical administration lay in the hands of the seven kings and their professional advisers or in the hands of well-established families. Then in 221 BC, after a long process of expansion and takeover, a radical change occurred in Chinese politics: the kingdom of Ch'in succeeded in eliminating the power of its six rivals and established a single rule that was acknowledged in their territories. According to later Chinese historians, this success was achieved and the Ch'in Empire was thereafter maintained by oppressive methods and the rigorous enforcement of a harsh penal code, but this view was probably coloured by later political prejudices. Whatever the quality of Ch'in Imperial government, the regime scarcely survived the death of the first emperor (210/209 BC). The choice of

his successor was subject to manipulation by statesmen, and local rebellions soon developed into large-scale warfare. Kao-tsu, whose family had not so far figured in Chinese history, emerged as the victor of two principal contestants for power. Anxious to avoid the reputation of having replaced one oppressive regime by another, he and his advisers endeavoured to display their own empire—of Han—as a regime whose political principles were in keeping with a Chinese tradition of liberal and beneficent administration. As yet, however, the concept of a single centralized government that could command universal obedience was still subject to trial. In order to exercise and perpetuate its authority, therefore, Kao-tsu's government perforce adopted the organs of government, and possibly many of the methods, of its discredited predecessor.

The authority of the Han emperors had been won in the first instance by force of arms, and both Kao-tsu and his successors relied on the loyal cooperation of military leaders and on officials who organized the work of civil government. In theory and to a large extent in practice, the emperor remained the single source from whom all powers of government were delegated. It was the Han emperors who appointed men to the senior offices of the central government and in whose name the governors of the commanderies (provinces) collected taxes, recruited men for the labour corps and army, and dispensed justice. And it was the Han emperors who invested some of their kinsmen with powers to rule as kings over certain territories or divested them of such powers in order to consolidate the strength of the central government.

The Imperial succession

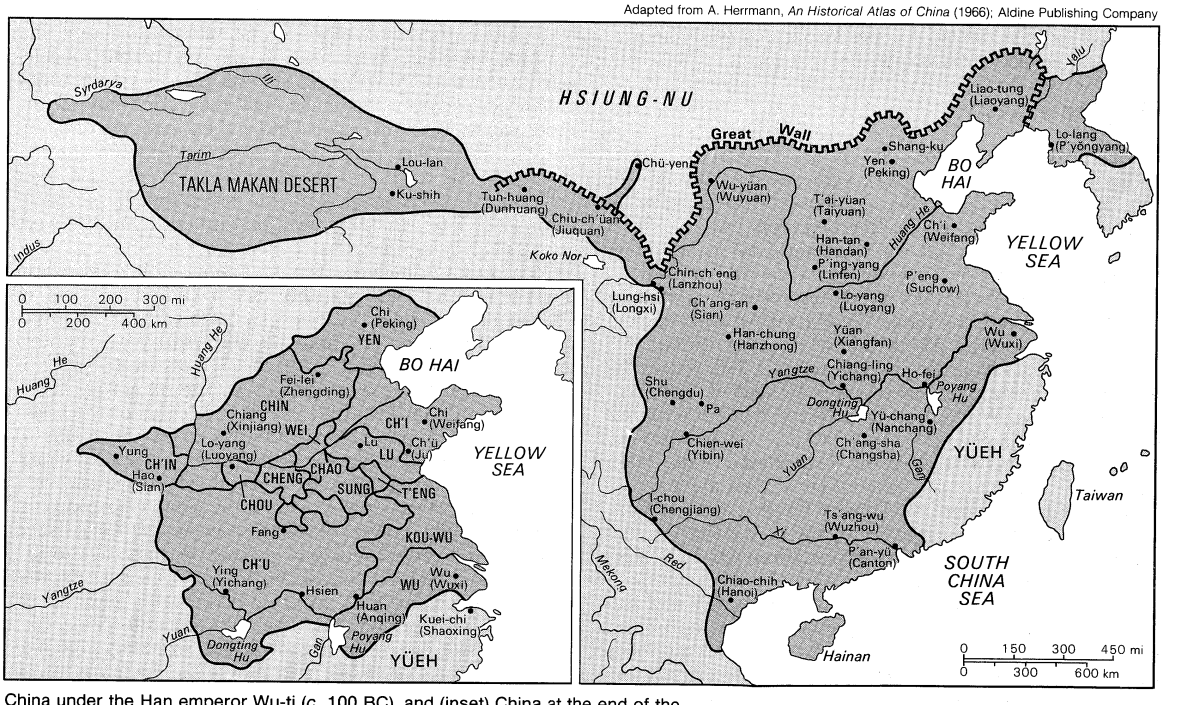
The succession of emperors was hereditary, but it was complicated to a considerable extent by a system of Imperial consorts and the implication of their families in politics. Of the large number of women who were housed in the palace as the emperor's favourites, one was selected for nomination as the empress; and while it was theoretically possible for an emperor to appoint any one of his sons heir apparent, this honour, in practice, usually fell on one of the sons of the empress. Changes could be made in the declared succession, however, by deposing one empress and giving the title to another favourite, and sometimes, when an emperor died without having nominated his heir, it was left to the senior statesmen of the day to arrange for a suitable successor. Whether or not an heir had been named, the succession was often open to question, as pressure could be exerted on an emperor over his choice. Sometimes a young or weak emperor was overawed by the expressed will of his mother or by anxiety to please a newly favoured concubine.

Throughout the Hsi Han and Tung Han periods, the succession and other important political considerations were affected by the members of the Imperial consorts' families. Often the father or brothers of an empress or concubine were appointed to high office in the central government; alternatively, senior statesmen might be able to curry favour with their emperor or consolidate their position at court by presenting a young female relative for the Imperial pleasure. In either situation the succession of emperors might be affected, jealousies would be aroused between the different families concerned, and the actual powers of a newly acceded emperor would be overshadowed by the women in his entourage or their male relatives. Such situations were particularly likely to develop if, as often happened, an emperor was succeeded by an infant son.

The Imperial succession was thus frequently bound up with the political machinations of statesmen, particularly as the court grew more sophisticated and statesmen acquired coteries of clients engaged in factional rivalry. On the death of the first emperor, Kao-tsu (195 BC), the palace came under the domination of his widow. Outliving her son, who had succeeded as emperor under the title of Hui-ti (reigned 195–188), the empress dowager Kao-hou arranged for two infants to succeed consecutively. During this time (188–180/179 BC) she issued Imperial edicts under her own name and by virtue of her own authority as empress dowager. She set a precedent that was to be followed in later dynastic crises—*e.g.*, when the throne was vacant and no heir had been appointed; in such cases, although statesmen or officials would in fact determine how to proceed, their decisions were implemented in the form of edicts promulgated by the senior surviving empress.

Kao-hou appointed a number of members of her own family to highly important positions of state and clearly hoped to substitute her own family for the reigning Liu family. But these plans were frustrated on her death (180/179) by men whose loyalties remained with the founding emperor and his family. Liu Heng, better known as Wen-ti, reigned from 180/179 to 157/156. He soon came to be regarded (with Kao-tsu and Wu-ti) as one of three outstanding emperors of the Hsi Han. He was credited with the ideal behaviour of a monarch reigning according to later Confucian doctrine; *i.e.*, he was supposedly ready to yield place to others, hearken to the advice and remonstrances of his statesmen, and eschew personal

Adapted from A. Herrmann, *An Historical Atlas of China* (1966); Aldine Publishing Company




China under the Han emperor Wu-ti (*c.* 100 BC), and (inset) China at the end of the Ch'un-ch'iu period (*c.* 500 BC).

extravagance. It can be claimed that his reign saw the peaceful consolidation of Imperial power, successful experimentation in operating the organs of government, and the steady growth of China's material resources.

Accomplishments of Wu-ti's reign

The third emperor of the Hsi Han to be singled out for special praise by traditional Chinese historians was Wu-ti (reigned 141/140–87/86 BC), whose reign was the longest of the whole Han period. His reputation as a vigorous and brave ruler derives from the long series of campaigns fought chiefly against the Hsiung-nu (northern nomads) and in Central Asia. But Wu-ti never took a personal part in the fighting. The policy of taking the offensive and extending Chinese influence into unknown territory resulted not from the Emperor's initiative but from the stimulus of a few statesmen, whose decisions were opposed vigorously at the time. Thanks to the same statesmen, Wu-ti's reign saw a more intensive use of manpower and exploitation of natural resources. This depended on more active administration by Han officials. Wu-ti participated personally in the religious cults of state far more actively than his predecessors and some of his successors. And it was during his reign that the state took new steps to promote scholarship and develop the civil service.

From *c.* 90 BC it became apparent that Han military strength had been overtaxed, and a policy of retrenchment was begun in military and economic policies. The last few years of the reign were darkened by a dynastic crisis arising out of jealousies between the Empress and heir apparent on the one hand, and a rival Imperial consort's family on the other. Intense and violent fighting took place in Ch'ang-an in 91, and the two families were almost eliminated. Just before Wu-ti's death a compromise was reached whereby an infant who came from neither family was chosen to succeed. This was Chao-ti (reigned 87/86–74/73). The stewardship of the empire was vested in the hands of a regent, Huo Kuang. This shrewd and circumspect statesman had seen service in government for some two decades, and even after his death his family retained a dominating influence in Chinese politics until 64 BC. Chao-ti had been married to a granddaughter of Huo Kuang; his successor, who was brought to the throne at the invitation of Huo Kuang and other statesmen, proved unfit for his august position and was deposed after a reign of 27 days. Huo Kuang, however, was able to contrive his replacement by a candidate whom he could control or manipulate. This was Hsüan-ti (reigned 74/73–48), who began to take a personal part in government after Huo Kuang's death in 68. The new emperor had a predilection for a practical rather than a scholastic approach to matters of state. While his reign was marked by a more rigorous attention to implementing the laws than had recently been fashionable, his edicts paid marked attention to the ideals of governing a people in their own interests and distributing bounties where they were most needed. The move away from the aggressive policies of Wu-ti's statesmen was even more noticeable during the next reign (Yüan-ti; 48–33/32).

The regency of Huo Kuang

In the reigns of Ch'eng-ti (33/32–7/6), Ai-ti (7/6–1 BC), and P'ing-ti (1 BC–AD 6) the conduct of state affairs and the atmosphere of the court were subject to the weakness or youth of the emperors, the lack of an heir to succeed Ch'eng-ti, and the rivalries between four families of Imperial consorts. It was a time when considerable attention was paid to omens. Changes that were first introduced in the state religious cults in 32 BC were alternately countermanded and reintroduced in the hope of securing material blessings by means of intercession with different spiritual powers. To satisfy the jealousies of a favourite, Ch'eng-ti went so far as to murder two sons born to him by other women. Ai-ti took steps to control the growing monopoly exercised by other families over state affairs. It was alleged at the time that the deaths of both Ch'eng-ti, who had enjoyed robust health, and P'ing-ti, not yet 14 when he died, had been arranged for political reasons.

In the meantime, the Wang family had come to dominate the court. Wang Cheng-chün, who had been the empress of Yüan-ti and mother of Ch'eng-ti, exercised considerable powers not only in her own capacity but also through several of her eight brothers. From 33/32 to 7/6 BC five members of the family were appointed in succession to the most powerful position in the government, and the status of other members was raised by the bestowal of nobilities. The Empress Dowager lived until AD 13, surviving the decline of the family's influence under Ai-ti, who sought to restore a balance at court by honouring the families of other consorts (the Fu and Ting families). Wang Mang, nephew of the empress dowager Wang, restored the family's position during the reign of P'ing-ti. After the latter died and an infant succeeded to the throne, Wang Mang was appointed regent, but in AD 9 he assumed the Imperial position himself, under the dynastic title of Hsin.

The regime of Wang Mang

Insofar as he took Imperial power from the Liu family, Wang Mang's short reign from 9 to 23 may be described as an act of usurpation. His policies were marked by both traditionalism and innovation. In creating new social distinctions, he tried to revert to a system allegedly in operation before the Imperial age, and some of his changes in the structure of government were similarly related to precedents of the dim past. He appealed to the poorer classes by instituting measures of relief, but his attempts to eliminate private landholding and abolish private slaveholding antagonized the more wealthy members of society. Experiments in new types of coinage and in controlling economic transactions failed to achieve their purpose of increasing the resources of state, which were depleted by enormously costly preparations for campaigns against the Hsiung-nu. The last years of his reign were dislocated by the rise of dissident bands in a number of provinces; several leaders declared themselves emperor in different regions, and in the course of the fighting Ch'ang-an was entered and damaged. Later it was captured by the Red Eyebrows, one of the most active of the robber bands; and Wang Mang was killed in a scene of violence played out within the palace buildings.

**Tung (Eastern) Han.** The Han house was restored by Liu Hsiu, better known as Kuang-wu ti, who reigned from AD 25 to 57/58. His claim had been contested by another member of the Liu house, Liu Hsüan—better known as Liu Keng-shih—who had been actually enthroned for two years, until his death in the course of turbulent civil fighting. Ch'ang-an had been virtually destroyed by warfare, and Kuang-wu ti established his capital at Lo-yang.

The new emperor completed the defeat of rival aspirants to the throne in 36. As had occurred in Hsi Han, dynastic establishment was followed by a period of internal consolidation rather than expansion. Kuang-wu ti resumed the structure of government of the Hsi Han emperors, together with the earlier coinage and system of taxation. The palace once more promoted the cause of scholarship. Eunuchs had come to the fore in the Han palace during Yüan-ti's reign, and several had succeeded in reaching powerful positions. Kuang-wu ti's policy was to rid the government of such influences, together with that of the families of Imperial consorts. Under Ming-ti (57/58–75/76) and Chang-ti (75/76–88), China was once more strong enough to adopt a positive foreign policy and to set Chinese armies on the march against the Hsiung-nu. To prevent the incursions of the latter, and possibly to encourage the growth of trade, Han influence was again brought to bear in Central Asia. Chinese prestige reached its zenith around 90 and fell markedly after 125.

Beginnings of Han decline

Dynastic decline can be dated from the reign of Ho-ti (88–105/106), when the court once more came under the influence of consorts' families and eunuchs. The succession of emperors became a matter of dexterous manipulation designed to preserve the advantages of interested parties. The weakness of the throne can be judged from the fact that, of the 12 emperors of Tung Han, no less than eight took the throne as boys aged between 100 days and 15 years. There was an increasing tendency for the growth of factions whose members, like the families of Imperial consorts and like the eunuchs, might choose to place their own interests above those of the state.

During the last 50 years of Tung Han, North China became subject to invasion from different sides; and, as was observed by several philosopher-statesmen, the administration became corrupt and ineffective. Powerful regional officials were able to establish themselves almost indepen-

dently of the central government. Rivalry between consorts' families and eunuchs led to a massacre of the latter in 189, and the rebel bands that arose included the Yellow Turbans, who were fired by beliefs in supernatural influences and led by inspired demagogues. Soldiers of fortune and contestants for power were putting troops in the field in their attempts to establish themselves as emperors of a single united China. By 207 Ts'ao Ts'ao had gained control over the north; and had he not been defeated by Sun Ch'üan at the battle of the Red Cliff, which later became famous in Chinese literature, he might well have succeeded in establishing a single dynastic rule. Other participants in the fighting included Tung Cho, Liu Pei, and Chu-ko Liang. The situation was resolved in 220 when Ts'ao P'ei, son of Ts'ao Ts'ao, accepted an instrument of abdication from Hsien-ti, last of the Han emperors (acceded 189). Ts'ao P'ei duly became emperor of a dynasty styled Wei, whose territories stretched over the northern part of China and whose capital was at Lo-yang. A year later, in 221, Liu Pei was declared emperor of the Shu-Han dynasty, thereby maintaining the fiction that as a member of the Liu family he was continuing its rule of the Han dynasty, albeit in the restricted regions of Shu in the southwest (capital at Ch'eng-tu). In the southeast there was formed the third of the San-kuo (Three Kingdoms), as the period from 220 to 280 has come to be described. This was the kingdom of Wu, with its capital at Chien-yeh, under the initial dispensation of Sun Ch'üan.

## THE ADMINISTRATION OF THE HAN EMPIRE

**The structure of government.** One of the main contributions of the Han dynasty to the future of Imperial China lay in the development of the civil service and the structure of central and provincial government. The evolutionary changes that subsequently transformed Han polity beyond recognition were not directed at altering the underlying principles of government but at applying them expediently to the changing dynastic, political, social, and economic conditions of later centuries. One of the problems faced by Han governments was the recruitment of able and honest men to staff the civil service of an empire; these men eventually became known in the West as mandarins. Despite the recent reform of the script, which facilitated the drafting of documents, considerable training was still needed before sufficient competence could be attained. Much of the training occurred in local-level bureaus, where aspirants for Imperial appointments served the equivalent of apprenticeships. Meritorious young men advanced from clerical positions to head various local bureaus. Having proved themselves in these positions, they were then eligible for recommendation or sponsorship, the standard means of recruitment to the civil service. Officials were invited to present candidates who possessed suitable qualities of intelligence and integrity, usually established in their service in local bureaus, and at certain times provincial units were ordered to send a quota of men to the capital at regular intervals. At times candidates were required to submit answers on questions of policy or administration. They might then be kept at the palace to act as advisers in attendance, or they might be given appointments in the central government or in the provinces, depending upon their success. But at this time there was no regular system of examination and appointment such as was evolved in the time of the Sui and T'ang dynasties.

Recruitment of civil servants

The recruitment system was important for two reasons directly related to the nature and development of Han society. First, the apprenticeship system assured that entry into the Imperial bureaucracy was based on administrative merit. Thus, men of little wealth could enter clerical positions and support themselves while preparing for higher-level careers. (This recruitment system differed strikingly from the later examination system that often required years of study in which to master the Confucian Classics and to develop writing skills.) Second, powerful families, increasingly in the Tung Han period, were able to dominate the clerical and other positions in the local bureaus, thereby limiting to those powerful families the candidates for Imperial bureaucratic service. Control of local positions in turn strengthened the powerful families by allowing them to manipulate tax and census registers. Such families created the social milieu from which the aristocratic families of the post-Han period were to emerge.

There was a total of 12 grades in the Han civil service, ranging from that of clerk to the most senior minister of state. No division in principle existed between men serving in the central offices or the provincial units. Promotion could be achieved from one grade of the service to the next, and, in theory, a man could rise from the humblest to the highest post. In theory and partly in practice, the structure of Han government was marked by an adherence to regular hierarchies of authority, by the division of specialist responsibilities, and by a duplication of certain functions. By these means it was hoped that excessive monopoly of power by individual officials would be avoided. The uppermost stratum of officials or statesmen comprised the chancellor, the Imperial counselor, and, sometimes, the commander in chief. These men acted as the emperor's highest advisers and retained final control over the activities of government. Responsibility was shared with nine ministers of state who cared for matters such as religious cults, security of the palace, adjudication in criminal cases, diplomatic dealings with foreign leaders, and the collection and distribution of revenue. Each minister of state was supported by a department staffed by directors and subordinates. There were a few other major agencies; these ranked slightly below the nine ministries and were responsible for specialist tasks. Functions were duplicated so as to check the growth of power. Occasionally, for example, two chancellors were appointed concurrently. Similarly, financial matters were controlled by two permanent ministries: the Department of Agriculture and Revenue and the Privy Treasury.

The foregoing structure of regular organs of government was known as the Outer Court. With the passage of time it became balanced by the growth of a secondary seat of power known as the Inner Court. This grew up from members of the secretariat and had started as a subordinate agency in the Privy Treasury. The secretariat officials had acquired direct access to the emperor and could thus circumvent the more formal approaches required by protocol of other officials. The secretariat rose to prominence during the latter part of the 1st century BC and was at times staffed by eunuchs. Its members were sometimes distinguished by the bestowal of privileged titles that conveyed a mark of Imperial favour without specific administrative responsibility. The highest of these titles was that of supreme commander, and, when this title was accompanied by the right or the Imperial instruction to assume leadership of the secretariat, the powers of the incumbent outweighed those of the highest ministers of the Outer Court. An official thus named could effectively control decisions of state, to the discomfiture of senior officials such as the chancellor. It was in this capacity that Wang Mang and his four predecessors had been able to assert their power without fear of check.

Provincial government

At the outset of the Han dynasty very large areas were entrusted as kingdoms to the emperor's kinsmen while the central government administered the interior provinces as commanderies. But by *c.* 100 BC the Imperial government had deprived the kingdoms of their strength, and most of their lands had been incorporated as commanderies under the central government. Although the kingdoms survived in a much reduced form until the end of the period, their administration came to differ less and less from that of the commanderies, which formed the regular provincial units. Each commandery was controlled by two senior officials, the governor and the commandant, who were appointed by the central government. Commanderies could be established at will: by dividing larger into smaller units, by taking over the lands of the kings, or by establishing organs of government in regions only recently penetrated by Chinese officials. Provincial government was not necessarily pervasive throughout the lands where commandery offices existed, but there was a steady advance in provincial government during the Han period. During Kao-tsu's reign 16 commanderies existed, but by the end of the Hsi Han there were 83 commanderies and 20 kingdoms.

Each of the commanderies consisted of some 10 or 20

prefectures, the size of which corresponded to that of English counties. The prefect's headquarters were situated in a walled town, from which his administration was extended and his officials were sent to collect taxes, settle disputes, or recruit able-bodied men for service. The prefectures were themselves subdivided into districts. The commanderies included a number of nobilities, the holders of which enjoyed a noble title and income from the taxes collected in them by central government officials. The nobles exercised no administrative, judicial, or other power over their nobilities. The number of nobilities varied considerably, sometimes totaling several hundred. The system was used as a political instrument for reducing the power of the kings, rewarding military officers and civil officials, and treating surrendered enemy leaders. Special arrangements were instituted for provincial government at the periphery of the empire. Agencies of a specialist nature were set up both there and in the provinces of the interior, with responsibilities for such matters as supervision of the salt and iron industries, manufacture of textiles, fruit growing, and sponsored agriculture, as well as control of passage in and out of the frontier.

From 106 BC the government tried to supervise the work of provincial officials more directly. A total of 14 regional inspectors was appointed, with orders to visit the commanderies and kingdoms of a specified area and to report to the central government on the efficiency of officials, the degree of oppression or corruption, and the state of popular affection or disaffection. Although the arrangement was not yet tantamount to the creation of a limited number (about 20) of large provinces, such as came about from about the 13th century, it may have facilitated the establishment of separatist provincial regimes at times of dynastic decline.

Command of the armed forces

The command of the armed forces was also arranged so as to avoid giving excessive powers to a single individual. General officers were usually appointed in pairs, and, in times of emergency or when a campaign was being planned with a defined objective, officers were appointed for a specific task; when their mission was fulfilled, their commands were brought to a close. At a lower level there existed a complement of colonels whose duties were defined so as to cover smaller scale activities. In addition, the governors and commandants of the commanderies were sometimes ordered to lead forces. The commandants were also responsible for training conscript soldiers and setting them to maintain internal discipline and to man the static lines of defense in the north and northwest.

The Han armies drew their recruits from conscripts, volunteers, and convicts. Conscripts, who formed the majority, were obliged to serve for two years, either under training or on active service. This duty devolved on all able-bodied males other than those who had acquired privileges of rank or those who could pay for substitutes. The latter practice was probably rare. In addition, men were liable for recall to the armed forces in times of emergency. Volunteers were the sons of privileged families and probably served as cavalrymen, and convicts were sometimes drafted to work out their terms of sentence in the army. There is ample evidence to show that Han commanders used to draw on Central Asian tribesmen as recruits, and the tribesmen were particularly valuable as skilled cavalrymen. A number of foreigners also served with distinction as officers. While little is known of the organization of armies on campaign, garrison forces were divided into separate commands consisting of perhaps four companies. Each company had a strength of some 40 or 50 sections, each of which comprised one officer and up to five men.

**The practice of government.** As the final arbiter of power, the emperor—and at times the empress dowager—issued edicts declaring the Imperial will. Such instructions often took the form of repeating officials' proposals with a note of approval. Some edicts were couched as comments on the current situation and called in general terms for an improvement in the quality of government or for more vigorous attempts to achieve a just administration. The emperor also issued formal deeds of investiture to kings or noblemen and letters of appointment for senior officials. Edicts were circulated to the relevant authorities for action, together with books of other regulations such as the statutes and ordinances, laying down entitlements for services rendered to the state and penalties for infringing its prohibitions. Officials could suggest methods of government by submitting written memorials, and there were occasions when an emperor called a conference of senior statesmen and asked their views on topical problems.

The Han governments regularly issued calendars to enable the court to follow a cosmically correct ritual schedule and officials to maintain their records correctly. Regular means of transport were kept for the use of officials traveling on business and for the conveyance of official mail from one office to another. Provincial and local officials were responsible for two regular counts without which government could not proceed: the census of the population and the register of the land and its production. Returns, which were submitted for the number of households and individuals and for acreage under the plow, eventually found their way to the capital. One count that has been preserved records the existence of some 12,233,000 households and 59,595,000 individuals in AD 2. Two other main forms of revenue collection were the land tax and the poll tax. The land tax was levied in kind at a 30th (sometimes a 15th) part of the produce, the assessment depending partly on the quality of the land. Poll tax was usually paid in cash and varied with the age and sex of the members of the household. Other taxes were levied in respect to wealth and by means of property assessments.

Use of the census

In addition to service in the army, able-bodied males were liable to one month's service annually in the state labour corps; tasks included building palaces and Imperial mausoleums, transporting staple goods such as grain and hemp, and constructing roads and bridges. Sometimes conscript labour was used to repair breaches in riverbanks or dikes, and men were sent to work in the salt and iron industries after these were taken over by the state.

State monopolies

The establishment of state monopolies for salt and iron was one of several measures taken in Wu-ti's reign to bring China's resources under the control of the government. Agencies were set up *c.* 117 BC to supervise mining, manufacturing, and distribution and to raise revenue in the process. The measure was criticized on the grounds both of principle and of expedient and was withdrawn for three years from 44 BC, and by the mid-1st century AD the industries had reverted in practice to private hands. Final measures to standardize the coinage and to limit minting to state agencies were taken in 112 BC; and, with the exception of Wang Mang's experiments, the copper coin of a single denomination, minted from Wu-ti's reign onward, remained the standard medium of exchange. Little is known of the work of other agencies established in Hsi Han to stabilize the prices of staple commodities and to regulate their transport. Such measures had been the answer of Wu-ti's government to the problem of moving goods from an area of surplus to one of shortage.

The government ordered migrations of the population for several reasons. At times, such a migration was intended to populate an area artificially—the city of Hsien-yang during the Ch'in Empire, for example, and the state-sponsored farms of the borderlands. Alternatively, if the defense of the periphery was impractical, the population was sometimes moved away from danger, and distressed folk were moved to areas where they could find a more prosperous way of life.

From about 100 BC it was evident to some statesmen that great disparities of wealth existed and that this was most noticeable in respect of landownership. Some philosophers looked back nostalgically to an ideal state in which land was said to have been allotted and held on a basis of equality, thereby eliminating the wide differences between rich and poor. It was only in Wang Mang's time that an attempt was made to abolish private landownership and private slaveholding. But the attempt failed because of powerful economic and social opposition, and the accumulation of land continued during Tung Han. In the last half century or so of the dynasty, country estates acquired retainers and armed defenders, almost independently of the writ of government. The great families thus came to exercise more power than appointed officials of state.

Han law

The Han, like the Ch'in, government ruled by dispensing rewards for service and exacting punishment for disobedience and crime. Rewards consisted of exemptions from tax; bounties of gold, meat, spirits, or silk; amnesties for criminals; and orders of honour. The latter were bestowed either individually or to groups. There was a scale of rank of 20 degrees, and, with the receipt of several of these awards cumulatively, one could rise to the eighth place in the scale. The more senior orders were given for specified acts of valour, charity, or good administration, usually to officials, and the highest of the orders was that of the nobility. In addition to conferring social status, the orders carried with them legal privileges and freedom from some tax and service obligations.

In theory, the laws of Han were binding on all members of the population, and some incidents testify to the punishment of the highest in the land. But some privileged persons were able to secure mitigation of sentences. Nobles, for example, could ransom themselves from most punishments by forfeiting their nobilities. Han laws specified a variety of crimes, including those of a social nature such as murder or theft, those that infringed the Imperial majesty, and offenses that were classed as gross immorality. There was a regular procedure for impeachment and trial, and some difficult cases could be referred to the emperor for a final decision. The punishments to which criminals were sentenced included exile, hard labour, flogging, castration, and death. In the most heinous cases the death sentence was carried out publicly, but senior officials and members of the Imperial family were usually allowed to avoid such a scene by committing suicide. After the death penalty a criminal's goods, including members of his family, were confiscated by the state. Such persons then became slaves of the state and were employed on menial or domestic tasks in government offices. Government slaves were sometimes given as rewards to meritorious officials.

## RELATIONS WITH OTHER PEOPLES

Simultaneously with the rise of the Ch'in and Han empires, some of the nomadic peoples of Central Asia, known as the Hsiung-nu, succeeded in achieving a measure of unity under a single leader. As a result, while the Chinese were consolidating their government, the lands lying to the north of the empire, and the northern provinces themselves, became subject to incursion by Hsiung-nu horsemen. One of the achievements of the Ch'in dynasty had been the unification of the several lines of defense into a single system of fortification, the Great Wall. By keeping that wall, or line of earthworks, manned, the Ch'in dynasty had been free of invasion. With the fall of Ch'in and China's subsequent weakness, the wall fell into a state of disrepair and lacked a garrison. Until about 135 BC Han governments were obliged to seek peaceful relations with the Hsiung-nu at the price of gold, silk, and even the hand of a Chinese princess. But with the initiation of strong policies by Wu-ti's governments, China took the offensive in an attempt to throw back the Hsiung-nu to Central Asia and to free the northern provinces from the threat of invasion and violence. By 119 BC campaigns fought to the north of Chinese territory had attained this objective, and after a short interval it was possible to send Han armies to advance in the northeast (modern Korea), the south (modern Vietnam), and the southwest. As a result of the campaigns fought from 135 BC onward, 18 additional commanderies were founded, and organs of Han provincial government were installed as outposts among peoples who were unassimilated to a Chinese way of life.

Relations with the Hsiung-nu

Chinese government was by no means universally accepted in these outlying regions. But despite large losses and expenditures incurred in fighting the Hsiung-nu, the Chinese were able to mount expeditions into Central Asia from *c.* 112 BC. The defensive walls were repaired and remanned, and by *c.* 100 BC they were extended to the northwest as far as Tun-huang. Chinese travelers, whether diplomats or merchants, were thus protected as far as the Takla Makan Desert. It was at this juncture that trade routes skirting the desert were established and came to be known collectively as the Silk Road.

The success of Chinese arms in these remote areas was short-lived. Long lines of communication made it impossible to set up garrisons or colonies in the forbidding country to the west of Tun-huang. Diplomatic moves were made to implant Chinese prestige more firmly among the communities that were situated around the Takla Makan Desert and controlled the oases, for it was necessary for the Chinese to win those peoples' support, thus denying it to the Hsiung-nu. In a few cases the Chinese resorted to violence or plots to remove a leader and to replace him with a candidate known to favour the Han cause. A more usual procedure was to marry one of the alien leaders to a Chinese princess, with the intention that he should in time be succeeded by an heir who was half-Chinese. These endeavours and the military ventures met with partial success. While the Chinese position in Central Asia was subject to question, relations with the Hsiung-nu leaders varied. The visit of a Hsiung-nu leader to Ch'ang-an in 51 BC was hailed as a mark of Chinese success, but the ensuing decades were not free from fighting. Chinese prestige declined toward the end of the Hsi Han and recovered only during the reigns of Ming-ti and Chang-ti, when the Han government was once more strong enough to take the field. Pan Ch'ao's campaigns in Central Asia (from AD 94) reestablished the Chinese position, but again the full strength of Chinese prestige lasted for only a few decades. During the Tung Han, China suffered invasion from the northeast as well as from the north. The settlement of Hsiung-nu tribesmen south of the wall was a disruptive factor in the 2nd and 3rd centuries, to the detriment of Imperial unity.

The Han expansion into Central Asia has been represented by the Chinese as a defensive measure designed to weaken the Hsiung-nu and to free China from invasion. Allowance must also be made for commercial motives. Some of Wu-ti's statesmen were well aware of the advantages of exporting China's surplus products in return for animals and animal products from Central Asia, and there is evidence that Chinese silk was exported at this time. No attempt can be made to estimate the volume of trade, and, as the transactions were conducted through Parthian middlemen, no direct contact was made by this means between Han China and the world of Rome and the Mediterranean. China's export trade was sponsored by the government and not entrusted to private merchants.

External trade

The Great Wall formed a boundary separating the Chinese provinces from the outside world. Traffic was controlled at points of access, not only to check incoming travelers to China but also to prevent the escape of criminals or deserters. At the same time, a ban was imposed on the export of certain goods such as iron manufactures and weapons of war. The wall also formed a protected causeway for travelers to the west. Watch stations were erected in sight of each other to signal the approach of the enemy, and the garrison troops were highly trained and disciplined. Meticulous records were kept to show how government stores were expended and rations issued; routine signals were relayed along the line and daily patrols were sent out to reconnoitre.

As a result of the campaigns and of diplomatic activity, China's immediate contacts with other peoples grew more brisk. Many of the Hsiung-nu and other neighbouring leaders who had surrendered to Han arms were given nobilities and settled in the interior of the empire. Chang Ch'ien had been the pioneer who had set out *c.* 130 BC to explore the routes into Central Asia and North China, and, as a result of his report and observations, Han advances were concentrated in the northwest. In AD 97 Chinese envoys were frustrated in an attempt to visit the western part of the world, but a mission from Rome reached China by ship in 166. The first record of official visitors arriving at the Han court from Japan is for the year AD 57.

## CULTURAL DEVELOPMENTS

The Han emperors and governments posed as having a temporal dispensation that had received the blessing of Heaven together with its instructions to spread the benefits of a cultured life as widely as possible. By a cultured life the Chinese had in mind a clear distinction between their own settled agriculture and the delights of the cities, as

opposed to the rough and hardy life spent in the saddle by the nomads of Central Asia. The growth of Han government both depended upon and encouraged the development of literary accomplishment, scholastic competence, religious activity, scientific discovery, and technological achievement.

The discovery of paper

Han administration required a proliferation of documents. Official returns were sometimes kept in duplicate, and each agency kept running files to record its business. Following the reform of the script that had evolved before the Han period, there developed a new style of writing suited to the compilation of official documents. These were written mostly on bulky and fragile wooden strips; silk was also used as a medium for writing. A major development in world history occurred in China in AD 105 when officials reported to the throne the manufacture of a new substance. Although archaeological evidence indicates the existence of paper for more than a century before this incident, the earlier materials were not completely superseded until some three or four centuries later. In the meantime, the demands of a growing civilization had increased the written vocabulary of the Chinese. The first Chinese dictionary, compiled *c.* AD 100, included more than 9,000 characters, with explanations of their meanings and the variant forms used in writing.

In an attempt to break with earlier tradition, the Ch'in government had taken certain steps to proscribe literature and learning. Han governments stressed their desire to promote these causes as part of their mission. In particular, they displayed a veneration for works with which Confucius had been associated, either as a collector of texts or as an editor. Beginning during the reign of Wen-ti, orders were given to search for books lost during the previous dynasty. Knowledge of texts such as the *Shih Ching* ("Classic of Poetry"), the *Shu Ching* ("Classic of History"), the *I Ching* ("Classic of Changes"), and the *Ch'un-ch'iu* ("Spring and Autumn") annals became a necessary accomplishment for officials and candidates for the civil service. To support an argument laid before the throne, statesmen would find a relevant quotation from

Rise of Classical education

these works; already in the 1st century BC the tradition was being formed whereby the civil service of Imperial China was nurtured on a Classical education. On two occasions (51 BC and AD 79) the government ordered official discussions about the interpretation of texts and the validity of differing versions; and in AD 175 a project was completed for inscribing an approved version on stone tablets, so as to allay scholastic doubts in the future. In the meantime, and still before the invention of paper, a collection of literary texts had been made for the Imperial library. The catalog of this collection, which dates from the start of the Christian Era, was prepared after comparing different copies and eliminating duplicates. The list of titles has been preserved and constitutes China's first bibliographical list. The works are classified according to subject, but many have been lost. The importance of these measures lies both in their intrinsic achievement and in the example they set for subsequent dynasties.

The prose style of Han writers was later taken as a model of simplicity, and, following the literary embellishments and artificialities of the 5th and 6th centuries, deliberate attempts were made to revert to its natural elegance. Examples of this direct prose may be seen in the Imperial edicts, in the memorials ascribed to statesmen, and, above all, in the text of the standard histories themselves, in which such documents of state were incorporated. The compilation of the standard histories was a private undertaking in Han times, but it already received Imperial patronage and assistance. History was written partly to justify the authority and conduct of the contemporary regime and partly as a matter of pride in Chinese achievement. Further examples of prose writing are the descriptions of protocol for the court. One of the earliest acts of the Han government (*c.* 200 BC) had been to order the formulation of such modes of behaviour as a means of enhancing the dignity of the throne, and one of the latest compilations (*c.* AD 175) that still survives is a list of such prescriptions, drawn up at a time when the dynasty was manifestly losing its majesty and natural authority. Some of the emperors were themselves composers of versified prose; their efforts have also been preserved in the standard histories.

Religious practices

The emperor was charged with the solemn duty of securing the blessings of spiritual powers for mankind. One of the nine ministries of state existed to assist in this work of mediation, but from the time of Wu-ti onward the emperor himself began to play a more active part in worship and sacrifice. The cults were initially addressed to the Five Elements—fire, water, earth, wood, and metal; to the Supreme Unity; and to the Lord of the Soil. In 31 BC these cults were replaced by sacrifices dedicated to Heaven and Earth. The sites of worship were transferred to the southern and northern outskirts of Ch'ang-an, and a new series of altars and shrines was inaugurated. On occasion the Han emperor paid his respects to supreme powers and reported on the state of the dynasty at the summit of Mt. T'ai. Wu-ti's desire for immortality or for the quickening of his deceased favourites led him to patronize a number of intermediaries who claimed to possess the secret of making contact with the world of the immortals. From such beliefs, and from a fear of the malevolent influences that the unappeased souls of the dead could wreak on mankind, a few philosophers, such as Wang Ch'ung (AD 27–*c.* 100), reacted by propounding an ordered and rational explanation of the universe. But their skepticism received little support. Sometime during the 1st century AD Buddhism had reached China, propagated in all probability by travelers who had taken the Silk Road from north India. The establishment of Buddhist foundations in China and the first official patronage of the faith followed shortly. From the 2nd century AD there arose a variety of beliefs, practices, and disciplines from which alchemy and scientific experiment were to spring and which were to give rise to Taoist religion.

Most of the cultural attainments of the Han period derived from the encouragement of the palace and the needs of officials. A textbook of mathematical problems was probably compiled to assist officials in work such as land assessment; fragments of a medical casebook were concerned with the care of troops and horses serving on the northwestern frontier. Water clocks and sundials were used to enable officials to complete their work on schedule. The palace demanded the services of artists and craftsmen to decorate Imperial buildings with paintings and sculptures and to design and execute jades, gold and silver wares, and lacquer bowls for use at the Imperial table. Intricate patterns in multicoloured silks were woven on looms in the Imperial workshops. On a more mundane level, technology served the cause of practical government. The state's ironwork factories produced precision-made instruments and weapons of war, and the state's agencies for the salt industry supervised the recovery of brine from deep shafts cut in the rocks of west China. Water engineers planned the construction of dikes to divert the flow of excess waters and the excavation of canals to serve the needs of transport or irrigation; and in many parts of the countryside there was seen a sight that was to remain typical of the Chinese landscape up to the 20th century—a team of two or three peasants sitting astride a beam and pedaling the lugs of the "dragon's backbone" so as to lever water from the sluggish channels below to the upper levels of the cultivated land. (J.L.D./Ed.)

## The Six Dynasties

### POLITICAL DEVELOPMENTS

**The division of China.** *San-kuo (Three Kingdoms; AD 220–280).* By the end of the 2nd century AD the Han Empire had virtually ceased to exist. The repression of the Taoist rebellions of the Yellow Turbans and related sects marked the beginning of a period of unbridled warlordism and political chaos, from which three independent centres of political power emerged. In the north all authority had passed into the hands of the generalissimo and "protector of the dynasty," Ts'ao Ts'ao; in AD 220 the last puppet emperor of the Han officially ceded the throne to Ts'ao Ts'ao's son, who thereby became the legitimate heir of the empire and the first ruler of the Wei dynasty. Soon afterward, two competing military leaders proclaimed them-

The Wei, Shu-Han, and Wu dynasties

selves emperor, one in the far interior (Shu-Han dynasty, in the present-day Szechwan Province) and one in the south, behind the formidable barrier of the Yangtze River (the empire of Wu, with its capital at Chien-yeh, present-day Nanking). The short and turbulent period of these "Three Kingdoms" (San-kuo), filled with bloody warfare and diplomatic intrigue, has ever since been glorified in Chinese historical fiction as an age of chivalry and individual heroism.

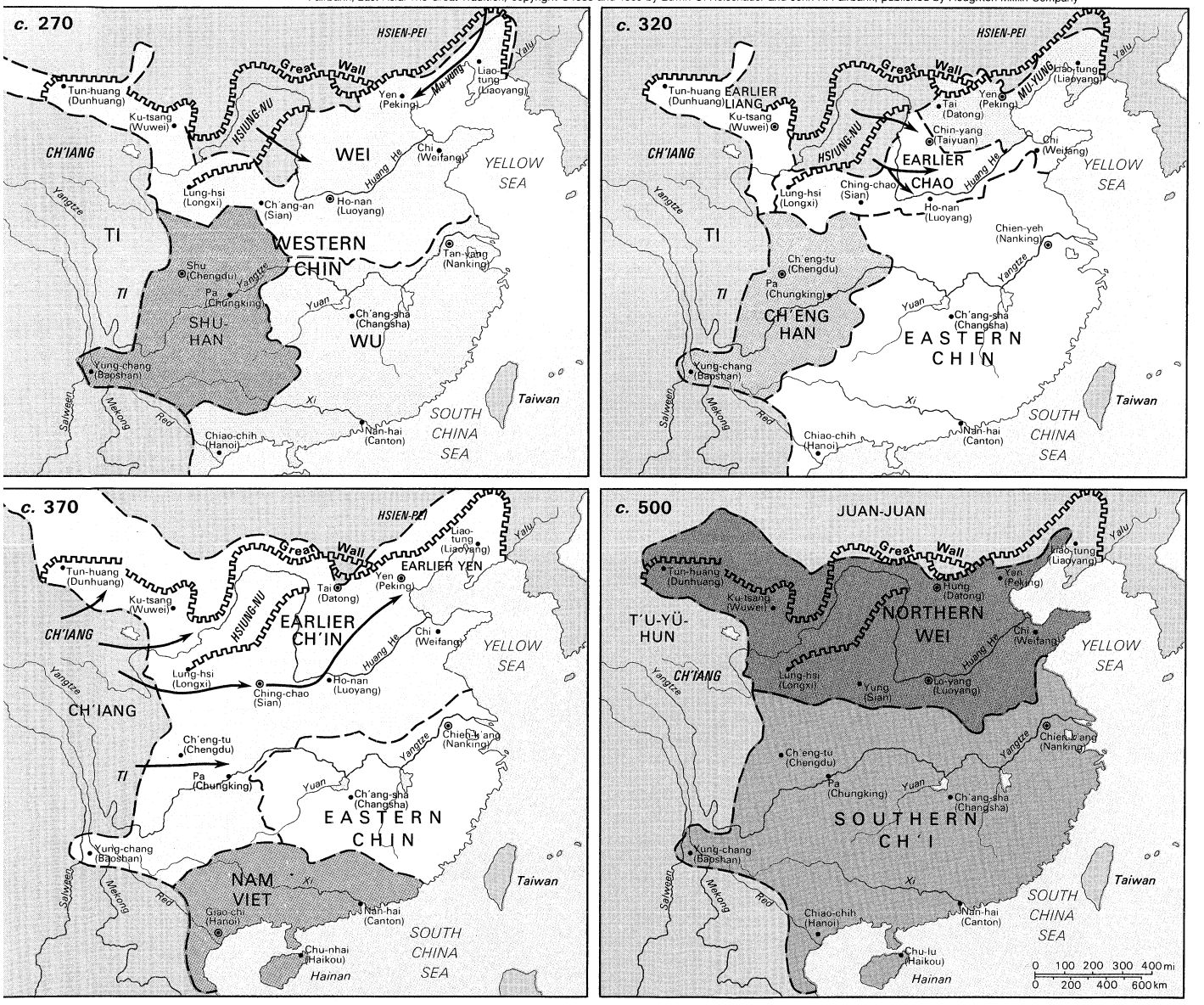
In fact, even Wei, the strongest of the three, hardly represented any real political power. The great socioeconomic changes that had started in the Tung (Eastern) Han period had transformed the structure of society to such an extent that all attempts to reestablish the centralized bureaucratic state—the ideal of the Ch'in and Han dynasties—were doomed to failure. While central authority declined, the great families—aristocratic clans of great landowners—survived the decades of civil war on their fortified estates under the protection of their private armies of serfs and clients and even increased their power. These conditions were to remain characteristic of medieval China. The Han system of recruiting officials on the basis of talent was replaced by a network of personal relations and patronage. The hierarchy of state officials and government institutions was never abolished, but it became monopolized by a few aristocratic clans, who filled the highest offices with their own members and the minor posts with their clients.

Wei succeeded in conquering Shu-Han in 263/264, but two years later a general of the dominant Ssu-ma clan overthrew the house of Wei (265/266) and founded the Hsi (Western) Chin dynasty (265). Wu could maintain itself until 280, when it was overrun by the Chin armies.

The role of Wu was extremely important: it marks the beginning of the progressive Sinicization of the region south of the Yangtze River, which before that time had been a frontier area, inhabited mainly by primitive tribes. The rise of Chien-k'ang (modern Nanking) as a great administrative and cultural centre on the lower Yangtze paved the way for future developments: after the north was lost to barbarian invaders (311), it was to become the capital of Chinese successor states and an important locus of Chinese culture for more than 250 years.

*The Hsi (Western) Chin (AD 265–316/317).* This was a period of relative order and prosperity, a short interlude between the turbulent age of the San-kuo and the devastating barbarian invasions. The empire had been nominally reunited (AD 280), and for a short time the central government attempted important fiscal and political reforms, mainly intended to curb the great families, which threatened the ruler's authority. Contacts with the oasis kingdoms of Central Asia and the Indianized states of the far south (Funan and Champa) were resumed, and in 285

From (all but bottom right) M. Penkala, *A Correlated History of the Far East*, maps by Edward Penkala, F.R.G.S.; (bottom right) E. Reischauer and J. Fairbank, *East Asia: The Great Tradition*, copyright ©1958 and 1960 by Edwin O. Reischauer and John K. Fairbank, published by Houghton Mifflin Company




China in the Six Dynasties period.

the Chin court even sent an envoy to distant Fergana in Central Asia to confer the title of king on its ruler—a grand Imperial gesture reminiscent of the great days of Han. But this ghost of the Han Empire disappeared almost as soon as it had been evoked. Within two decades the Chin disintegrated through the struggles of rival clans. There followed an internecine war between the various Ssu-ma princes, collapse of the central government, decentralized military control of the provinces, famine, large-scale banditry, and messianic peasant movements.

**The era of "barbarian" invasions and rule.** For the first time, the power vacuum was filled by non-Chinese forces. In 304 a Sinicized Hsiung-nu chieftain, Liu Yüan, assumed the title of king of Han and started the conquest of northern China. Operating from bases in western and southern Shansi, the Hsiung-nu armies, supported by local Chinese rebels, conquered the ancient homeland of Chinese civilization; the fall and destruction of the two capitals, Lo-yang (311) and Ch'ang-an (316), ended Chinese dynastic rule in the north for centuries. Although in the far northeast, in present-day Kansu, and in the inaccessible interior (Szechwan), Chinese local kingdoms did occasionally succeed in maintaining themselves for some time, the whole North China Plain itself became the scene of a bewildering variety of barbarian states, collectively known in Chinese historiography as the Shih-liu kuo (Sixteen Kingdoms).

The "barbarian" states of the north and the refugee states of the south

*The Tung (Eastern) Chin (317–420) and later dynasties in the south (420–589).* During the whole medieval period it was the lower Yangtze region, the former territory of Wu, that remained the stronghold of a series of "legitimate" Chinese dynasties with Chien-k'ang as their capital. In 317 a member of the Chin Imperial family had set up a refugee regime at Chien-k'ang, consisting mainly of members of the exiled northern aristocracy. From the beginning, the Chin court was completely at the mercy of the "great families." Government in the Chinese south became a kind of oligarchy exercised by ever-changing groups and juntas of aristocratic clans. The so-called Six Dynasties (actually five: Tung Chin, 317–420; Liu-Sung, 420–479; Nan (Southern) Ch'i, 479–502; Nan Liang, 502–557; and Nan Ch'en, 557–589; the earlier kingdom of Wu, 222–280, is counted as the sixth) were politically and militarily weak and constantly plagued by internal feuds and revolts. Their annihilation (in 589) was postponed only by the internal division of the north and by the protection afforded by the Yangtze. To the very end, their opposition to the north remained alive, but occasional attempts to reconquer the ancient homeland were doomed to failure. The final reunification of China was to start from the northern plains, not from Chien-k'ang.

Although politically insecure, these dynasties were characterized by cultural brilliance: in literature, art, philosophy, and religion, they constituted one of the most creative periods in Chinese history. They reached their highest flowering under the long and relatively stable reign of the great protector of Buddhism, Wu-ti, the first emperor of the Nan Liang dynasty (reigned 502–549).

*The Shih-liu kuo (Sixteen Kingdoms) in the north (303–439).* The term Sixteen Kingdoms traditionally denotes the plethora of short-lived non-Chinese dynasties that from 303 came to rule the whole or parts of northern China. Many ethnic groups were involved, including ancestors of the Turks (such as the Hsiung-nu, possibly related to the Huns of late Roman history, and the Chieh), the Mongolians (Hsien-pei), and the Tibetans (Ti and Ch'iang). Most of these nomadic peoples, relatively few in number, had to some extent been Sinicized long before their ascent to power. In fact, some of them, notably the Ch'iang and the Hsiung-nu, had already since late Han times been allowed to live in the frontier regions within the Great Wall.

The barbarian rulers thus set up semi-Sinicized states, in which the foreign element constituted a military aristocracy and the nucleus of the armed forces. Since in administrative matters they lacked experience, and since their own tribal institutions were not adapted to the complicated task of ruling a large agrarian society, they had to make use of traditional Chinese ways of government. In doing so, they faced the dilemma that has ever since confronted foreign rulers on Chinese soil: the tension that existed between the need to preserve their own ethnic identity (and their position as Herrenvolk), on the one hand, and, on the other, the practical necessity of using Chinese literati and members of prominent Chinese families in order to rule at all. In spite of various and sometimes highly interesting experiments, most of these short-lived empires did not survive this tension. Significantly, the only one that proved to have more lasting power and that was able to unify the whole of northern China, the T'o-pa, or Northern Wei (386–534/535), was largely Sinicized within one century. In the late 5th century the court even forbade the use of the original T'o-pa language, dress, customs, and surnames. This policy of conscious acculturation was further symbolized by the transfer of the Northern Wei capital from the northern frontier region to the ancient Imperial residence of Lo-yang.

Sinicization of the foreign conquerors

Thus, toward the end of the period of division, the north had become more homogeneous as the result of a long process of adaptation. The most important factor in this process may have been the rehabilitation of the Chinese agrarian economy under the Northern Wei, stimulated by fiscal reform and redistribution of land (*c.* AD 500). The landed gentry again became the backbone of society, and the primitive rulers of nomadic origin simply had to conform to their way of life. Another factor may be sought in the intrinsic superiority of Chinese upper-class culture: in order to play the role of the "Son of Heaven," the barbarian court had to adopt the complicated rules of Chinese ritual and etiquette. In order to surround themselves with an aura of legitimacy, the foreign conquerors had to express themselves in terms of Chinese culture. In doing so, they invariably lost their own identity. History has repeated itself again and again: in this respect the 4th- and 5th-century Chieh and T'o-pa were but the forerunners of the Ch'ing, or Manchu, rulers in the 19th century.

In the early 6th century the Wei was divided between the Sinicized court and a faction of the nobility desperate to preserve its T'o-pa identity. Soon after 520 the Wei Empire disintegrated into rival northeastern and northwestern successor states. Northern China again became a battlefield for several decades. The Pei (Northern) Chou (557–581), strategically based in the rich basin of the Wei River, reunified the north (577). Four years later Yang Chien (better known by the name Wen-ti), a general of mixed Chinese and barbarian descent (but claiming to be a pure-blooded Chinese), usurped the throne and founded the Sui dynasty. In 589, having consolidated his regime, he crossed the Yangtze River and overthrew the last of the Chinese dynasties at Chien-k'ang. After almost four centuries of division and political decay, China was again united under one central government, which, in spite of its short duration, would lay the foundation of the great T'ang Empire.

## INTELLECTUAL AND RELIGIOUS TRENDS

**Confucianism and philosophical Taoism.** The social and political upheaval of the late 2nd and 3rd centuries AD was accompanied by intense intellectual activity. During the Han period, Confucianism had been slowly adopted as an ideology and had gradually come to provide the officially accepted norms, morals, and ritual and social behaviour regulating the relations between ruler and subject.

By the beginning of the 3rd century, however, Confucianism had lost its prestige: it had obviously failed to save the empire from disintegration or to safeguard the privileges of the ruling elite. Disappointed members of the scholar-official class started to look elsewhere. Thus, in the 3rd century there was a revival of various all-but-forgotten schools of thought: Legalism, with its insistence upon harsh measures, intended to reestablish law and order; Mohism and the ancient school of Dialecticians; and, above all, a renewed interest in the earliest Taoist philosophers, Lao-tzu and Chuang-tzu. In general, this movement did not mean a return to ancient Taoist quietism and consequently a rejection of Confucianism. With the breakdown of the elaborate scholastic doctrine that had formed the official Han ideology, Confucianism had been deprived of its metaphysical superstructure, and this

vacuum was now filled by a whole set of philosophical ideas and speculations, largely of Taoist provenance.

*Ming-chiao* and Hsüan Hsüeh

Within this movement, two trends came to dominate the intellectual life of the cultured minority. One of these was closely related to the practical affairs of government and stressed the importance of social duties, ritual, law, and the study of human characteristics. This mixture of Confucian and Legalist notions was called *ming-chiao,* "the doctrine of names" ("names" in ancient Confucian parlance designating the various social functions—father, ruler, subject, etc.—that an individual could have in society). The other trend was marked by a profound interest in ontological and metaphysical problems: the quest for a permanent substratum (called *t'i,* "substance") behind the world of change (called *yung,* "function"). It started from the assumption that all temporally and spatially limited phenomena—anything "nameable"; all movement, change, and diversity; in short, all "being"—is produced and sustained by one impersonal principle, which is unlimited, unnameable, unmoving, unchanging, and undiversified. This important movement, which found its scriptural support both in Taoist and in drastically reinterpreted Confucian sources, was known as Hsüan Hsüeh ("Dark Learning"); it came to reign supreme in cultural circles, especially at Chien-k'ang during the period of division, and represented the more abstract, unworldly, and idealistic tendency in early medieval Chinese thought.

The proponents of Hsüan Hsüeh undoubtedly still regarded themselves as true Confucians. To them, Confucius was not simply the great teacher who had fixed the rules of social behaviour for all time but was the enlightened sage who had inwardly recognized the ultimate reality but had kept silent about it in his worldly teachings, knowing that these mysteries could not be expressed in words. Hence, his doctrine was supposed to be an expedient, a mere set of ad hoc rules intended to answer the practical needs of the times. This concept of "hidden saintliness" and the "expedient" character of the canonical teachings came to play a very important role in upper-class Buddhism.

Hsüan Hsüeh is sometimes referred to by the term Neo-Taoism, but this confuses the issue. It was both created by and intended for literati and scholar-officials—not Taoist masters and hermits. The theories of such thinkers as Hsi K'ang (223–262), who, with their quest for immortality and their extreme antiritualism, were much nearer to the spirit of Taoism, hardly belong to the sphere of Hsüan Hsüeh, and the greatest Taoist author of this period, Ko Hung (*c.* 283–330), was clearly opposed to these mystic speculations.

The popularity of Hsüan Hsüeh was closely related to the practice of "pure conversation" (*ch'ing-t'an*), a special type of philosophical discourse much in vogue among the cultured upper class from the 3rd century onward. In the earliest phase, the main theme of such discussion—a highly formalized critique of the personal qualities of well-known contemporaries—still had a concrete function in political life ("characterization" of persons was the basis of recommendation of clients for official posts and had largely taken the place of the earlier methods of selection of officials by court examinations). By the 4th century, however, *ch'ing-t'an* meetings had evaporated into a refined and very exclusive pastime of the aristocratic elite, a kind of salon in which "eloquent gentlemen" expressed some philosophical or artistic theme in elegant and abstruse words. It is obvious that much of Hsüan Hsüeh had become divorced from the realities of life and afforded an escape from it.

True Confucianism had thus lost much of its influence. In the north the not yet Sinicized "barbarian" rulers were interested in Confucianism mainly as a system of court ritual; ideologically, they were more attracted by the magical powers of Buddhist and Taoist masters. In the south the disillusioned aristocratic exiles, doomed by circumstances to lead a life of elegant inactivity, had little use for a doctrine that preached the duties of government and the regulation of human society as its highest goals, although many families preserved Confucian learning and clung to Confucian mores. In this period of internal division and political weakness, Confucianism had to hibernate; soon after the Sui had reunited the empire, it would wake up again.

**Taoism.** The suppression of the Yellow Turbans and other Taoist religious movements in AD 184 had left the Taoist church decapitated. With the elimination of its highest leadership, the movement had fallen apart into many small religious communities or parishes, each led by a local Taoist master (*tao-shih*), assisted by a council of wealthy Taoist laity. Under such circumstances, local Taoist masters could easily become leaders of independent sectarian movements. They could also, in times of unrest, use their charismatic power to play a leading part in local rebellions. In the early medieval period, Taoism at the grass-roots level continued to play this double role: it had an integrating function by providing spiritual consolation and ritualized forms of communal activity, but it could also play a disintegrating role as a potential source of subversive movements. The authorities naturally were well aware of this. Taoist rebellions periodically broke out in this period; and, although some masters occasionally became influential at court, the governments, both northern and southern, maintained a cautious reserve toward the Taoist religion. It was never stimulated and patronized to an extent comparable with Buddhism.

Development of sectarian movements

It would be wrong to speak of Taoism as a popular religion. Taoism counted its devotees even among the highest nobility. In view of the expensive ceremonies, the costly ingredients used in Taoist alchemy (notably cinnabar), and the almost unlimited amount of spare time required from the serious practitioner, one may assume that only the well-to-do were able to follow the road toward salvation. But they were mostly individual seekers; in the 3rd and 4th centuries there gradually grew a distinction between individual (and mainly upper-class) Taoism and the popular, collective creed of the simple devotees. In fact, Taoism has always been a huge complex of many different beliefs, cults, and practices. Most of these can be traced back to Tung Han times; after the 3rd century they were influenced more and more by Buddhism.

The basic ideal of Taoist religion—the attainment of bodily immortality in a kind of indestructible "astral body" and the realization of the state of *hsien,* or Taoist "immortal"—remained alive. It was to be pursued by a series of individual practices: dietary control, gymnastics, good deeds, and meditation and visualization of the innumerable gods and spirits that were supposed to dwell inside the microcosmos of the body. Famous literati, such as the poet Hsi K'ang (223–262) and the calligrapher Wang Hsi-chih (321–379), devoted much of their life to such practices. They combined various methods, ranging from mystic self-identification with the all-embracing Tao to the use of charms and experiments in alchemy.

During the 4th century the development of Taoism seems to have reached a new stage. At that time in southern China there was already an ancient school of esoteric learning, exemplified by Ko Hung. The retreat of the Chin to southern China in the early 4th century brought to that region the organized religion and priesthood that had arisen in the north and west during the Tung Han. In this context, new priestly cults arose in the south, their teaching connected with a series of revelations, the first through Yang Hsi, which led to the formation first of the Shang-ch'ing sect and later to the rival Ling-pao sect. By the end of the period of division, Taoism had its own canons of scriptural writings, much influenced by Buddhist models but forming a quite independent religious tradition.

Popular Taoist ceremonies

The other, collective, and more popular form of Taoism, practiced in the communities or parishes throughout the country, was characterized by communal ceremonies (*chai,* "fasting sessions," and *ch'u,* "banquets") held by groups of Taoist families under the guidance of the local master, both on fixed dates and on special occasions. The purpose of such meetings was the collective elimination of sins (evil deeds being considered as the main cause of sickness and premature death) through incantations, deafening music, fasting, and the display of penance and remorse. The gatherings sometimes lasted several days and nights, and, according to the indignant reports of their Buddhist adversaries, they were ecstatic and sometimes even orgias-

tic. The allegation of sexual excesses and promiscuity may have been stimulated by the fact, unknown in Confucian and Buddhist ritual, that in Taoist meetings both men and women took part.

The Taoist parish as an organization and the Taoist master who led it relied on two sources of income: the presents made by devotee families at ceremonial gatherings and the regular "Heavenly tax," or yearly contribution of five bushels of rice, which every family was due to pay on the seventh day of the seventh month. The office of Taoist master was hereditary, within one family; in the early centuries Taoist priests usually married. Because Buddhist influence also increased at this humble level, however, the *tao-shih* more and more came to resemble the Buddhist clergy, especially since most Taoist priests, at least from the 5th century onward, went to live in Taoist monasteries with their wives and children. In the 6th century, when Buddhism became paramount, some Taoist leaders introduced celibacy; in Sui times the unmarried state had become general, and the Taoist clergy with its monks and nuns had evolved into a counterpart of the Buddhist *saṅgha*. Unlike Buddhist monasteries, the Taoist monasteries and clergy never developed great economic power.

In spite of their resemblance or, perhaps, because of it, the two creeds were bitterly opposed throughout the period. Taoist masters were often involved in anti-Buddhist propaganda and persecution. As an answer to Buddhist claims of superiority, Taoist masters even developed the curious theory that the Buddha had been only a manifestation of Lao-tzu, who had preached to the Indians a debased form of Taoism, which naturally should not be reintroduced into China; this theme can be traced in Buddhist and Taoist polemic literature from the 4th to the 13th century.

The rise of Buddhism

**Buddhism.** The Buddhist age of China began in the 4th century. Several factors contributed to the extraordinary expansion and absorption of the foreign religion after about 300, both in the Chinese south and in the occupied north. A negative factor was the absence of a unified Confucian state, which naturally would be inclined to suppress a creed whose basic tenets (notably the monastic life and the pursuit of individual salvation outside family and society) were clearly opposed to the ideals of Confucianism. The popularity of Hsüan Hsüeh was a positive and powerful factor. Especially in the south, Mahāyāna Buddhism, thoroughly amalgamated with Hsüan Hsüeh, was preached by cultured monks in the circles of the Chien-k'ang aristocracy, where it became extremely popular.

Another stimulus for the growth of Buddhism was the relative security and prosperity of monastic life. In a countryside devastated by war and rebellion, innumerable small peasants preferred to give up their independence and to avoid the scourges of heavy taxation, forced labour, and deportation by joining the large estates of the nobility as serfs, where they would get at least a minimum of protection. This process of tax evasion and the consequent extension of the manorial system also stimulated the growth of Buddhist monasteries as landowning institutions, peopled with both monks and families of hereditary temple serfs. By the beginning of the 6th century the monasteries had become an economic power of the first order, which, moreover, enjoyed special privileges (such as exemption from taxes). This, indeed, became a main source of tension between clergy and government and occasionally led to anti-Buddhist movements and harsh restrictive measures imposed upon the Buddhist church (446–452 and again in 574–578).

The monastic life attracted many members of the gentry as well. In these times of turmoil, the official career was beset with dangers, and the monastery offered a hiding place to literati who tried to keep clear of the intrigues and feuds of higher official circles; thus the ancient Chinese ideal of the retired scholar merged with the new Buddhist ideal of the monastic life. Many large monasteries thereby became centres of learning and culture and so became even more attractive to members of minor gentry families, for whom the higher posts in government in any event would be unattainable. Buddhist institutions offered a kind of "internal democracy"—a fact of great social importance in the history of class-ridden medieval China. Finally, Buddhism was patronized by most of the "barbarian" rulers in the north. At first, they were mainly attracted by the pomp and magical power of Buddhist ritual. Later, other motivations were added to this. Unwilling to rely too much upon Chinese ministers with their following of clan members and clients, they preferred to make use of Buddhist masters, who as unmarried individuals were totally dependent on the ruler's favour. Ideologically, Buddhism was less "Chinese" than Confucianism, especially in the north, where the connections with Central Asia constantly reinforced its international and universalistic character. This peculiar "Sino-barbarian" nature of northern Buddhism with its foreign preachers and its huge translation projects strongly contrasts with the south, where Buddhism in the 4th century was already fully domesticated.

Because of all these circumstances, the large-scale development of Chinese Buddhism started only after the barbarian invasions of the early 4th century. In the 3rd century the picture did not basically differ from Han times: there are indications that Buddhism was still largely a religion of foreigners on Chinese soil (apart from some activity involving the translation of Buddhist scriptures). But by the 4th century the picture was changing. At the southern Chinese court in Chien-k'ang there was forming a clerical elite of Chinese monks and propagators of a completely Sinicized Buddhism, strongly amalgamated with Hsüan Hsüeh, and their sophisticated creed was being spread among the southern gentry. Starting at Chien-k'ang and in northern Chekiang (the Hang-chou region), this trend was further developed in the late 4th and early 5th centuries in other centres throughout the middle and lower Yangtze Basin. The highest flowering of this uniquely "Chinese" type of Buddhism took place in the early 5th century.

In the north the climax of Buddhist activity and Imperial patronage occurred under the Wei, especially after the beginning of their policy of conscious Sinicization. The T'o-pa court and the great families vied with each other in building temples and granting land and money to the monasteries; the monumental cave temples at Yün-kang are lasting proof of this large-scale Imperial protection. This also had its dark side: in the north the Buddhist clergy became closely tied with secular government, and the lavish treatment of the church was counterbalanced by repeated attempts at government control. It may also be noted that the north remained open to influences brought by traveling monks from Central Asia, and an enormous amount of Indian Buddhist texts of all schools and eras was translated.

Popular Buddhism

Little is known of the beginnings of popular Buddhism. Among the masses there was, to judge from Taoist materials, an intense mingling of Buddhist and popular Taoist notions and practices, such as communal festivals and the worship of local Taoist and Buddhist saints. At this level, simple devotionalism was no doubt far more influential than the scriptural teachings. It is also possible that the oral recital of Buddhist scriptures (mainly edifying tales) had already inspired the development of vernacular literature.

In any event, the constant amalgamation of Buddhism, Taoism, and the innumerable local cults whose history went back into high antiquity was to continue for centuries, eventually producing an amorphous mass of creeds and practices collectively known as Chinese popular religion. (E.Z./D.C.T.)

## The Sui dynasty

The Sui dynasty (581–618), which reunified China after nearly four centuries of political fragmentation during which the north and south had developed in different ways, played a part far more important than its short span would suggest. In the same way that the Ch'in rulers of the 3rd century BC had unified China after the Chan-kuo (Warring States) period, so the Sui brought China together again and set up many institutions that were to be adopted by their successors, the T'ang. Like the Ch'in, however, the Sui overstrained their resources and fell. And also as in the case of the Ch'in, traditional history has judged the Sui somewhat unfairly; it has stressed the harshness of the

Sui regime and the megalomania of its second emperor, giving too little credit for its many positive achievements.

eign of Ven-ti

Wen-ti, the founder (reigned 581–604) of the Sui dynasty, was a high-ranking official at the Pei (Northern) Chou court, a member of one of the powerful northwestern aristocratic families that had taken service under the successive non-Chinese royal houses in northern China and had intermarried with the families of their foreign masters. The Pei Chou had recently (577) reunified northern China by the conquest of the rival northeastern dynasty of Pei Ch'i. But political life in the northern courts was extremely unstable. The succession of an apparently deranged and irresponsible young emperor to the Chou throne in 578 set off a train of court intrigues, plots, and murders. Wen-ti was able to install a child as puppet emperor and then to seize the throne for himself.

In control of all of northern China and in command of formidable armies, he immediately set about establishing order within his frontiers. He built himself a grand new capital, Ta-hsing, close to the site of the old Ch'in and Han capitals, a city erected very quickly with a prodigal use of compulsory labour. This great city remained (under the name Ch'ang-an) the capital of the Sui and T'ang dynasties and the principal seat of government until the beginning of the 10th century.

Wen-ti also took quick action to protect the frontiers of his new state. China during the 6th century had a formidable northern neighbour in the Turks (T'u-chüeh), who controlled the steppe from the borders of Manchuria to the frontiers of the Byzantine and Sāsānian empires. At the time of Wen-ti's seizure of power, the Turks were splitting into two great empires, an eastern one dominating the Chinese northern frontier from Manchuria to Kansu, and a western one stretching in a vast arc north of the Tarim Basin into Central Asia. Wen-ti encouraged this split by supporting the khan of the western Turks, Tardu. Throughout his reign Wen-ti also pursued a policy of encouraging factional strife among the eastern Turks. At the same time he strengthened his defenses in the north by repairing the Great Wall. In the northwest in the area around the Koko Nor (Blue Lake), he defeated the T'u-yü-hun people, who from time to time raided the border territories.

By the late 580s Wen-ti's state was stable and secure enough for him to take the final step toward the reunification of the whole country. In 587 he dethroned the Emperor of the Hou (Later) Liang, the state that had ruled the middle Yangtze Valley as a puppet of the Pei Chou since 555. In 589 he overwhelmed the last southern dynasty, the Ch'en, which had put up only token resistance. Several rebellions against the Sui regime subsequently broke out in the south, but these were easily quelled. Wen-ti now ruled over a firmly reunited empire.

### WEN-TI'S INSTITUTIONAL REFORMS

Wen-ti's achievement consisted of much more than strengthening and reunifying the empire. He provided it with uniform institutions and established a pattern of government that survived into the T'ang dynasty and beyond. A hardworking administrator, he employed a number of extremely able ministers who combined skill in practical statecraft with a flexible approach to ideological problems. They revived the Confucian state rituals to win favour with the literati and to establish a link with the empire of the Han; and at the same time they fostered Buddhism, the dominant religion of the south, attempting to establish the Emperor's image as an ideal Buddhist saint-king.

Wen-ti's lasting success, however, was in practical politics and institutional reforms. In the last days of the Pei Chou he had been responsible for a revision of the laws, and one of his first acts on becoming emperor was to promulgate

he legal ode

a penal code, the New Code of 581. In 583 his ministers compiled a revised code, the K'ai-huang code, and administrative statutes. These were far simpler than the laws of the Pei Chou, and more lenient. Considerable pains were taken to ensure that local officials studied and enforced the new laws. Toward the end of Wen-ti's reign, when neo-Legalist political advisers gained ascendancy at court, the application of the laws became increasingly strict. The K'ai-huang code and statutes have not survived, but they provided the pattern for the T'ang code, the most influential body of law in the history of the Far East.

The central government under Wen-ti developed into a complex apparatus of ministries, boards, courts, and directorates. The conduct of its personnel was supervised by another organ, the censorate. The emperor presided over this apparatus, and all orders and legislation were issued in his name. He was assisted by the heads of the three central ministries who acted as Counselors on State Affairs (I kuo-cheng). This system later provided the basic framework for the central government of the early T'ang.

Even more important, he carried out a sweeping reform and rationalization of local government. The three-level system of local administration inherited from Han times had been reduced to chaos during the 5th and 6th centuries by excessive subdivision; there were innumerable local districts, some of them very small and dominated by single families. Wen-ti created a simplified structure in which a much reduced number of counties was directly subordinated to prefectures. He also rationalized the chaotic rural administrative units into a uniform system of *hsiang*. Appointments to the chief offices in prefectures and counties were now made by the central government, rather than filled by members of local influential families as had been the practice. This reform ensured that local officials would be agents of the central government. It also integrated local officials into the normal pattern of bureaucratic promotion and in time produced a more homogeneous civil service.

Local government

Since the registration of population had fallen into chaos under the Pei Chou, a careful new census was carried out during the 580s. It recorded the age, status, and landed possessions of all the members of each household in the empire. On the basis of this census, the land allocation system employed under the successive northern dynasties since the end of the 5th century was reimposed. The tax system also followed the old model of head taxes levied in grain and silk at a uniform rate. The taxable age was raised, and the annual period of labour service to which all taxpayers were liable was reduced.

Wen-ti's government, in spite of his frontier campaigns and vast construction works, was very economical and frugal. By the 590s he had accumulated great reserves, and when the Ch'en territories were incorporated into his empire he was in a position to exempt the new population from 10 years of taxes to help ensure their loyalty.

The military system, likewise, was founded upon that of the northern dynasties; the Imperial forces were organized in militias. They served regular annual turns of duty but lived at home during the rest of the year and were largely self-supporting. Many troops were settled in military colonies on the frontiers to make the garrisons self-sufficient. Only when there was a campaign did the costs of the military establishment soar.

### INTEGRATION OF THE SOUTH

The second Sui emperor, Yang-ti (reigned 604–617), has been depicted as a supreme example of arrogance, extravagance, and personal depravity who squandered his patrimony in megalomaniac construction projects and unwise military adventures. This mythical Yang-ti was to a large extent the product of the hostile record written of his reign shortly after his death. His reign began well enough, continuing the trends begun under Wen-ti. A further revision of the law code with a general reduction of penalties was carried out in 607.

Reign of Yang-ti

Yang-ti's principal achievement was the integration of the south more firmly into a unified China. There is little evidence that the south was ever completely brought into line with all the administrative practices of the north; the land allocation system seems unlikely to have been enforced there, and it is probable that the registration of the population, the essential foundation for the whole fiscal and military system, was only incompletely carried out in the old Ch'en territories. But Yang-ti himself was personally very much involved with the south. Married to a princess from the southern state of Liang, he had spent 591–600 as viceroy for the southern territories; their

successful integration into the Sui Empire after the initial wave of risings was largely due to his administration and to the generally clement policies employed in the former Ch'en territories.

His identification with the southern interest was one of the reasons he began the establishment of an examination system, based upon the Confucian Classical curriculum, as a means of drawing into the bureaucracy scholars from the southern and northeastern elites that had preserved traditions of Confucian learning. Hitherto, the court had been dominated by the generally less cultivated aristocratic families of mixed blood from northwestern China.

Yang-ti also attempted to weaken the predominance of the northwest by building a second great capital city at Lo-yang on the border of the eastern plains. This capital was not only distant from the home territories of the northwestern aristocrats but also easily provisioned from the rich farmlands of Hopeh and Honan. The new city was constructed in a great hurry, employing vast numbers of labourers both in building and in transporting the timber and other materials required. Yang-ti also built new palaces and an immense Imperial park, again with a prodigal use of labour.

Development of the canals

Another grandiose plan aimed at unifying the empire was to develop still further the canal system his father had begun in the metropolitan region and to construct a great canal linking Lo-yang with the Huai River and with the southern capital, Chiang-tu (Yang-chou), on the Yangtze. Much of this route followed existing rivers and ancient canals, but it was still an immense undertaking that employed masses of forced labourers working under appalling conditions. In 605 the canal system was opened between the capital at Lo-yang and the Yangtze, and in 610 it was extended south of the Yangtze to Hang-chou. At the same time, in preparation for campaigns in Manchuria and on the Korean frontier, another great canal was built northward from Lo-yang to the vicinity of modern Peking. By 611 the whole eastern plain had a canal system linking the major river systems of northern China and providing a trunk route from the Yangtze Delta to the northern frontier. The construction of these waterways was inordinately expensive, caused terrible suffering, and left a legacy of widespread social unrest; but in the long term the transportation system was to be a most important factor for maintaining a unified empire. Further hardship was caused by the mass levies of labour required to rebuild and strengthen the Great Wall in Shansi in 607 and 608 as a precaution against the resurgent eastern Turks.

FOREIGN AFFAIRS UNDER YANG-TI

In addition to these farsighted construction works, Yang-ti also pursued a very active foreign policy. An expedition to the south established sovereignty over the old Chinese settlement in Tongking and over the Champa state of Lin-yi in central Nam Viet. Several expeditions were sent to Taiwan, and relations with Japan were opened. The T'u-yü-hun people were driven out of Kansu and Tsinghai, and Sui colonies were established along the great western trade routes. The rulers of the various petty local states of Central Asia and the king of Kao-ch'ang (T'u-lu-p'an) became tributaries. A prosperous trade with Central Asia and the West grew up.

Relations with the Turks

The principal foreign threat was still posed by the Turks. These had now been completely split into the eastern Turks, who occupied most of the Chinese northern frontier, and the immensely powerful western Turks, whose dominions stretched westward to the north of the Tarim Basin as far as Sāsānian Persia and Afghanistan. During the early part of Yang-ti's reign the western Turks, whose ruler, Ch'u-lo, was half-Chinese, were on good terms with the Sui. In 610, however, Yang-ti supported a rival, She-kuei, who drove out Ch'u-lo. The latter took service, with an army of 10,000 followers, at Yang-ti's court. When Sui power began to wane after 612, the western Turks under She-kuei gradually replaced the Sui garrisons in Central Asia and established control over the states of the Tarim Basin. The eastern Turks had remained on good terms with the Sui, their khans being married to Chinese princesses. In 613 P'ei Chü, Yang-ti's principal agent in dealing with the foreign states of the north, attempted unsuccessfully to dethrone the eastern Turkish khan and split up his khanate. Relations with the Turks rapidly deteriorated, and in the last years of his reign Yang-ti had to contend with a hostile and extremely powerful neighbour.

Campaigns in Korea

His most costly venture was a series of campaigns in Korea. At that time Korea was divided into three kingdoms, of which the northern one, Koguryŏ, was the most important and powerful. It was hostile to the Chinese and refused to pay homage to Yang-ti. Yang-ti made careful preparations for a punitive campaign on a grand scale, including construction of the Yung-chi-ch'ü canal from Lo-yang to Peking. In 611 the canal was completed; a great army and masses of supplies were collected, but terrible floods in Hopeh delayed the campaign. During 612, 613, and 614 Yang-ti campaigned against the Koreans. The first two campaigns were unsuccessful and were accompanied by the outbreak of many minor rebellions in Shantung and southern Hopeh. The severe repression that followed led to outbreaks of disorder throughout the empire. In 614 yet another army was sent into Korea and threatened the capital at P'yŏngyang, but it had to withdraw without a decisive victory. These futile campaigns distracted Yang-ti's attention from the increasingly vital internal problems of his empire, involved an immense loss of life and matériel, and caused terrible hardships among the civilian population. They left the Sui demoralized, militarily crippled, and financially ruined.

At this point, Yang-ti decided to secure his relations with his northern neighbours. His envoy, P'ei Chü, had continued to intrigue against the eastern Turkish khan, in spite of the fact that the Sui were no longer in a position of strength. When in the summer of 615 Yang-ti went to inspect the defenses of the Great Wall, he was surrounded and besieged by the Turks at Yen-men; he was rescued only after a month of peril.

Fall of the Sui dynasty

Rebellions and uprisings soon broke out in every region of the empire. Late in 616 Yang-ti decided to withdraw to his southern capital of Chiang-tu, and much of northern China was divided among rebel regimes contending with one another for the succession to the empire. Yang-ti remained nominally emperor until the spring of 618, when he was murdered by members of his entourage at Chiang-tu. But by 617 the real powers in China had become the various local rebels: Li Mi in the area around Lo-yang, Tou Chien-te in the northeast, Hsüeh Chü in the far northwest, and Li Yüan (who remained nominally loyal but had established a local position of great power) in Shansi. At the beginning of 617 Li Yüan inflicted a great defeat on the eastern Turks and thus consolidated his local power in the impregnable mountainous area around T'ai-yüan. In the summer of 617 he raised an army and marched on the capital with the aid of the Turks and other local forces; Ch'ang-an fell at year's end. Hsüeh Chü's northwestern rebels were crushed, and the armies of Li Yüan occupied Szechwan and the Han River valley. A Sui prince, Kung-ti, was enthroned as "emperor" in 617, while Yang-ti was designated "retired emperor." In the summer of 618, after Yang-ti's death, Li Yüan (known by his temple name, Kao-tsu) deposed his puppet prince and proclaimed himself emperor of a new dynasty, the T'ang, which was to remain in power for nearly three centuries.

## The T'ang dynasty

EARLY T'ANG (618–626)

Kao-tsu's struggle for control

When Kao-tsu became emperor (reigned 618–626), he was still only one among the contenders for control of the empire of the Sui. It was several years before the empire was entirely pacified. After the suppression of Hsüeh Chü and the pacification of the northwest, the T'ang had to contend with three principal rival forces: the Sui remnants commanded by Wang Shih-ch'ung at Lo-yang, the rebel Li Mi in Honan, the rebel Tou Chien-te in Hopeh, and Yü-wen Hua-chi, who had assassinated the previous Sui emperor Yang-ti and now led the remnants of the Sui's southern armies. Wang Shih-ch'ung set up a grandson of Yang-ti at Lo-yang as the new Sui emperor. Yü-wen Hua-chi led his armies to attack Lo-yang, and Wang Shih-

ch'ung persuaded Li Mi to return to his allegiance with the Sui and help him fight Yü-wen Hua-chi. Li Mi defeated Yü-wen Hua-chi's armies but depleted his own forces seriously. Wang Shih-ch'ung, seeing the chance to dispose of his most immediate rival, took over Lo-yang and routed Li Mi's forces. Li Mi fled to Ch'ang-an and submitted to the T'ang. In the spring of 619 Wang Shih-ch'ung deposed the puppet Sui prince at Lo-yang and proclaimed himself emperor.

The T'ang armies gradually forced him to give ground in Honan, and by 621 Kao-tsu's son Li Shih-min was besieging him in Lo-yang. At this time Wang Shih-ch'ung attempted to form an alliance with the most powerful of all the Sui rebels, Tou Chien-te, who controlled much of Hopeh and who had completed the defeat of Yü-wen Hua-chi's forces in 619. He held the key area of southern Hopeh, where he had successfully resisted both the T'ang armies and the forces of Wang and Li. Tou now agreed to come to the aid of the beleaguered Wang, but in the spring of 621 Li Shih-min attacked his army before it could lift the siege, routed it, and captured Tou. Wang then capitulated. The T'ang had thus disposed of its two most powerful rivals and extended its control over most of the eastern plain, the most populous and prosperous region of China.

This was not the end of resistance to the T'ang conquest. Most of the surrendered rebel forces had been treated leniently, and their leaders were often confirmed in office or given posts in the T'ang administration. Tou and Wang, however, were dealt with severely, Tou being executed and Wang murdered on his way to exile. At the end of 621 Tou's partisans in the northeast again rebelled under Liu Hei-t'a and recaptured most of the northeast. He was finally defeated by a T'ang army under the crown prince Chien-ch'eng at the beginning of 623. The prolonged resistance in Hopeh and the comparatively harsh T'ang conquest of the region were the beginning of resistance and hostility in the northeast that continued to some degree throughout the T'ang dynasty.

Resistance was not confined to the northeast. Liu Wu-chou in the far north of Shansi, who had been a constant threat since 619, was finally defeated and killed by his former Turkish allies in 622. In the south during the confusion at the end of the Sui, Hsiao Hsien had set himself up as emperor of Liang, controlling the central Yangtze region, Kiangsi, Kwangtung, and Annam. The T'ang army descended the Yangtze from Szechwan with a great fleet and defeated Hsiao Hsien's forces in two crucial naval battles. In 621 Hsiao Hsien surrendered to the T'ang, who thus gained control of the central Yangtze and the far south. The southeast was occupied by another rebel, Li Tzu-t'ung, based in Chekiang. He, too, was decisively defeated near modern Nanking at the end of 621. As had been the case with Hsiao Hsien's dominions, the southeast was incorporated into the T'ang Empire with a minimum of fighting and resistance. A last southern rebellion by Fu Kung-shih, a general who set up an independent regime at Nanking in 624, was speedily suppressed. After a decade of war and disorder, the empire was completely pacified and unified under the T'ang house.

**Administration of the state.** The T'ang unification had been far more prolonged and bloody than the Sui conquest. That the T'ang regime lasted for nearly three centuries rather than three decades, as with the Sui, was largely the result of the system of government imposed on the conquered territories. The emperor Kao-tsu's role in the T'ang conquest was understated in the traditional histories compiled under his successor T'ai-tsung (Li Shih-min; reigned 626–649), which portrayed T'ai-tsung as the prime mover in the establishment of the dynasty. T'ai-tsung certainly played a major role in the campaigns, but Kao-tsu was no figurehead. Not only did he direct the many complex military operations, but he also established the basic institutions of the T'ang state, which proved practicable not only for a rapidly developing Chinese society but also for the first centralized states in societies as diverse as those of Japan, Korea, Nam Viet, and the southwestern kingdom of Nan Chao.

The T'ang pattern of central control

The structure of the new central administration resembled that of Wen-ti's time, with its ministries, boards, courts, and directorates. There was no radical change in the dominant group at court. Most of the highest ranks in the bureaucracy were filled by former Sui officials, many of whom had been the new emperor's colleagues when he was governor in T'ai-yüan, or by descendants of officials of the Pei Chou, Pei Ch'i, or Sui, or of the royal houses of the northern and southern dynasties. The T'ang were related by marriage to the Sui royal house, and a majority of the chief ministers were related by marriage to either the T'ang or Sui Imperial family. The emperor's court was composed very largely of men of similar social origins. At this level the T'ang in its early years, like the Sui before it, continued the pattern of predominantly aristocratic rule that had dominated the history of the northern courts.

Kao-tsu also continued the pattern of local administration established under the Sui and maintained the strict control exercised by the central government over provincial appointments. In the first years after the T'ang conquest, many prefectures and counties were fragmented to provide offices for surrendered rebel leaders, surrendered Sui officials, and followers of the emperor. But these new local districts were gradually amalgamated and reduced in number, and by the 630s the pattern of local administration was very similar to that under the Sui. The merging of the local officials into the main bureaucracy, however, took time; ambitious men still looked upon local posts as "exile" from the main current of official promotion at the capital. Until well into the 8th century many local officials continued to serve for long terms, and the ideal of a regular circulation of officials prevailed only gradually.

Local government in early T'ang times had a considerable degree of independence, but each prefecture was in direct contact with the central ministries. In the spheres of activity that the administration regarded as crucial—registration, land allocation, tax collection, conscription of men for the army and for corvée duty, and the maintenance of law and order—prefects and county magistrates were expected to follow centrally codified law and procedure. They were, however, permitted to interpret the law to suit local conditions. Local influences remained very strong in the prefectures and counties. Most of the personnel in these divisions were local men, many of them members of families of petty functionaries.

**Fiscal and legal system.** Kao-tsu had inherited a bankrupt state, and most of his measures were aimed at simple and cheap administration. His bureaucracy was very small, at both the central and local levels. The expenses of government were largely met out of endowments of land attached to each office, the rents from which paid office expenses and salaries; by interest on funds of money allocated for similar purposes; and by services of taxpayers who performed many of the routine tasks of government as special duties, being exempted from tax in return.

Land distribution

Land distribution followed the equal allocation system used under the northern dynasties and the Sui. Every taxable male was entitled to a grant of land, part of which was to be returned when he ceased to be a taxpayer at 60, part of which was hereditary. The disposal of landed property was hedged around with restrictive conditions. Great landed estates were limited to members of the Imperial clan and powerful officials, to various state institutions, and to the Buddhist foundations. Although some land was hereditary, and more and more passed into the hereditary category with the passage of time, the lack of primogeniture meant that landholdings were fragmented among all the sons in each generation and thus tended to be small. It is unlikely that the system was ever enforced to the letter in any region, and it was probably never enforced at all in the south. But as a legal system governing registration of landed property and restricting its disposal, it remained in force until An Lu-shan's rebellion in the 8th century.

The tax system based on this land allocation system was also much the same as that under the Sui and preceding dynasties. Every adult male paid a head tax in grain and cloth and was liable to 20 days of work for the central government (normally commuted into a payment in cloth) and to a further period of work for the local authorities. Revenues were collected exclusively from the rural popu-

lation, the trade sector and the urban communities being exempt, and the system bore more heavily upon the poor, since it ignored the taxpayer's economic status.

The Sui had made a somewhat desultory attempt to provide China with a unified coinage. Kao-tsu set up mints and began the production of a good copper currency that remained standard throughout the T'ang era. But cash was in short supply during most of the 7th century and had to be supplemented by standard-sized lengths of silk. Counterfeiting was rife, particularly in the Yangtze Valley where the southern dynasties had supported a more highly monetized economy, and where the governments had exploited commerce as a source of revenue.

Law code

Kao-tsu also undertook a new codification of all centralized law, completed in 624. It comprised a code that embodied what were considered basic, unchanging normative rules, prescribing fixed penalties for defined offenses; statutes, comprising the general body of universally applicable administrative law; regulations, or codified legislation supplementary to the code and statutes; and ordinances, detailed procedural laws supplementing the statutes and issued by the departments of the central ministries. Under the early T'ang this body of codified law was revised every 20 years or so. The systematic effort to maintain a universally applicable codification of law and administrative practice was essential to the uniform system of administration that the T'ang succeeded in imposing throughout its diverse empire. The T'ang code proved remarkably durable. It was still considered authoritative as late as the 14th century and was used as a model by the Ming. It was also adopted with appropriate modifications in Japan in the early 8th century and by the Koreans and the Vietnamese at a much later date.

Kao-tsu thus laid down, in the very first years of the 7th century, institutions that survived until the middle of the 8th century. These provided strong central control, a high level of administrative standardization, and very economical administration.

## THE PERIOD OF T'ANG POWER (626–755)

Two of Kao-tsu's sons were rivals for the succession: the crown prince Chien-ch'eng and Li Shih-min, the general who had played a large part in the wars of unification. Their rivalry, and the factional strife it generated, reached a peak in 625–626 when it appeared that Chien-ch'eng was likely to succeed. In a military coup, Li Shih-min murdered Chien-ch'eng and another of his brothers and forced his father to abdicate in his favour. He succeeded to the throne in 626 and is known by his temple name, T'ai-tsung.

Reign of T'ai-tsung

**The "era of good government."** The reign of T'ai-tsung (626–649), known traditionally as the "era of good government of Chen-kuan," was not notable for innovations in administration. Generally, his policies developed and refined those of his father's reign. The distinctive element was the atmosphere of his administration and the close personal interplay between the sovereign and his unusually able team of Confucian advisers. It approached the Confucian ideal of a strong, able, energetic, yet fundamentally moral king seeking and accepting the advice of wise and capable ministers, advice that was basically ethical rather than technical. Some important changes in political organization were begun during his reign and were continued throughout the 7th century. The court remained almost exclusively the domain of men of aristocratic birth. But T'ai-tsung attempted to balance the regional groups among the aristocracy so as to prevent the dominance of any single region. They comprised the Kuan-lung group from the northwest, the Tai-pei group from Shansi, the Shantung group from Hopeh, and the southern group from the Yangtze Valley. The most powerful Hopeh clans were excluded from high office, but T'ai-tsung employed members of each of the other groups and of the lesser northeastern aristocracy in high administrative offices as well as in his consultative group of scholars.

A second change was the use of the examination system on a large scale. The Sui examinations had already been reestablished under Kao-tsu, who had also revived the Sui system of high-level schools at the capital. Under T'ai-tsung the schools were further expanded and new ones established. Measures were taken to standardize their curriculum; an official orthodox edition of the Classics with a standard commentary was completed in 638. The schools at the capital were mostly restricted to the sons of the nobility and of high-ranking officials. Other examination candidates, however, came from the local schools. The examinations were in principle open to all, but they provided relatively few new entrants to the bureaucracy. Most officials still entered service by other means—by hereditary privilege as sons of officials of the upper ranks or by promotion from the clerical service or the guards. The examinations demanded a high level of education in the traditional curriculum and were largely used as an alternative method of entry by younger sons of the aristocracy and by members of lesser families with a scholar-official background. Moreover, personal recommendation, the lobbying of examiners, and often a personal interview by the emperor played a large part. Even in late T'ang times, not more than 10 percent of officials were recruited by the examinations. The main effect of the examination system in T'ang times was to bring into being a highly educated court elite within the bureaucracy, to afford access to the upper levels of the bureaucracy for members of locally prominent clans, and in the long term to break the monopoly of political power held by the upper aristocracy. The employment of persons dependent for their position on the emperor and the dynasty, rather than upon birth and social standing, made it possible for the T'ang emperors to establish their own power and independence.

The examination system

In the early years there was a great debate as to whether the T'ang ought to reintroduce the feudal system used under the Chou and the Han, by which authority was delegated to members of the Imperial clan and powerful officials and generals who were enfeoffed with hereditary territorial jurisdictions. T'ai-tsung eventually settled on a centralized form of government through prefectures and counties staffed by members of a unified bureaucracy. The T'ang retained a nobility, but its "fiefs of maintenance" were merely lands whose revenues were earmarked for its use and gave it no territorial authority.

T'ai-tsung continued his father's economic policies, and government remained comparatively simple and cheap. He attempted to cut down the bureaucratic establishment at the capital and drastically reduced the number of local government divisions. The country was divided into 10 provinces, which were not permanent administrative units but "circuits" for occasional regional inspections of the local administrations; these tours were carried out by special commissioners, often members of the censorate, sent out from the capital. This gave the central government an additional means of maintaining standardized and efficient local administration. Measures to ensure tax relief for areas stricken by natural disasters, and the establishment of relief granaries to provide adequate reserves against famine, helped to ensure the prosperity of the countryside. T'ai-tsung's reign was a period of low prices and general prosperity.

Foreign policy

T'ai-tsung was also successful in his foreign policy. In 630 the eastern Turks were split by dissension among their leadership and by the rebellion of their subject peoples. Chinese forces invaded their territories, totally defeated them, and captured their khan, and T'ai-tsung was recognized as their supreme sovereign, the "Heavenly khan." Many of the surrendered Turks were settled on the Chinese frontier, and many served in the T'ang armies. A similar policy of encouraging internal dissension was later practiced against the western Turks, who split into two separate khanates for a while. In 642–643 a new khan reestablished a degree of unified control with Chinese support and agreed to become a tributary of the Chinese. To seal the alliance, T'ai-tsung married him to a Chinese princess.

The eclipse of Turkish power enabled T'ai-tsung to extend his power over the various small states of the Tarim Basin. By the late 640s a Chinese military administration had extended westward even beyond the limits of modern Sinkiang. To the north, in the region of the Orhon River and to the north of the Ordos Desert, the T'ang armies

defeated the Hsüeh-yen-t'o, former vassals of the eastern Turks, who became T'ang vassals in 646. The T'u-yü-hun in the region around Koko Nor (Blue Lake) caused considerable trouble in the early 630s. T'ai-tsung invaded their territory in 634 and defeated them, but they remained unsubdued and invaded Chinese territory several times.

The Chinese western dominions now extended farther than in the great days of the Han. Trade developed with the West, with Central Asia, and with India. The Chinese court received embassies from Sāsānian Persia and from the Byzantine Empire. The capital was thronged with foreign merchants and foreign monks and contained a variety of non-Chinese communities. The great cities had their Zoroastrian, Manichaean, and Nestorian temples, along with the Buddhist monasteries that had been a part of the Chinese scene for centuries.

T'ai-tsung's only failure in foreign policy was in Korea. The northern state of Koguryŏ had sent tribute regularly, but in 642 there was an internal coup; the new ruler attacked Silla, another T'ang vassal state in southern Korea. T'ai-tsung decided to invade Koguryŏ, against the advice of most of his ministers. The T'ang armies, in alliance with the Khitan in Manchuria and the two southern Korean states Paekche and Silla, invaded Koguryŏ in 645 but were forced to withdraw with heavy losses. Another inconclusive campaign was waged in 647, and the very end of T'ai-tsung's reign was spent in building a vast fleet and making costly preparations for a final expedition.

T'ai-tsung's last years also saw a decline in the firm grasp of the Emperor over politics at his court. In the 640s a bitter struggle for the succession developed when it became clear that the designated heir was mentally unstable. The court split into factions supporting various candidates. The final choice, Li Chih, prince of Chin (reigned 649–683; temple name Kao-tsung) was a weak character, but he had the support of the most powerful figures at court.

**Rise of the empress Wu-hou.** Kao-tsung was 21 when he ascended the throne. In his first years he was dominated by the remaining great statesmen of T'ai-tsung's court, above all by the Emperor's uncle Chang-sun Wu-chi. But real power soon passed from Kao-tsung into the hands of the empress Wu-hou, one of the most remarkable women in Chinese history. Wu-hou had been a low-ranking concubine of T'ai-tsung. She was taken into Kao-tsung's palace and, after a series of complex intrigues, managed in 655 to have the legitimate empress, Wang, deposed and herself appointed in her place. The struggle between the two was not simply a palace intrigue. Empress Wang, who was of noble descent, had the backing of the old northwestern aristocratic faction and of the great ministers surviving from T'ai-tsung's court. Wu-hou came from a family of lower standing from Tai-yüan. Her father had been one of Kao-tsu's original supporters, her mother a member of the Sui royal family. She seems to have been supported by the eastern aristocracy, by the lesser gentry, and by the lower ranking echelons of the bureaucracy.

Skillful intrigue of the empress Wu-hou

But her success was largely the result of her skill in intrigue, her dominant personality, and her utter ruthlessness. The deposed empress and another Imperial favourite were savagely murdered, and the next half century was marked by recurrent purges in which she hounded to death one group after another of real or imagined rivals. The good relationship between the emperor and his court, which had made T'ai-tsung's reign so successful, was speedily destroyed. Political life became precarious and insecure, at the mercy of the Empress' unpredictable whims. The first victims were the elder statesmen of T'ai-tsung's reign, who were exiled, murdered, or driven to suicide in 657–659. In 660 Kao-tsung suffered a stroke. He remained in precarious health for the rest of his reign, and Wu-hou took charge of the administration.

Although utterly unscrupulous in politics, she backed up her intrigues with policies designed to consolidate her position. In 657 Lo-yang was made the second capital. The whole court and administration were frequently transferred to Lo-yang, thus removing the centre of political power from the home region of the northwestern aristocracy. Ministries and court offices were duplicated; Lo-yang had to be equipped with all the costly public buildings needed for a capital. After Kao-tsung's death, Wu-hou took up permanent residence there. Kao-tsung and Wu-hou were obsessed by symbolism and religion. One favourite magician, holy man, or monk followed another. State rituals were radically changed. For symbolic reasons the names of all offices were altered, and the emperor took the new title of "Heavenly emperor."

The bureaucracy was rapidly inflated to a far greater size than in T'ai-tsung's time, many of the new posts being filled by candidates from the examination system who now began to attain the highest offices and thus to encroach on what had been the preserves of the aristocracy. Another blow at the aristocracy was struck by the compilation in 659 of a new genealogy of all the empire's eminent clans, which ranked families according to the official positions achieved by their members rather than by their traditional social standing. Needless to say, the first family of all was that of Wu-hou. The lower ranks of the bureaucracy, among whom the Empress found her most solid support, were encouraged by the creation of new posts, by greater opportunities for advancement, and by salary increases.

The Chinese were engaged in foreign wars throughout Kao-tsung's reign. Until 657 they waged continual war against the western Turks, finally defeating them and placing their territories as far as the valley of the Amu Darya under a nominal Chinese protectorate in 659–661. Kao-tsung also waged repeated campaigns against Koguryŏ in the late 650s and 660s. In 668 the T'ang forces took P'yŏngyang, the capital, and Koguryŏ was also placed under a protectorate. But by 676 rebellions forced the Chinese to withdraw to southern Manchuria, and all Korea was increasingly dominated by the rapidly expanding power of the southern Korean state of Silla. The eastern Turks, who had been settled along the northern border, rebelled in 679–681 and were quelled only after widespread destruction and heavy losses to the Chinese forces.

The most serious foreign threat in Kao-tsung's reign was the emergence of a new and powerful force to the west, the Tibetans (T'u-fan), a people who had exerted constant pressure on the northern border of Szechwan since the 630s. By 670 the Tibetans had driven the T'u-yü-hun from their homeland in the Koko Nor basin. The northwest had to be increasingly heavily fortified and garrisoned to guard against their repeated raids and incursions. After a series of difficult campaigns, they were finally checked in 679.

When Kao-tsung died in 683 he was succeeded by the young Chung-tsung, but Wu-hou was made empress dowager and immediately took control over the central administration. Within less than a year she had deposed Chung-tsung, who had shown unexpected signs of independence, and replaced him by another puppet emperor, Jui-tsung, who was kept secluded in the Inner Palace while Wu-hou held court and exercised the duties of sovereign.

In 684 disaffected members of the ruling class under Li Ching-yeh raised a serious rebellion at Yang-chou in the south, but this was speedily put down. The Empress instituted a reign of terror among the members of the T'ang royal family and officials, employing armies of agents and informers. Fear overshadowed the life of the court. The Empress herself became more and more obsessed with religious symbolism. She manipulated Buddhist scripture to justify her becoming sovereign and in 688 erected a Ming T'ang ("Hall of Light")—the symbolic supreme shrine to Heaven described in the Classics—a vast building put up with limitless extravagance. In 690 the Empress proclaimed that the dynasty had been changed from T'ang to Chou. She became formally the empress in her own right, the only woman sovereign in China's history. Jui-tsung, the Imperial heir, was given her surname, Wu; everybody with the surname Wu in the empire was exempted from taxation. Every prefecture was ordered to set up a temple in which the monks were to expound the notion that the Empress was an incarnation of Buddha. Lo-yang became the "holy capital," and the state cult was ceremoniously transferred there from Ch'ang-an. The remnants of the T'ang royal family who had not been murdered or banished were immured in the depths of the palace.

Wu-hou enthroned

Destructive and demoralizing as the effects of her policies must have been at the capital and at court, there

is little evidence of any general deterioration of administration in the empire. By 690 the worst excesses of her regime were past. In the years after she had proclaimed herself empress she retained the services and loyalty of a number of distinguished officials. The court, however, was still unstable, with unending changes of ministers, and the Empress remained susceptible to the influence of a series of worthless favourites. After 700 she gradually began to lose her grip on affairs.

The external affairs of the empire had meanwhile taken a turn for the worse. The Tibetans renewed their warfare on the frontier. In 696 the Khitan in Manchuria rebelled against their Chinese governor and overran part of Hopeh. The Chinese drove them out, with Turkish aid, in 697. The Chinese reoccupied Hopeh under a member of the Empress' family and carried out brutal reprisals against the population. In 698 the Turks in their turn invaded Hopeh and were driven off only by an army under the nominal command of the deposed emperor, Chung-tsung, who had been renamed heir apparent in place of Jui-tsung. The military crisis had forced the Empress to abandon any plan to keep the succession within her own family.

New taxes

The expenses of the empire began to call for new taxes. These took the form of a household levy—a graduated tax based on a property assessment upon everyone from the nobility down, including the urban population—and a land levy collected on an acreage basis. These new taxes were to be assessed on the basis of productivity or wealth, rather than on a uniform per capita basis. Some tried to evade taxes by illegally subdividing their households to reduce their liabilities. There was a large-scale migration of peasant families fleeing from oppression and heavy taxation in the Hopeh and Shantung area. This migration of peasants, who settled as unregistered squatters on vacant land in central and southern China and no longer paid taxes, was accelerated by the Khitan invasion in the late 690s. Attempts to stop it were ineffectual.

By 705 the Empress, who was now 80, had allowed control of events to slip from her fingers. The bureaucratic faction at court, tired of the excesses of her latest favourites, forced her to abdicate in favour of Chung-tsung. The T'ang was restored.

Chung-tsung, however, also had a domineering wife, the empress Wei, who initiated a regime of utter corruption at court, openly selling offices. When the Emperor died in 710, probably poisoned by her, she tried to establish herself as ruler as Wu-hou had done before her. But Li Lung-chi, the future Hsüan-tsung, with the aid of Wu-hou's formidable daughter, T'ai-p'ing, and of the palace army, succeeded in restoring his father, Jui-tsung (the brother of Chung-tsung), to the throne. The Princess now attempted to dominate her brother, the Emperor, and there followed a struggle for power between her and the heir apparent. In 712 Jui-tsung ceded the throne to Hsüan-tsung but retained in his own hands control over the most crucial areas of government. A second coup, in 713, placed Hsüan-tsung completely in charge and resulted in Jui-tsung's retirement and in the princess T'ai-p'ing's suicide.

**Prosperity and progress.** Hsüan-tsung's reign (712–756) was the high point of the T'ang dynasty. It was an era of wealth and prosperity that saw institutional progress along with a flowering of the arts. Political life was at first dominated by the bureaucrats recruited through the examination system who had staffed the central government under Wu-hou. But a gradual revival of the power of the great aristocratic clans tended to polarize politics, a polarization that was sharpened by the Emperor's employment of a series of aristocratic specialists who reformed the empire's finances from 720 onward, often in the teeth of bureaucratic opposition.

Reign of Hsüan-tsung

After 720 a large-scale reregistration of the population produced a greatly increased number of taxpayers and restored state control over vast numbers of unregistered families. The new household and land taxes were expanded. In the 730s the canal system, which had been allowed to fall into neglect under Wu-hou and her successors, was repaired and reorganized so that the administration could transport large stocks of grain from the Yangtze region to the capital and to the armies on the northern frontiers. The south was at last financially integrated with the north. By the 740s the government had accumulated enormous reserves of grain and wealth. The tax and accounting systems were simplified, taxes and labour services reduced.

Some important institutional changes accompanied these reforms. The land registration, reorganization of transport, and coinage reform were administered by specially appointed commissions holding extraordinary powers, including the authority to recruit their own staff. These commissions were mostly headed by censors, and they and the censorate became centres of aristocratic power. The existence of these new offices reduced the influence of the regular ministries, enabling the emperor and his aristocratic advisers to circumvent the normal channels and procedures of administration.

Dominance of the aristocracy

After 736 the political dominance of the aristocracy was firmly reestablished. An aristocratic chief minister, Li Lin-fu, became a virtual dictator, his powers increasing as Hsüan-tsung in his later years withdrew from active affairs into the pleasures of palace life and the study of Taoism. In the latter part of his reign Hsüan-tsung, who had previously strictly circumscribed the power of the palace women to avoid a recurrence of the disasters of Wu-hou's time and had also excluded members of the royal family from politics, faced a series of succession plots. In 745 he fell deeply under the influence of a new favourite, the Imperial concubine Yang Kuei-fei. In 751–752 one of her relatives, Yang Kuo-chung, thanks to her influence with the Emperor, rapidly rose to rival Li Lin-fu for supreme power. After Li's death in 752 Yang Kuo-chung dominated the court. He had not, however, Li's great political ability nor his experience and skill in handling people.

**Military reorganization.** The most important new development in Hsüan-tsung's reign was the growth in the power of the military commanders. During Kao-tsung's reign the old militia system had proved inadequate for frontier defense and had been supplemented by the institution of permanent armies and garrison forces quartered in strategic areas on the frontiers. These armies were made up of long-service veterans, many of them non-Chinese cavalry troops, settled permanently in military colonies. Although these armies were adequate for small-scale operations, for a large-scale campaign an expeditionary army and a headquarters staff had to be specially organized and reinforcements sent in by the central government. This cumbersome system was totally unsuitable for dealing with the highly mobile nomadic horsemen on the northern frontiers.

Hostility of the Turks and Tibetans

At the beginning of Hsüan-tsung's reign the Turks threatened to become again a major power, rivaling China in Central Asia and along the borders. Mo-ch'o, the Turkish khan who had invaded Hopeh in the aftermath of the Khitan invasion in the time of Wu-hou and had attacked the Chinese northwest at the end of her reign, turned his attention northward. By 711 he controlled the steppe from the Chinese frontier to Transoxiana and appeared likely to develop a new unified Turkish empire. When he was murdered in 716 his flimsy empire collapsed. His successor, Bilge, tried to make peace with the Chinese in 718, but Hsüan-tsung preferred to try to destroy his power by an alliance with the southwestern Basmil Turks and with the Khitan in Manchuria. Bilge, however, crushed the Basmil and attacked Kansu in 720. Peaceful relations were established in 721–722. Bilge's death in 734 precipitated the end of Turkish power. A struggle among the various Turkish subject tribes followed, from which the Uighur emerged as victors. In 744 they established a powerful empire that was to remain the dominant force on China's northern border until 840. Unlike the Turks, however, the Uighur pursued a consistent policy of alliance with the T'ang. On several occasions Uighur aid, even though offered on harsh terms, saved the dynasty from disaster.

The Tibetans were the most dangerous foe during Hsüan-tsung's reign, invading the northwest annually from 714 on. In 727–729 the Chinese undertook large-scale warfare against them, and in 730 a settlement was concluded. But in the 730s fighting broke out again, and the Tibetans began to turn their attention to the T'ang territories in the Tarim Basin. Desultory fighting continued on the border

of Kansu until the end of Hsüan-tsung's reign. From 752 onward the Tibetans acquired a new ally in the Nan Chao state in Yunnan, which enabled them to exert a continuous threat along the entire western frontier.

The frontier commanders

In the face of these threats, Hsüan-tsung organized the northern and northwestern frontiers from Manchuria to Szechwan into a series of strategic commands or military provinces under military governors who were given command over all the forces in a large region. This system developed gradually and was formalized in 737 under Li Lin-fu. The frontier commanders controlled enormous numbers of troops: nearly 200,000 were stationed in the northwest and Central Asia and more than 100,000 in the northeast; there were well over 500,000 in all. The military governors soon began to exercise some functions of civil government. In the 740s a non-Chinese general of Sogdian and Turkish origin, An Lu-shan, became military governor first of one and finally of all three of the northeastern commands, with 160,000 troops under his orders. An Lu-shan had risen to power largely through the patronage of Li Lin-fu. When Li died, An became a rival of Yang Kuo-chung. As Yang Kuo-chung developed more and more of a personal stranglehold over the administration at the capital, An Lu-shan steadily built up his military forces in the northeast. The armed confrontation that followed (see below) nearly destroyed the dynasty.

During the 750s there was a steady reversal of T'ang military fortunes. In the far west the overextended Imperial armies had been defeated by the Arabs in 751 on the Talas River. In the southwest a campaign against the new state of Nan Chao in Yunnan had led to the almost total destruction of an army of 50,000 men. In the northeast the Chinese had lost their grip on the Manchuria–Korea border with the emergence of the new state of Parhae in place of Koguryŏ, and the Khitan and Hsi peoples in Manchuria constantly caused border problems. The Tibetans in the northwest were kept in check only by an enormously expensive military presence. The principal military forces were designed essentially for frontier defense.

Thus the end of Hsüan-tsung's reign found the state in a highly unstable condition, with central government dangerously dependent on a small group of men operating outside the regular institutional framework, and with an overwhelming preponderance of military power in the hands of potentially rebellious commanders on the frontiers, against whom the Emperor could put into the field only a token force of his own and the troops of those commanders who remained loyal.

## THE LATE T'ANG (755–907)

Rebellion ed by An Lu-shan

The rebellion of An Lu-shan in 755 marked the beginning of a new period. At first the rebellion had spectacular success. It swept through the northeastern province of Hopeh, captured the eastern capital, Lo-yang, early in 756, and took the main T'ang capital, Ch'ang-an, in July of the same year. The Emperor fled to Szechwan, and on the road his consort Yang Kuei-fei and other members of the Yang faction who had dominated his court were killed. Shortly afterward the heir apparent, who had retreated to Ling-wu in the northwest, himself usurped the throne. The new emperor, Su-tsung (reigned 756–762), was faced with a desperately difficult military situation. The rebel armies controlled the capital and most of Hopeh and Honan. In the last days of his reign Hsüan-tsung had divided the empire into five areas, each of which was to be the fief of one of the Imperial princes. Prince Yung, who was given control of the southeast, was the only one to take up his command; during 757 he attempted to set himself up as the independent ruler of the crucially important economic heart of the empire in the Huai and Yangtze valleys but was murdered by one of his generals.

An Lu-shan himself was murdered by a subordinate early in 757, but the rebellion was continued, first by his son and then by one of his son's generals, Shih Ssu-ming, and his son Shih Ch'ao-i. Not until 763 was it finally suppressed. The rebellion had caused great destruction and hardship, particularly in Honan. The final victory was made possible partly by the employment of Uighur mercenaries, whose insatiable demands remained a drain on the treasury well into the 770s; partly by the failure of the rebel leadership after the death of the able Shih Ssu-ming; and partly by the policy of clemency adopted toward the rebels after the decisive campaign in Honan in 762. The need for a speedy settlement was urged by the growing threat of the Tibetans in the northwest. The latter, allied with the Nan Chao kingdom in Yunnan, had exerted continual pressure on the western frontier and in 763 occupied the whole of modern Kansu. Late in 763 they actually took and looted the capital. They continued to occupy the Chinese northwest until well into the 9th century. Their occupation of Kansu signaled the end of Chinese power for almost a millennium in Sinkiang.

**Provincial separatism.** The post-rebellion settlement not only pardoned several of the most powerful rebel generals but also appointed them as Imperial governors in command of the areas they had surrendered. Hopeh was divided into four new provinces, each under surrendered rebels, while Shantung became the province of An Lu-shan's former garrison army from P'ing-lu in Manchuria, which had held an ambivalent position during the fighting.

Breakdown of the central machinery

Within these provinces the central government held little power. The leadership was decided within each province, and the central government in its appointments merely approved faits accomplis. Succession to the leadership was frequently hereditary. For all practical purposes, the northeastern provinces remained semi-independent throughout the later part of the T'ang era. They had been among the most populous and productive parts of the empire, and their semi-independence was not only a threat to the stability of the central government but also represented a huge loss of revenue and potential manpower.

Provincial separatism also became a problem elsewhere. With the general breakdown of the machinery of central administration after 756, many of the functions of government were delegated to local administrations. The whole empire was now divided into provinces (*tao*), which formed an upper tier of routine administration. Their governors had wide powers over subordinate prefectures and counties. The new provincial governments were of two main types. In northern China (apart from the semi-autonomous provinces of the northeast, which were a special category) most provincial governments were military, their institutions closely modeled on those set up on the northern frontier under Hsüan-tsung. The military presence was strongest in the small frontier-garrison provinces that protected the capital, Ch'ang-an, from the Tibetans in Kansu and in the belt of small, heavily garrisoned provinces in Honan that protected China—and the canal from the Huai and Yangtze valleys, on which the central government depended for its supplies—from the semiautonomous provinces. Military governments were also the rule in Szechwan, which continued to be menaced by the Tibetans and Nan Chao, and in the far south in Ling-nan. In central and southern China, however, the provincial government developed into a new organ of the civil bureaucracy. The civil governors of the southern provinces were regularly appointed from the bureaucracy, and it became customary to appoint to these posts high-ranking court officials who were temporarily out of favour.

All the new provinces had considerable latitude of action, particularly during the reigns of Su-tsung and Tai-tsung, when central power was at a low ebb. There was a general decentralization of authority. The new provinces had considerable independence in the fields of finance, local government, law and order, and military matters.

Reign of Tai-tsung

Under Tai-tsung (reigned 762–779/780) the court was dominated by the Emperor's favourite, Yüan Tsai, and by the eunuchs who now began to play an increasing role in T'ang politics. A succession of eunuch advisers not only rivaled in influence the chief ministers but even exerted influence over the military in the campaigns of the late 750s and early 760s. Under Tai-tsung many of the regular offices of the administration remained unfilled, while the irregularities encouraged by Yüan Tsai and his clique in the appointment of officials led to an increasing use of eunuchs in secretarial posts and to their increasing dominance over the Emperor's private treasury.

The central government did achieve some success in

finance. The old fiscal system with its taxes and labour services had been completely disrupted by the breakdown of authority and by the vast movements of population. The revenues became more and more dependent upon additional taxes levied on cultivated land or on property. Increasingly the government attempted to raise revenue from the urban population. But its survival depended upon the revenues it drew from central China, from the Huai Valley, and from the lower Yangtze. These revenues were sent to the capital by means of a reconstructed and improved canal system maintained out of the new government monopoly on salt. By 780 the salt monopoly was producing a major part of the state's central revenues, in addition to maintaining the transportation system. The salt and transportation administration was controlled by an independent commission centred in Yang-chou near the mouth of the Yangtze, and this commission gradually took over the entire financial administration of southern and central China.

Taxes and rebellion

The weak Tai-tsung was succeeded by a tough, intelligent activist emperor, Te-tsung (reigned 779/780–805), who was determined to restore the fortunes of the dynasty. He reconstituted much of the old central administration and decided on a showdown with the forces of local autonomy. As a first step, in 780 he promulgated a new system of taxation under which each province was assessed a quota of taxes, the collection of which was to be left to the provincial government. This was a radical measure, for it abandoned the traditional concept of head taxes levied at a uniform rate throughout the empire and also began the assessment of taxes in terms of money.

The semi-independent provinces of the northeast saw this as a threat to their independence, and, when it became apparent that Te-tsung was determined to carry out consistently tough policies toward the northeast, reducing their armies and even denying them the right to appoint their own governors, the Hopeh provinces rebelled. From 781 to 786 there was a wave of rebellions not only in the northeastern provinces but also in the Huai Valley and in the area of the capital itself. These brought the T'ang even closer to disaster than had the An Lu-shan rising. The situation was saved because at a crucial moment the rebels fell out among themselves and because the south remained loyal. In the end, the settlement negotiated with the governors of Hopeh virtually endorsed the preceding status quo, although the court made some marginal inroads with the establishment of two small new provinces in Hopeh.

After this disaster, Te-tsung pursued a very careful and passive policy toward the provinces. Governors were left in office for long periods, and hereditary succession continued. But the latter part of Te-tsung's reign was a period of steady achievement. The new tax system was gradually enforced and proved remarkably successful; it remained the basis of the tax structure until Ming times. Revenues increased steadily, and Te-tsung left behind him a wealthy state. Militarily, he was also generally successful. The Tibetan threat was contained. Nan Chao was won from its alliance with the Tibetans. The garrisons of the northwest were strengthened. At the same time, Te-tsung built up large new palace armies, giving the central government a powerful striking force—numbering 100,000 men by the end of his reign. Command was given to eunuchs considered loyal to the throne. The death of Te-tsung in 805 was followed by the very brief reign of Shun-tsung, an invalid monarch whose court was dominated by the clique of Wang Shu-wen and Wang P'ei. They planned to take control of the palace armies from the eunuchs but failed.

Reign of Hsien-tsung

**The struggle for central authority.** Under Hsien-tsung (reigned 805/806–820) the T'ang regained a great deal of its power. Hsien-tsung, a tough and ruthless ruler who kept a firm hand on affairs, is notable chiefly for his successful policies toward the provinces. Rebellions in Szechwan (806) and the Yangtze Delta (807) were quickly put down. After an abortive campaign (809–810) that was badly bungled by a favourite eunuch commander, the court was again forced to compromise with the governors of Hopeh. A fresh wave of trouble came in 814–817 with a rebellion in Huai-hsi, in the upper Huai valley, that threatened the canal route. This uprising was crushed and the province divided up among its neighbours. The P'ing-lu army in Shantung rebelled in 818 and suffered the same fate. Hsien-tsung thus restored the authority of the central government throughout most of the empire. His success was based largely upon the palace armies. The fact that these were controlled by eunuchs placed a great measure of power in the Emperor's hands. Under his weak successors, however, eunuchs' influence in politics proved a disaster.

Hsien-tsung's restoration of central authority involved more than military dominance. It was backed by a series of institutional measures designed to strengthen the power of the prefects and county magistrates, as against their provincial governors, by restoring to them the right of direct access to central government and giving them some measure of control over the military forces quartered within their jurisdiction. In an important financial reform, the provincial government no longer had first call on all the revenue of the province, as some revenue went directly to the capital. The government also began the policy, continued throughout the 9th and 10th centuries, of cutting down and fragmenting the provinces. It strengthened its control over the provincial administrations through a system of eunuch army supervisors, who were attached to the staff of each provincial governor. These eunuchs played an increasingly important role, not merely as sources of information and intelligence but as active agents of the emperor, able to intervene directly in local affairs.

The balance of power within the central government had also been considerably changed. The emperor Te-tsung had begun to delegate a great deal of business, in particular the drafting of edicts and legislation, to his personal secretariat, the Hanlin Academy. Although the members of the Hanlin Academy were handpicked members of the bureaucracy, their positions as academicians were outside the regular official establishment. This eventually placed the power of decision and the detailed formulation of policy in the hands of a group dependent entirely upon the emperor, thus threatening the authority of the regularly constituted ministers of the court. The influence of the eunuchs also began to be formalized and institutionalized in the palace council; this provided the emperor with another personal secretariat, which controlled the conduct of official business and had close links with the eunuchs' command of the powerful palace armies.

The eunuchs

The eunuchs' influence in politics steadily increased. Hsien-tsung was murdered by some of his eunuch attendants, and henceforth the chief eunuchs of the palace council and the palace armies were a factor in nearly every succession to the T'ang throne; in some cases they had their own candidates enthroned in defiance of the previous emperor's will. The emperor Wen-tsung (reigned 826/827–840) sought to destroy the dominance of the eunuchs; his abortive schemes only demoralized the bureaucracy, particularly after the Sweet Dew coup of 835, which misfired and led to the deaths of several ministers and a number of other officials. But the apogee of the eunuchs' power was brief, ending with the accession of Wu-tsung in 840. Wu-tsung and his minister, Li Te-yü, managed to impose some restrictions on the eunuchs' power, especially in the military.

In the second half of the 9th century the central government became progressively weaker. In I-tsung's reign (859–873) there was a resurgence of the eunuchs' power and constant fratricidal strife between eunuchs and officials at court. From the 830s onward the first signs of unrest and banditry had appeared in the Huai Valley and Honan. From 856 trouble spread to the Yangtze Valley and the south. There were major uprisings led by K'ang Ch'üan-t'ai in southern Anhwei in 858 and by Ch'iu Fu in Chekiang in 859. The situation was complicated by a costly war against the Nan Chao kingdom on the borders of the Chinese protectorate in Annam, which later spread to Szechwan and dragged on from 858 until 866. After the suppression of the invaders, part of the garrison force that had been sent to Ling-nan mutinied and, under its leader, Pang Hsün, fought and plundered its way back to Honan, where it caused widespread havoc in 868 and 869, cutting the canal linking the capital to the loyal Yangtze and Huai provinces. In 870 war broke out again with Nan Chao.

I-tsung was succeeded by Hsi-tsung, a boy of 11 who was the choice of the palace eunuchs. He reigned from 873 to 888. Honan had repeatedly suffered serious floods. In 874 following a terrible drought, a wave of peasant risings began. The most formidable of them was led by Huang Ch'ao, who in 878 marched south and sacked Canton and then marched to the north, where he took Lo-yang in late 880 and Ch'ang-an in 881. Although Huang Ch'ao attempted to set up a regime in the capital, he proved cruel and inept. Hemmed in by loyal armies and provincial generals, in 883 he was forced to abandon Ch'ang-an and withdraw to Honan and then to Shantung, where he died in 884. His forces were eventually defeated with the aid of Sha-t'o Turks, although the T'ang court was left virtually powerless, its emperor a puppet manipulated by rival military leaders. The dynasty lingered on until 907, but the last quarter century was dominated by the generals and provincial warlords. With the progressive decline of the central government in the 880s and 890s, China fell apart into a number of virtually independent kingdoms. Unity was not restored until long after the establishment of the Sung dynasty.

End of the T'ang dynasty

## CULTURAL DEVELOPMENTS

**The influence of Buddhism.** The T'ang emperors officially supported Taoism because of their claim to be descended from Lao-tzu, but Buddhism continued to enjoy great favour and lavish Imperial patronage through most of the period. The famous pilgrim Hsüan-tsang, who went to India in 629 and returned in 645, was the most learned of Chinese monks and introduced new standards of exactness in his many translations from Sanskrit. The most significant development in this time was the growth of new indigenous schools that adapted Buddhism to Chinese ways of thinking. Most prominent were the syncretistic T'ien-t'ai school, which sought to embrace all other schools in a single hierarchical system (even reaching out to include Confucianism), and the radically antitextual, antimetaphysical southern Ch'an (Zen) school, which had strong roots in Taoism. The popular preaching of the salvationist Pure Land sect was also important. After the rebellion of An Lu-shan, a nationalistic movement favouring Confucianism appeared, merging with the efforts of T'ien-t'ai Buddhism to graft Buddhist metaphysics onto Classical doctrine and lay the groundwork for the Neo-Confucianism of the Sung era.

In 843–845 the emperor Wu-tsung, a fanatical Taoist, decided to suppress Buddhism. One of his motives was economic. China was in a serious financial crisis, which Wu-tsung and his advisers hoped to solve by seizing the lands and wealth of the monasteries. The suppression was far-reaching: 40,000 shrines and temples—all but a select few—were closed; 260,000 monks and nuns were returned to lay life; vast acreages of monastic lands were confiscated and sold and their slaves manumitted. The suppression was short-lived, but irreparable damage was done to Buddhist institutions. Buddhism had already begun to lose intellectual momentum, and this attack upon it as a social institution marked the beginning of its decline in China.

The monastic communes

There were several types of monastic community. Official temples set up by the state had large endowments of land and property and large communities of monks who chose their own abbot and other officers. There were vast numbers of small village temples, shrines, and hermitages; these were often privately established, had little property, and were quite vulnerable to state policies. There were also private temples or "merit cloisters" established by great families, often in order that the family might donate its property and have it declared tax-exempt.

The monastic community was free of all obligations to the state. It was able to hold property without the process of division by inheritance that made the long-term preservation of individual and family fortunes almost impossible in T'ang times. It acquired its wealth from those taking monastic vows, from gifts of pious laymen, and from grants of lands by the state. The lands were worked by monastic slaves, by dependent families, by lay clerics who had taken partial vows but lived with their families, or by tenants. Monasteries also operated oil presses and mills. They were important credit institutions, supplying loans at interest and acting as pawnshops. They provided lodgings for travelers, operated hospitals and infirmaries, and maintained the aged. One of their most important social functions was primary education. The temples maintained their own schools, training the comparatively large proportion of the male population, which, although not educated to the standards of the Confucian elite or of the clergy, was nevertheless literate.

T'ang poets

**Trends in the arts.** In literature the greatest glory of the T'ang period was its poetry. By the 8th century poets had broken away from the artificial diction and matter of the court poetry of the southern dynasties and achieved a new directness and naturalism. The reign of Hsüan-tsung (712–756)—known as Ming-huang, the Brilliant Emperor—was the time of the great figures of Li Po, Wang Wei, and Tu Fu. The rebellion of An Lu-shan and Tu Fu's bitter experiences in it brought a new note of social awareness to his later poetry. This appears again in the work of Po Chü-i (772–846), who wrote verse in clear and simple language. Toward the end of the dynasty a new poetic form, the *tz'u,* in a less regular metre than the five-word and seven-word *lu-shih* and meant to be sung, made its appearance. The *ku-wen,* or "ancient style," movement grew up after the rebellion of An Lu-shan, seeking to replace the euphuistic *p'ien-wen* ("parallel prose") then dominant. It was closely associated with the movement for a Confucian revival. The most prominent figures in it were Han Yü and Liu Tsung-yüan. At the same time came the first serious attempts to write fiction, the so-called *ch'uan-ch'i,* or "tales of marvels." Many of these T'ang stories later provided themes for the Chinese drama.

Visual arts

The patronage of the T'ang emperors and the general wealth and prosperity of the period encouraged the development of the visual arts. Though few T'ang buildings remain standing, contemporary descriptions give some idea of the magnificence of T'ang palaces and religious edifices and the houses of the wealthy. Buddhist sculpture shows a greater naturalism than in the previous period, but there is some loss of spirituality. Few genuine originals survive to show the work of T'ang master painters such as Wu Tao-hsüan, who worked at Hsüan-tsung's court (712–756). As a landscape painter the poet Wang Wei was a forerunner of the *wen-jen,* or "literary man's," school of mystical nature painting of later times. The minor arts of T'ang, including ceramics, metalwork, and textiles, give expression to the colour and vitality of the life of the period. Printing appeared for the first time during T'ang. Apparently invented to multiply Buddhist scriptures, it was used by the end of the dynasty for such things as calendars, almanacs, and dictionaries.

## SOCIAL CHANGE

**Decline of the aristocracy.** The late T'ang period saw the beginnings of social changes that did not reach their culmination until the 11th century. The most important of these was the change in the nature of the ruling class. Although from the early T'ang era the examination system had facilitated the recruitment into the higher ranks of the bureaucracy of persons from lesser aristocratic families, most officials continued to come from the established elite. Social mobility increased after the An Lu-shan rebellion: provincial governments emerged, their staffs in many cases recruited from soldiers of very lowly social origins, and specialized finance commissions were established, a large part of their personnel often recruited from the commercial community. The contending factions of the 9th-century court also employed irregular appointments to secure posts for their clients and supporters, many of whom came from comparatively lowly backgrounds.

Although the old aristocracy retained a grip on political power until very late in the dynasty, its exclusiveness and hierarchical pretensions were rapidly breaking down. It was finally extinguished as a separate group in the Wu-tai (Five Dynasties) period (907–960), when the old strongholds of aristocracy in the northeast and northwest became centres of bitter military and political struggles. The aristocratic clans that survived did so by merging into the new official-literati class; this class was based not

on birth alone but on education, office holding, and the possession of landed property.

At the same time there was a return to semiservile relationships at the base of the social pyramid. Sheer economic necessity led many peasants either to dispose of their lands and become tenants or hired labourers of rich neighbours or to become dependents of a powerful patron. Tenancy, which in early T'ang times had most often been a temporary and purely economic agreement, now developed into a semipermanent contract requiring some degree of personal subordination from the tenant.

The new provincial officials and local elites were able to establish their fortunes as local landowning gentry largely because after 763 the government ceased to enforce the system of state-supervised land allocation. In the aftermath of the An Lu-shan and later rebellions, large areas of land were abandoned by their cultivators; other areas of farmland were sold off on the dissolution of the monastic foundations in 843–846. The landed estate managed by a bailiff and cultivated by tenants, hired hands, or slaves became a widespread feature of rural life. Possession of such estates, previously limited to the established families of the aristocracy and the serving officials, now became common at less exalted levels.

**Population movements.** Censuses taken during the Sui and T'ang dynasties provide some evidence as to population changes. Surviving figures for 609 and for 742, representing two of the most complete of the earlier Chinese population registrations, give totals of some 9,000,000 households, or slightly more than 50,000,000 persons. Contemporary officials considered that only about 70 percent of the population was actually registered, so that the total population may have been as much as 70,000,000.

Between 609 and 742 a considerable redistribution of population took place. The population of Hopeh and Honan fell by almost a third because of the destruction suffered at the end of the Sui era and in the invasions of the 690s and because of epidemics and natural disasters. The population of Ho-tung (modern Shansi) and of Kuan-chung and Lung-yu (modern Shensi and Kansu) also fell, though not so dramatically. The population of the south, particularly the southeastern region around the lower Yangtze, took a leap upward, as did that of Szechwan.

Whereas under the Sui the population of the Great Plain (Hopeh and Honan) had accounted for more than half of the empire's total, by 742 this had dropped to 37 percent. The Huai–Yangtze area, which had contained only about 8 percent of the total in 609, now contained one-quarter of the entire population, and Szechwan's share jumped from 4 percent to 10 percent of the total, exceeding the population of the metropolitan province of Kuan-chung. The increase in the south was almost entirely concentrated in the lower Yangtze Valley and Delta and in Chekiang.

The southward migration

Although there are no reliable population figures from the late T'ang era, the general movement of population toward the south certainly continued; there were considerable increases in the population of the area south of the Yangtze, in modern Kiangsi and Hunan, and in Hupeh. The chaos of the last decades of the T'ang dynasty completed the ruin of the northwest. After the destruction of the city of Ch'ang-an in the Huang Ch'ao rebellion, no regime ever again established its capital in that region.

**Growth of the economy.** The 8th and 9th centuries were a period of growth and prosperity. The gradual movement of the population away from the north, with its harsh climate and dry farming, into the more fertile and productive south meant a great proportional increase in productivity. The south still had large areas of virgin land. Fukien, for example, was still only marginally settled along the coastline at the end of T'ang times. During the latter half of the T'ang, the Huai and lower Yangtze became a grain-surplus area, replacing Hopeh and Honan. From 763 to the mid-9th century, great quantities of grain were shipped from the south annually as tax revenue. New crops, such as sugar and tea, were grown widely. The productivity of the Yangtze Valley was increased by double-cropping land with rice and winter wheat and by developing new varieties of grain. Whereas in early T'ang times the chief silk-producing areas had been in the northeast, after the An Lu-shan rebellion silk production began to increase rapidly in Szechwan and the Yangtze Delta region.

Expansion of the grain crop

A boom in trade soon followed. The merchant class threw off its traditional legal restraints. In early T'ang times there had been only two great metropolitan markets, in Ch'ang-an and Lo-yang. Now every provincial capital became the centre of a large consumer population of officials and military, and the provincial courts provided a market for both staple foodstuffs and luxury manufactures. The diversification of markets was still more striking in the countryside. A network of small rural market towns, purely economic in function and acting as feeders to the county markets, grew up. At these periodic markets, held at regular intervals every few days, traveling merchants and peddlers dealt in the everyday needs of the rural population. By the end of the T'ang period these rural market centres had begun to form a new sort of urban centre, intermediate between the county town, with its administrative presence and its central market, and the villages.

The growth of trade brought an increasing use of money. In early T'ang times silk cloth had been commonly employed as money in large transactions. When the central government lost control of the major silk-producing region in Hopeh and Honan, silk was replaced in this use by silver. The government neither controlled silver production nor minted a silver coinage. Silver circulation and assay were in the hands of private individuals. Various credit and banking institutions began to emerge: silversmiths took money on deposit and arranged for transfers of funds; a complex system of credit transfers arose by which tea merchants would pay the tax quota for a district, sometimes even for a whole province, out of their profits from the sale of the crop at Ch'ang-an and receive reimbursement in their home province.

The increasing use of money and of silver also affected official finance and accounting. Taxes began to be assessed in money. The salt monopoly was collected and accounted for entirely in money. The government also began to look to trade as a source of revenue—to depend increasingly on taxes on commercial transactions, levies on merchants, transit taxes on merchandise, and sales taxes.

The merchants

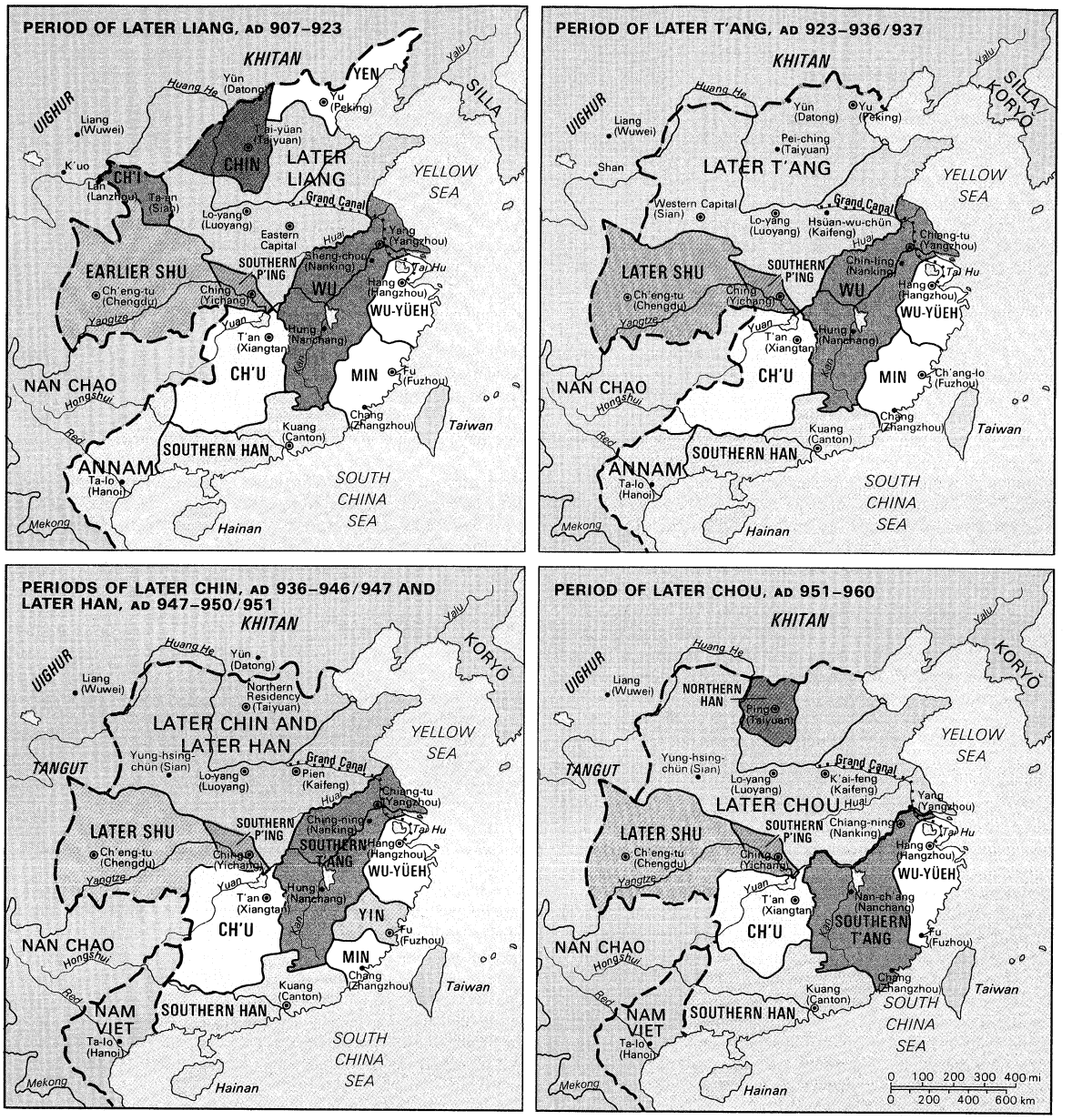
The most prosperous of the merchants were the great dealers in salt, the tea merchants from Kiangsi, the bankers of the great cities and particularly of Ch'ang-an, and the merchants engaged in overseas trade in the coastal ports. Foreign trade was still dominated by non-Chinese merchants. Yang-chou and Canton had large, Arab trading communities. The northern coastal traffic was dominated by the Koreans. Overland trade to Central Asia was mostly in the hands of Sogdian and later of Uighur merchants. Central Asian, Sogdian, and Persian merchants and peddlers carried on much local retail trade and provided restaurants, wine shops, and brothels in the great cities. Only in the 9th century did the foreign influence in trade begin to recede.

In the late T'ang many officials began to invest their money (and official funds entrusted to them) in commercial activities. High officials took to running oil presses and flour mills, dealing in real estate, and providing capital for merchants. The wall between the ruling class and the merchants that had existed since the Han period was rapidly breaking down in the 9th century, and the growth of urbanization, which characterized the Sung period (960–1279), had already begun on a wide scale. (D.C.T.)

## The Five Dynasties and the Ten Kingdoms

The period of political disunity between the T'ang and the Sung lasted little more than half a century, from 907 to 960. During this brief era, when China was truly a multistate system, five short-lived regimes succeeded one another in control of the old Imperial heartland in northern China, hence the name Wu-tai (Five Dynasties). During these same years, 10 relatively stable regimes occupied sections of southern and western China, so the period is also referred to as that of the Shih-kuo (Ten Kingdoms).

Most of the major developments of this period were extensions of changes already under way during the late T'ang, and many were not completed until after the




China during the Five Dynasties and Ten Kingdoms period *c.* AD 907–960.
Adapted from A. Herrmann, *An Historical Atlas of China* (1966), Aldine Publishing Company

founding of the Sung dynasty. For example, the process of political disintegration had begun long before Chu Wen brought the T'ang dynasty to a formal end in 907. The developments that eventually led to reunification, the rapid economic and commercial growth of the period, and the decline of the aristocratic clans had also begun long before the first Sung ruler, T'ai-tsu, reconquered most of the empire, and they continued during the reigns of his successors on the Sung throne.

### THE WU-TAI (FIVE DYNASTIES)

None of the Wu-tai regimes that dominated North China ever forgot the ideal of the unified empire. Each sought, with gradually increasing success, to strengthen the power of the central authorities. Even Chu Wen, who began the Wu-tai by deposing the last T'ang emperor in 907, sought to extend his control in the north. While consolidating his strength on the strategic plains along the Huang Ho (Yellow River) and connecting them with the vital transportation system of the Grand Canal, he made the significant choice of locating his base at Pien (modern K'ai-feng, in Honan); it later became the Pei (Northern) Sung capital. Pien's lack of historical prestige was balanced by the presence of the ancient capital, Lo-yang, a short distance to the west, which was still the nation's cultural centre. Chu Wen's short-lived Hou (Later) Liang dynasty, founded in 907, was superseded by the Hou T'ang in 923, by the Hou Chin in 936, by the Hou Han in 947, and by the Hou Chou in 951. These rapid successions of dynasties came to an end only with the rise in 960 of the Sung dynasty, which finally succeeded in establishing another lasting empire and in taking over much, though not all, of the former T'ang Empire.

Changes in the political process

Beneath the surface, however, were the continuous efforts of a reintegrative political process that heralded the coming of a new empire and helped to shape its political system. In this respect the successive rulers moved like a relay team along the tortuous road back to unification. These militarists expanded their personal power by recruiting peoples of relatively humble social origins to replace the aristocrats. Such recruits owed personal allegiance to their masters, on whose favours their political positions remained dependent, thus presaging the rise of absolutism.

Rather than being discarded, the T'ang administrative form underwent expedient alterations so that the new types of officials, promoted because of merit from regional posts to palace positions, could use the military administration to supervise the nearby provinces and gradually bring

them under direct control. Top priority went to securing fiscal resources from the salt monopoly, tribute transport, and in particular new tax revenues, without which military domination would have been hard to sustain and political expansion impossible. Eventually, a pattern of centralizing authority emerged. Fiscal and supply officials of the successive regimes went out to supervise provincial finances and the local administration. The minor militarists, heretofore the local governors in control of their own areas, were under double pressure to submit to reintegrative measures. They faced the inducement of political accommodation, which allowed them to keep their residual power, and the military threat of palace army units commanded by special commissioners, which were sent on patrol duty into their areas. The way was thus paved, in spite of occasional detours and temporary setbacks, for the ultimate unification.

The seemingly chaotic period was in fact less chaotic than other rebellious times, except from the standpoint of the aristocrats, who lost their preeminent status along with their large estates, which were usually taken over piecemeal by their former managers. The aristocratic era in Chinese history was gone forever; a new bureaucratic era was about to begin.

#### THE SHIH-KUO (TEN KINGDOMS)

From the time of the T'ang dynasty until the Ch'ing dynasty, which arose in the 17th century, China consisted of two parts: the north, militarily strong, and the south, economically and culturally wealthy. Between 907 and 960, 10 independent kingdoms emerged in China, mainly in the south: the Wu (902–937), the Nan (Southern) T'ang (937–975/976), the Nan P'ing (924/925–963), the Ch'u (927–951), the Ch'ien (Former) Shu (907–925), the Hou (Later) Shu (934–965), the Min (909–945/946), the Pei (Northern) Han (951–979), the Nan Han (917–971), and the Wu-Yüeh (907–978), the last located in China's most rapidly advancing area—in and near the lower Yangtze Delta.

Regional developments

Some of these separate regimes achieved relative internal stability; none attained enough strength to strive to unify China. Nonetheless, the regional developments in South China, in the upper Yangtze region in southwest China, and in the lower Yangtze region in southeast China were of great interest. In South China, the Min kingdom in modern Fukien and the Nan Han in modern Kwangtung and Kwangsi reflected sharp cultural differences. Along the coast, sea trade expanded, promoting both urban prosperity and cultural diversity. On land, wave after wave of refugees moved southward, settling along rivers and streams and in confining plains and mountain valleys, using a frontier agriculture but with highly developed irrigation and land reclamation. Usually they pushed aside the aboriginal minorities, earlier settlers, and previous immigrant groups. This process turned South China into a cultural chessboard of great complexity, with various subcultural pieces sandwiched between one another. Many eventually evolved along different lines.

In southwest China the valley of modern Szechwan presented an interestingly different picture of continuous growth. Usually protected from outside disturbances and invasions by the surrounding mountains, it enjoyed peace and prosperity except for one decade of instability between the Ch'ien Shu and Hou Shu. The beautiful landscape inspired poets who infused a refreshing vitality into old-style poetry and essays. In this region, a stronghold of Taoist religion, the people inserted into Confucian scholarship an admixture of Taoist philosophy. Buddhism also flourished. These intellectual trends in Szechwan foreshadowed an eclectic synthesis of the three major teachings—Confucianism, Taoism, and Buddhism.

The Buddhist monasteries owned large estates and were usually among the first to introduce new and better technology. Growing commerce created a demand for money. The ensuing shortage of copper was met by an increasing output of iron through more efficient methods and an elementary division of labour in production. When the limited number of copper coins could no longer meet the growing volume of trade, iron currency briefly went into circulation. Increasing commerce also resulted in the development of various paper credit instruments, the best known being drafts for transmitting funds called *fei-ch'ien* ("flying money"). Somewhat later the private assay shops in Szechwan began to issue certificates of deposit to merchants who had left valuables at the shops for safekeeping. These instruments, which began to circulate, were the direct ancestors of the paper money that emerged in the early 11th century.

During the Wu-tai printing became common. The most famous, and monumental, cultural production of the period was the editing and printing of the Confucian Classics and the Buddhist *Tipiṭaka,* but the period also saw the rise of a printing industry that produced works for private buyers. The best printing in the country during the Wu-tai and the Sung dynasty came from the regions of Szechwan and Fukien.

Leadership of the Yangtze Delta

From the Wu-tai onward southeast China, especially its core region of the Yangtze Delta, began to lead the country in both economic prosperity and cultural refinement. In this region fertile soil, irrigation networks, and highly selective crops combined to create the best model of intensive farming. Interlocking streams, rivers, and lakes fed an ever-increasing number of markets, market towns, cities, and metropolitan areas, where many farm products were processed into an ever-expanding variety of consumer goods. Such development enhanced regional trade, stimulated other regions to adopt specialization, and promoted overseas commerce.

The Sung conquerors from the north recognized the high level of cultural development in this region. After the surrender of the last Nan T'ang ruler, himself a renowned poet, the unexcelled royal library in the capital at Nanking was moved to the north; along with it went many officials who were skilled in art, literature, and bibliography. The surrender of the Wu-Yüeh kingdom, slightly farther south, followed the same pattern. Moreover, refined culture developed away from the coast in such inland mountainous areas as modern Kiangsi, which shortly thereafter produced internationally coveted porcelain and where many great artists and scholar-officials attained positions of cultural leadership. Thereafter, southeast China retained its cultural excellence. At the end of the Pei Sung period the Nan Sung based itself in the lower Yangtze Delta and located its capital at Hang-chou, the former capital of the Wu-Yüeh.

As traditional histories stress, this period of disunity definitely had its dark side: militarism, wars, disintegration of the old order, and an inevitable lowering of moral standards. The dark side, however, stemmed largely from underlying changes that were transforming China into a new pattern that would last for nearly a millennium.

## The "barbarians": Tangut, Khitan, and Juchen

On the frontier the far-reaching influence of T'ang culture affected various nomadic, seminomadic, and pastoral peoples.

#### THE TANGUT

In the northwest the Tangut, a Tibetan people, inhabited the region between the far end of the Great Wall in modern Kansu and the Huang Ho bend in Inner Mongolia. Their semioasis economy combined irrigated agriculture with pastoralism. Their control over the terminus of the famous Silk Road made them middlemen in trade between Central Asia and China. They adopted Buddhism as a state religion, in government and education followed the T'ang model, and devised a written script for their own language. This richly mixed culture blossomed, as evidenced by the storing at the Tun-huang caves of an unparalleled collection of more than 30,000 religious paintings, manuscripts, and books in Chinese, Tibetan, Uighur, and other languages. In 1038 the Tangut proclaimed their own kingdom of Hsi Hsia, which survived for nearly two centuries with remarkable stability despite a series of on-and-off border clashes with the neighbouring states in North China. The kingdom's end came with the Mongols, the first nomads to conquer all of China.

THE KHITAN

To the north at the time of the Wu-tai rose the seminomadic but largely pastoral Khitan, who were related to the eastern Mongols. The word Khitan (or Khitai) is the source of Cathay, the name for North China in medieval Europe (as reported by Marco Polo), and of Kitai, the Russian name for China. The Khitan founded the Liao Empire (907–1125) by expanding from the border of Mongolia into both southern Manchuria and the 16 prefectures below the Great Wall. This area south of the line of the Great Wall was to remain out of Chinese political control for more than 400 years. Its control by a non-Chinese state posed a dangerous security problem for the Pei Sung. More importantly in the long run, this region acted for centuries as a centre for the mutual exchange of culture between the Chinese and the northern peoples.

The Liao made Yen-ching (modern Peking) their southern capital, thus starting that city's history as a capital, and claimed to be the legitimate successors to the T'ang. They incorporated their own tribes under respective chieftains and, with other subdued tribes in the area, formed a confederation, which they transformed into a hereditary monarchy. Leadership always remained in the hands of the ruling tribe, the Yeh-lü, who for the sake of stability shifted to the Chinese clan system of orderly succession.

The Liao economy was based on horse and sheep raising and on agriculture. Millet was the main crop, and salt, controlled by government monopoly, was an important source of revenue. There were also such other riches as iron produced by smelters. The Liao employed an effective dual system of administration to guard against the danger of being absorbed by Sinicization. They had one administration for their own people that enforced tribal laws, maintained traditional rites, and largely retained the steppe style of food and clothing. The Liao deliberately avoided the use of Chinese and added to their particular branch of the Mongolian language two types of writing—a smaller one that was alphabetical and a larger one related to Chinese characters. A second administration governed the farming region by the old T'ang system, using T'ang official titles, an examination system, Chinese-style tax regulations, and the Chinese language. The laws of the second administration enforced the established way of life, including such practices as ancestral worship among the Chinese subjects. The status of Chinese subjects varied. Some were free subjects who might move upward into the civil service; others might be held in bondage and slavery.

Dual system of administration

Though honouring the Confucian philosophy, the Liao rulers patronized Chinese Buddhism. Their achievements were generally military and administrative rather than cultural, but they did provide a model for their successors, the Chin, who in turn influenced the Mongols and, through them, succeeding Chinese dynasties.

THE JUCHEN

The Liao were eventually overthrown by the Juchen (Ju-chen in Wade–Giles, Ruzhen in Pinyin), another seminomadic and semipastoral people who originated in Manchuria, swept across North China, ended the Pei Sung, and established the Chin dynasty (1115–1234). This new and much larger empire in North China followed the Liao pattern of dual government and of some acculturation but at a much higher cultural level.

The Juchen, in establishing their Chinese-style Chin Empire, occupied a broader geographic region in the farming country than had any previous nomadic or pastoral conquerors. The migration of their own people in large numbers notwithstanding, they were proportionally a smaller minority than were the Khitan, for the Chin ruled a much larger Chinese population. Because they formed a small minority in their own empire, their tribesmen were kept in a standing army that was always prepared for warfare. Though quartered among their farming subjects, they were expected to respond to the command of their captains at short notice. In the military service the Juchen language was kept alive, and no Chinese-style names, clothing, or customs were permitted. They realized that protecting their separate ethnic and cultural identity was indispensable to maintaining military superiority.

Politically, however, it was necessary for the Juchen rulers to familiarize themselves with the higher culture of their Chinese subjects in order to manage state affairs. While limiting Chinese participation in the government, they shrewdly deflected the interests of their subjects toward the pursuit of such peaceful arts as printing, scholarship, painting, literature, and, significantly, the development of drama for widespread entertainment. (These trends continued under the Mongols and enriched the Chinese culture.) In spite of the Juchen efforts, time was on the side of the majority culture, which gradually absorbed the minority. The transplanted tribesmen, after settling on farmland, could not avoid being affected by the Chinese way of life, particularly during long periods of peace.

Juchen assimilation

Economically, the Juchen were no match for the Chinese. In time a number of Juchen became tenants on Chinese-owned land; some were reduced to paupers. Their economic decline altered social relations. Eventually they were permitted intermarriage, usually with parties wealthier than themselves. Their military strength also declined. It became normal for military units to be undermanned. Captains of "hundreds" often could put no more than 25 men into the field, and captains of "thousands" had no more than four or five such nominal "hundreds" under them. Their ruling class followed a parallel decline. The interests of the ruling group shifted from government affairs to Confucian studies, Chinese Classics, and T'ang- and Sung-style poetry. The rulers found little use for the two styles of Juchen script that their ancestors had devised.

Eventually the Juchen, much weakened, were brought down by the Mongols, led by Genghis Khan and his successors. (See below *The Yüan, or Mongol, dynasty: The Mongol conquest of China.*)

## The Sung dynasty

PEI (NORTHERN) SUNG (960–1127)

The Pei Sung (also known simply as the Sung) was the last major Chinese dynasty to be founded by a coup d'état. Its founder, Chao K'uang-yin (known by his temple name, T'ai-tsu), the commander of the capital area of K'ai-feng and inspector general of the Imperial forces, usurped the throne from the Hou Chou, the last of the Wu-tai.

**Unification.** Though a militarist himself, T'ai-tsu ended militarism as well as usurpation. Even his own coup was skillfully disguised to make it appear that the popular acclaim of the rank and file left him with no choice. Masterful in political maneuvering, T'ai-tsu, as emperor (reigned 960–976), did not destroy other powerful generals as had many previous founding rulers. Instead, he persuaded them to give up their commands in exchange for honorary titles, sinecure offices, and generous pensions—an unheard-of arrangement in Chinese history. The Sung founder and his successors reduced the military power of the generals and used a variety of techniques to keep them weak, but Sung rulers continued to support their social importance by frequently marrying members of the Imperial clan to members of leading military families.

End of militarism and of usurpation

With a shrewd appreciation of the war-weariness among the population, T'ai-tsu stressed the Confucian spirit of humane administration and the reunification of the whole country. To implement this policy, he took power from the military governors, consolidated it at court, and delegated the supervision of military affairs to able civilians; but no official was regarded as above suspicion. A pragmatic civil service system evolved, with a flexible distribution of power and elaborate checks and balances. Each official had a titular office, indicating his rank but not his actual function, a commission for his normal duties, and additional assignments or honours. This seemingly confusing formula enabled the ruler to remove an official to a lower position without demotion of rank, to give an official a promotion in rank but an insignificant assignment, or to pick up a low-ranking talent and test him on a crucial commission. Councillors controlled only the civil administration because the division of authority made the military commissioner and the finance commissioner separate entities, reporting directly to the ruler, who coordinated all important decisions. In decision making, the Emperor

received additional advice from academicians and other advisers—collectively known as opinion officials—whose function was to provide separate channels of information and to check up on the administrative branches.

Similar checks and balances existed in the diffused network of regional officials. The empire was divided into circuits, which were units of supervision rather than administration. Within these circuits officials, or intendants, were charged with overseeing the civil administration. Below these intendants were the actual administrators. These included prefects, whose positions were divided into several grades according to an area's size and importance. Below the prefects there were district magistrates (also called subprefects), in charge of areas corresponding roughly in size to counties. The duties of these subprefects were catholic, for they were supposed to see to all aspects of the welfare of the people in their area. This was the lowest level of major direct Imperial rule (though there were some petty officials on levels below the district). Because the members of the formal civil service level of the government were so few, actual administration in the yamen, or administrative headquarters, depended heavily on the clerical staff. Beyond the yamen walls control was in the hands of an officially sanctioned but locally staffed sub-bureaucracy.

Following Confucian ideals, the founder of the Sung dynasty lived modestly, listened to his ministers, and curbed excessive taxation. The rising prestige of his regime preceded his conquests. He also absorbed the best military units under his own command and disciplined them in the same Confucian style. His superior force notwithstanding, he embarked upon a reunification program by mixing war with lenient diplomatic or accommodative terms that assured defeated rivals of generous treatment. A well-planned strategy first took Szechwan in the southwest in 965, the extreme south in 971, and the most prosperous lower Yangtze area in the southeast one year before his death, making the reunification nearly complete. The Wu-Yüeh, the sole survivor among the Shih-kuo (Ten Kingdoms), chose to surrender without a war in 978.

The sudden death of the founder of the Sung dynasty left a speculative legend of assassination, though it was probably caused by his heavy drinking. The legend stemmed from the fact that his young son was denied the orderly succession. Instead, the Emperor's younger brother, who had acquired much experience at his side, seized the throne. With reunification accomplished in the south, the new emperor, T'ai-tsung (reigned 976–997/998), turned northward to fight the Khitan Empire, only to suffer a disastrous defeat (986). A relative shortage of horses and grazing grounds to breed them, in contrast to the strong Khitan cavalries, was not the only reason for the defeat. It also resulted from a deliberate policy of removing generals from their armies, subordinating officers to civilians, concentrating strength in Imperial units, and converting most provincial armies into labour batallions.

Defeat by the Khitan Empire

The Sung never achieved a military prowess comparable to that of the Han or the T'ang. Despite the occasional bellicosity of its officials, the Sung government failed to penetrate Indochina or to break the power of the Hsi Hsia of Tibet. As a result, Sung China became increasingly isolated, especially from Central Asia, whence much cultural stimulus had come under preceding dynasties. Combined with a natural pride in internal advancements, China's cultural ethnocentrism deepened.

**Consolidation.** Under the third emperor, Chen-tsung (reigned 997/998–1022/23), the Sung achieved consolidation. A threatening Khitan offensive was directly met by the Emperor himself. A few battles assured neither side of victory. The two empires pledged peaceful coexistence in 1004 through an exchange of sworn documents that foreshadowed modern international treaties. The Khitan gave up its claim to a disputed area it had once occupied below the Great Wall, and the Sung agreed to a yearly tribute: 100,000 units (a rough equivalent of ounces) of silver and 200,000 rolls of silk. It was a modest price for the Sung to pay for securing the frontier.

The Emperor thereafter sought to strengthen his absolutist image by claiming a Taoist charisma. Prompted by magicians and ingratiating high officials, he proclaimed that he had received a sacred document directly from Heaven. He ordered a grand celebration with elaborate rites, accompanied by reconstructed music of ancient times; and he made a tour to offer sacrifices at Mt. T'ai, following a Han dynasty precedent.

After the Emperor's death, friction arose between his widow, the empress dowager, who was acting as regent, and Jen-tsung (reigned 1022/23–1063/64), his teenage son by a palace lady of humble rank. Following the death of the Empress Dowager, the Emperor divorced his empress, who had been chosen for him by and had remained in sympathy with the Empress Dowager. The divorce was unjustifiable in Confucian morality and damaged the Imperial image.

By this time the bureaucracy was more highly developed and sophisticated than that of the early Sung. Well-regulated civil service examinations brought new groups of excellent scholar-officials who, though a numerical minority, dominated the higher policy-making levels of government. The sponsorship system, which discouraged favouritism by putting responsibility upon the sponsors for the official conduct of their appointees, also ensured deserving promotions and carefully chosen appointments. Many first-rate officials—especially those from the south whose families had no previous bureaucratic background—upheld Confucian ideals. These new officials were critical not only of palace impropriety but also of bureaucratic malpractices, administrative sluggishness, fiscal abuses, and socioeconomic inequities. Respecting absolutism, they focused their attacks upon a veteran chief councillor, whom the Emperor had trusted for years. Factionalism developed because many established scholar-officials, mostly from the north, with long bureaucratic family backgrounds, stood by their leader, the same chief councillor.

Growth of the bureaucracy

A series of crises proved the complaints of the idealists justified. After half a century of complacency, peace and prosperity began to erode. This became apparent in the occurrence of small-scale rebellions near the capital itself; in the disturbing inability of local governors to restore order themselves; and in a dangerous penetration of the northwestern border by Hsi Hsia, which rejected its vassal status and declared itself an independent kingdom. The Khitan took advantage of the changing military balance by threatening another invasion. The idealistic faction, put into power under these critical circumstances in 1043–44, effectively stopped the Hsi Hsia on the frontier by reinforcing a chain of defense posts and made it pay due respect to the Sung as the superior empire (though the Sung no longer claimed suzerainty). Meanwhile, the peace with Khitan was reensured by an increase in yearly tribute.

The court also instituted administrative reforms, stressing the need for emphasizing statecraft problems in civil service examinations, eliminating patronage appointments for family members and relatives of high officials, and enforcing strict evaluation of administrative performance. It also advocated reduction of compulsory labour, land reclamation and irrigation construction, organization of local militias, and a thorough revision of codes and regulations. Though mild in nature, the reforms hurt vested interests. Shrewd opponents undermined the reformers by misleading the Emperor into suspecting that they had received too much power and were disrespectful of him personally. With the crises eased, the Emperor found one excuse after another to send most reformers away from court. The more conventionally minded officials were returned to power.

Despite a surface of seeming stability, the administrative machinery once again fell victim to creeping deterioration. When some reformers eventually returned to court, beginning in the 1050s, their idealism was modified by the political lesson they had learned. Eschewing policy changes and tolerating colleagues of varying opinions, they made appreciable progress by concentrating upon the choice of better personnel, proper direction, and careful implementation within the conventional system; but many fundamental problems remained unsolved. Mounting military expenditures did not bring greater effectiveness; an expanding and more costly bureaucracy could not reverse

Creeping deterioration of the administration

the trend of declining tax yields. Income no longer covered expenditures. During the brief reign of Ying-tsung (1063/64–1067/68), relatively minor disputes and symbolically important issues concerning ceremonial matters embroiled the bureaucracy in mutual and bitter criticism.

**Reforms.** Shen-tsung (reigned 1067/68–1085/86) was a reform emperor. Originally a prince reared outside the palace, familiar with social conditions and devoted to serious studies, he did not come into the line of Imperial succession until adoption had put his father on the throne before him. Shen-tsung responded vigorously (and rather unexpectedly, from the standpoint of many bureaucrats) to the problems troubling the established order, some of which were approaching crisis proportions. Keeping above partisan politics, he made the scholar-poet Wang An-shih his chief councillor and gave him full backing to make sweeping reforms. Known as the new laws, or new policies, these reform measures attempted drastic institutional changes. In sum, they sought administrative effectiveness, fiscal surplus, and military strength. Wang's famous "Ten Thousand Word Memorial" outlined the philosophy of the reforms. Contrary to conventional Confucian views, it upheld assertive governmental roles; but its ideal remained basically Confucian—economic prosperity would provide the social environment essential to moral well-being.

Economic gains from the reforms

Never before had the government undertaken so many economic activities. The Emperor empowered Wang to institute a top-level office for fiscal planning, which supervised the Commission of Finance, previously beyond the jurisdiction of the chief councillor. The government squarely faced the reality of a rapidly spreading money economy by increasing the supply of currency. The state became involved in trading, buying specific products of one area for resale elsewhere (thereby facilitating the exchange of goods), stabilizing prices whenever and wherever necessary, and making a profit itself. This did not displace private trading activities. On the contrary, the government extended loans to small urban and regional traders through state pawnshops—a practice somewhat like modern government banking but unheard-of at the time. Far more important, if not controversial, the government made loans at the interest rate, low for the period, of 20 percent to the whole peasantry during the sowing season, thus assuring their farming productivity and undercutting their dependency upon usurious loans from the well-to-do. The government also maintained granaries in various cities to ensure adequate supplies on hand in case of emergency need. The burden upon rich and poor alike was made more equitable by a graduated tax scale based upon a reassessment of the size and the productivity of the landholdings. Similarly, compulsory labour was converted to a system of graduated tax payments, which were used to finance a hired-labour service program that at least theoretically controlled underemployment in farming areas. Requisition of various supplies from guilds was also replaced by cash assessments, with which the government was to buy what it needed at a fair price.

Wang's reforms achieved increased military power as well. To remedy the Sung's military weakness and to reduce the immense cost of a standing professional army, the villages were given the duty of organizing militias, under the old name of *pao-chia,* to maintain local order in peacetime and to serve as army reserves in wartime. To reinforce the cavalry, the government procured horses and assigned them to peasant households in northern and northwestern areas, in consideration for which one member of the family had the privilege of serving in the army with his horse. Various weapons were also developed. As a result of these efforts, the empire eventually scored some minor victories along the northwestern border.

The gigantic reform program required an energetic bureaucracy, which Wang attempted to create—with mixed results—by means of a variety of policies: promotion of a nationwide state school system; establishment or expansion of specialized training in such utilitarian professions as the military, law, and medicine, which were neglected by Confucian education; placing a strong emphasis on supportive interpretations of Classics, some of which Wang himself supplied rather dogmatically; demotion and dismissal of dissenting officials (thus creating conflicts in the bureaucracy); and provision of strong incentives for better performances by clerical staffs, including merit promotion into bureaucratic ranks.

Opposition to the reforms

The magnitude of the reform program was matched only by the bitter opposition to it. Determined criticism came from the groups hurt by the reform measures: large landowners, big merchants, and moneylenders. Noncooperation and sabotage arose among the bulk of the bureaucrats, drawn as they were from the landowning and otherwise wealthy classes. Geographically, the strongest opposition came from the traditionally more conservative northern areas. Ideologically, however, the criticisms did not necessarily coincide with either class background or geographic factors. They were best expressed by many leading scholar-officials, some of whom were northern conservatives while others were brilliant talents from Szechwan. Both the Emperor and Wang failed to reckon with the fact that, by its very nature, the entrenched bureaucracy could tolerate no sudden change in the system to which it had become accustomed. It also reacted against the overconcentration of power at the top, which neglected the art of distributing and balancing power among government offices; against the overexpansion of governmental power in society; and against the tendency to apply policies relatively uniformly in a locally diverse empire.

Without directly attacking the Emperor, the critics attacked the reformers for deviations from orthodox Confucianism. It was wrong, the opponents argued, for the state to pursue profits, to assume inordinate power, and to interfere in the normal life of the common people. It was often true as charged that the reforms—and the resulting changes in government—brought about the rise of unscrupulous officials, an increase in high-handed abuses in the name of strict law enforcement, unjustified discrimination against many scholar-officials of long experience, intense factionalism, and resulting widespread miseries among the population—all of which were in contradiction to the claims of the reform objectives. Particularly open to criticism was the rigidity of the reform system, which allowed little regional discretion or desirable adjustment for differing conditions in various parts of the empire.

Effects of the reforms

In essence the reforms augmented growing trends toward both absolutism and bureaucracy. Even in the short run, the cost of the divisive factionalism that the reforms generated had disastrous effects. To be fair, Wang was to blame for his overzealous if not doctrinaire beliefs, his low tolerance for criticism, and his persistent support of his followers even when their errors were hardly in doubt. Nonetheless, it was the Emperor himself who was ultimately responsible. Determined to have the reform measures implemented, he ignored loud remonstrances, disregarded friendly appeals to have certain measures modified, and continued the reforms after Wang's retirement.

The traditional historians, by studying documentary evidence alone, overlooked the fact that scholar-officials rarely openly criticized an absolutist emperor, and they generally echoed the critical views of the conservatives in assigning the blame to Wang—a revisionist Confucian in public, a profound Buddhist practitioner in his old age, and a great poet and essayist.

**Decline and fall.** Careful balancing of powers in the bureaucracy, through which the rulers acted and from which they received advice and information, was essential to good government in China. The demonstrated success of this principle in early Pei Sung so impressed later scholars that they described it as the art of government. It became a lost art under Shen-tsung, however, in the reform zeal and more so in the subsequent eagerness to do away with the reforms.

Anti-reform period

The reign of Che-tsung (1085/86–1100) began with a regency under another empress dowager, who recalled the conservatives to power. An antireform period lasted until 1093, during which time most of the reforms were rescinded or drastically revised. Though men of integrity, the conservatives offered few constructive alternatives. They achieved a relaxation of tension and a seeming stability but did not prevent old problems from recurring. Some conservatives objected to turning back the clock, especially by

swinging to the opposite extreme, but they were silenced. Once the young emperor took control, he undid what the Empress Dowager had done; the pendulum swung once again to a restoration of the reforms, a period that lasted to the end of the Pei Sung. In such repeated convulsions, the government could not escape dislocation, and the society became demoralized. Moreover, the restored reform movement was a mere ghost without its original idealism. Enough grounds were found by conservatives out of power to blame the reforms for the fall of the dynasty.

Che-tsung's successor, Hui-tsung (reigned 1100–1125/26), was a great patron of the arts and an excellent artist himself, but such qualities did not make him a good ruler. Indulgent in pleasures and irresponsible in state affairs, he misplaced his trust in favourites. Those in power knew how to manipulate the regulatory system to obtain excessive tax revenues. At first the complacent emperor granted more support to government schools everywhere; the objection that this move might flood the already crowded bureaucracy was dismissed, seeing the significant gains it would bring in popular support among scholar-officials. Then the Emperor built a costly new Imperial garden. When his extravagant expenditures put the treasury in deficit, he rescinded scholarships in government schools. Support for him among scholar-officials soon vanished.

More serious was carelessness in war and diplomacy. The Sung disregarded the treaty and coexistence with the Liao Empire, allied itself with the expanding Juchen from Manchuria, and made a concerted attack on the Liao. The Sung commander, contrary to long-held prohibition, was a favoured eunuch; under him and other unworthy generals, military expenditures ran high, but army morale was low. The fall of Liao was cause for court celebration. But because the Juchen had done most of the fighting, they accused the Sung of not doing its share and denied it certain spoils of the conquest. The Juchen soon turned upon the Sung. At this point the Emperor chose to abdicate, giving himself the title of Taoist "emperor emeritus" and leaving affairs largely in the unprepared hands of his son, Ch'in-tsung (reigned 1125/26–1127), while seeking safety and pleasure himself by touring the Yangtze region.

Alliance with the Juchen

During this period the government became increasingly ineffective. The reform movement had enlarged both the size and duties of the clerical staff. The antireform period brought a cutback but also a confusion that presented manipulative opportunities to some clerks. Supervision was difficult because officials stayed only a few years, whereas clerks remained in office for long periods. Bureaucratic laxity spread quickly to the clerical level. Bribes for appointments went either to them or through their hands. It was they who made cheating possible at examinations, using literary agents as intermediaries between candidates and themselves.

The Juchen swept across the Huang Ho plain and found the internally decayed Sung an easy prey. During their long siege of K'ai-feng (1126) they repeatedly demanded ransoms in gold, silver, jewels, other valuables, and general supplies. The court, whose emergency call for help brought only undermanned reinforcements and untrained volunteers, met the invaders' demands and ordered the capital residents to follow suit. Finally, an impoverished mob plundered the infamous Imperial garden for firewood. The court remained convinced that financial power could buy peace, and the Juchen lifted the siege briefly. But once aware that local resources were exhausted and that the regime, even with the return of the Emperor Emeritus, no longer had the capability of delivering additional wealth from other parts of the country, the invaders changed their tactics. They captured the two emperors and the entire Imperial house, exiled them to Manchuria, and put a tragic end to the Pei Sung.

End of the Pei Sung

### NAN (SOUTHERN) SUNG (1127–1279)

The Juchen could not extend their conquest beyond the Yangtze River. The Huai River valley, with its winding streams and crisscrossed marshlands, made cavalry operations difficult. Though the invaders penetrated this region and raided several areas below the Yangtze, they found the weather there too warm and humid for them. Moreover, the farther they went, the stronger the resistance they met. These areas had been leading the country in productivity and population and therefore in defense capability. Besides, the Juchen felt concerned about the areas in the rear that they had already occupied, where one after another of their puppet rulers had failed to secure popular support and the Juchen had been forced to consolidate control by setting up their own administration, following the Liao model of dual government.

**Survival and consolidation.** Despite the fall of the Pei Sung, the majority of scholar-officials refused to identify themselves with the alien conquerors. The same was generally true at the grass-roots level, among numerous roving bands of former volunteer militias, army units that had disintegrated, and bandits who had arisen during the disorder. As time went on, both civilians and military men turned toward the pretender to the throne, Kao-tsung. He was the only son of the former emperor Hui-tsung who had been absent from K'ai-feng and thus spared captivity.

As the founder of the Nan Sung, Kao-tsung devoted his long reign (1127–1162/63) to the arduous task of putting the pieces together. He rediscovered the lost arts of his ancestors: recruiting bureaucrats, securing fiscal resources, and extending centralized control. Since he started with no more than a few thousand troops, he had to place a much greater reliance on sophisticated politics, which he often artfully disguised. By praising the old, established ways of his predecessors, he pleased the conservatives who remained opposed to the reform system. In reality he modified the system he had inherited where it had obviously failed and pragmatically retained the parts that were working. Though he honoured the scholar-officials who had refused to serve under the puppet rulers, he was also glad to have those who had compromised their integrity in so serving. While he denounced the notorious favourites who had misled his father, he used the excuse of being broad-minded in picking many of their former subordinates for key positions, especially those experienced in raising tax revenues. A new network of officials called the fiscal superintendent generals was set up in each region, but they reported directly to court. Urban taxes were increased; they were easier to collect than rural revenues, and prosperous cities did not suffer much from the imposition. The high priority placed on fiscal matters, though not publicized as in the previous reform period in order to avoid a bad image, persisted throughout the Nan Sung, which was a long era of heavy taxation.

Use of sophisticated politics

Some officials, anxious for the recovery of the central plains, wished to have the capital located in Nanking, or farther up the Yangtze in central China. Kao-tsung discreetly declined such advice because these locations were militarily exposed. Instead, he chose Hang (present-day Hang-chou), renaming it Lin-an ("Temporary Safety"). Protected by the coastline and by the mountain ranges at its back, it was a securer retreat. It was popularly referred to as the place of Imperial headquarters (Hsing-tsai), later known to Marco Polo as Quinsai. Economically, it had the advantage of being at the corner of the lower Yangtze delta, the wealthy core of the new empire.

The Nan Sung, through continuous development, eventually became wealthier than the Pei Sung had been. Though its capital was near the sea—the only such case among Chinese empires—and international trade increased, the country was not sea-oriented. Kao-tsung maintained a defensive posture against periodic Juchen incursions from the north and meanwhile proceeded to restore Imperial authority in the hinterland as far west as the strategic Szechwan and in parts of Shensi to its immediate north.

No less important was the need for adequate military forces. Neither conscription nor recruitment would suffice. Because his position was militarily weak but financially strong, Kao-tsung adopted the *chao-an* policy, which offered peace to the various roving bands. The government granted them legitimate status as regular troops, and it overlooked their minor abuses in local matters. Thus, the size of Imperial forces swelled, and the problem of internal security was largely settled. The court then turned its attention to the control of these armies, which was inseparable from the issue of war or peace with the Juchen.

Quest for peace and security

Kao-tsung did not want to prolong the war; he valued most the security of his realm. A few minor victories did not convince him that he could hope to recover North China. Rather, he saw war as a heavy drain on available resources, with the risk of eventual defeat. Nor did he feel comfortable with the leading generals, on whom he would have to rely in case the war went on. He had to get around the critics at court, however, who found the Juchen peace terms humiliating and unacceptable: in addition to an enormous yearly tribute, the Juchen demanded that the Nan Sung formally admit, with due ceremonials, its inferior status as a vassal state. The shrewd emperor found an impeccable excuse for accepting the terms by claiming filial piety; he sought the return of his mother from captivity. To this, no Confucian could openly object. Significantly, Kao-tsung refrained from asking the release of former emperor Ch'in-tsung; such a move would have called into question the legitimacy of his succession.

A dramatic crisis occurred in 1141. On the eve of concluding peace negotiations, Kao-tsung decided to strip the three leading generals of their commands. The generals, summoned to the capital on the pretext of rewarding their merits, were promoted to military commissioners, while their units were reorganized into separate entities directly under Imperial control. Two of the generals reconciled themselves to the nominal honours and sizable pensions, but the third, Yüeh Fei, openly criticized the peace negotiations. He was put to death on a trumped-up charge of high treason. He later became the subject of a great legend, in which he was seen as a symbol of patriotism. At the time, however, his elimination signified full internal and external security for the court.

**Relations with the Juchen.** In spite of Kao-tsung's personal inclination, his artful guiding hand, and the success of his efforts to consolidate the empire, the impulse remained strong among many idealistic Confucians to attempt to recover the central plains. Even when silenced, they were potentially critical of court policies. Kao-tsung eventually decided to abdicate. He left the matter to his adopted heir, but he retained control from behind the throne. The new emperor, Hsiao-tsung (reigned 1162/63–1189/90), sympathetic to the idealists, appointed several of them to court positions and command posts. Information about a Juchen palace coup and alleged unrest in the Juchen Empire, particularly in the parts recently occupied, led to a decision to resume the war. An initial Sung attack was repulsed with such heavy losses that even regrouping took some time to accomplish. Sporadic fighting went on for nearly two years in the Huai Valley, reflecting a military stalemate. This resulted in a significant change in the new peace formula of 1165. The vassal state designation was dropped, and the Nan Sung attained a nearly equal footing with the Juchen, although it had to defer to the latter empire as the senior one.

Resumption of war

After the death of Kao-tsung in 1187, Hsiao-tsung followed the precedent of abdication. The international peace was kept during the brief reign of his son, Kuang-tsung (reigned 1189/90–1194/95), but it was broken again in 1205, during the reign of his grandson, Ning-tsung (reigned 1194/95–1224/25). The 40-year span of continuous peace dimmed the memory of difficulties in waging war. A new generation, nurtured by a flourishing Confucian education, tended to underestimate enemy strength and to think once more about recovering the central plains. The Nan Sung again initiated a northward campaign, and again it met with defeat. The event left no doubt that the Juchen Empire's hold over North China was far beyond the military capability of the southern empire alone. It was also obvious that the Chinese population in North China consisted of new generations brought up under alien domination and accustomed to it.

The Juchen not only retained their military edge over the Nan Sung but also revived their ambition of southward expansion. An offer was made to the governor of Szechwan, who decided to turn against the Sung court in faraway Lin-an and to become king of a vassal state allied with the Juchen. The civilian officials around him, however, took quick action and ended his separatist rebellion. Though a passing danger, it highlighted the fact that the Nan Sung consolidation was not entirely secure; peace was preferred.

**The court's relations with the bureaucracy.** Kao-tsung set the style for all subsequent Nan Sung emperors. The first two emperors in the Pei Sung, both strong militarists, had towered above the relatively modest bureaucracy they had created; most of their successors had found little difficulty in maintaining a balance in the bureaucracy. The circumstances under which the Nan Sung came into being, however, were quite different. Kao-tsung faced tough competition in building up a loyal bureaucracy, first with the two puppet rulers in the North and then from the dual administration the Juchen Empire had set up. He became keenly aware that a cautious handling of bureaucrats was essential. Later, the attempted rebellion in Szechwan taught his successors the same lesson.

Kao-tsung was an attentive student of history who consciously emulated the restoration by the Tung (Eastern) Han (AD 25–220) and defined his style as the "gentle approach." This meant using bureaucratic tactics to deal with the bureaucrats themselves. The gentle approach proved helpful in maintaining a balance at court and thus protecting councillors and Imperial favourites from the criticism of "opinion-officials." Absolutism had grown since the middle of the Pei Sung; the emperors had delegated much more power than before to a few ranking councillors. Similarly, Imperial favourites—*e.g.*, eunuchs, other personal attendants of the emperor, and relatives of the consorts—gained influence.

Kao-tsung's handling of bureaucrats

The opinion-officials by virtue of their rank or conviction wished to speak against those who abused power and influence; as a result of the factionalism that had plagued the late Pei Sung, their effectiveness had declined and never recovered. But as long as absolutism was qualified by Confucian values and the monarch cherished a Confucian image, he had to learn to deal with some adverse opinions; he often resorted to sophisticated delaying tactics. Skilled at bureaucratic manipulation, the Nan Sung emperors listened to criticism with ostensible grace, responded appreciatively, and made it known that they had done so, but they did not take concrete action. Sometimes an emperor would either order an investigation or express a general agreement with the criticism, thereby preventing the critics from making an issue of it by repeated remonstrances. On other occasions the emperors would listen to the critics and commend them for their courage, but, to avoid stirring up a storm, the court would explicitly forbid the circulating of private copies of the criticisms among other scholar-officials. More subtly, the court would sometimes announce an official version of such criticism, leaving out the most damaging part. Likewise, rectifying edicts that followed the acceptance of criticism often had little substance. Reconciliation at court was another technique; an emperor would deliberately, if not evasively, attribute criticism to probable misunderstanding, assemble the parties in dispute, ask them to compose their differences, caution those under attack to mend their ways, and suggest to the critics that their opinions, though valid, should be modified. The handling of severe critics who refused to change their stand required different tactics. Seemingly accepting their adverse opinion, the court might reward them by promotion to a higher position, whose functions did not include the rendering of further advice. Rarely did the court demote or punish opinion-officials, especially those with prestige; sometimes it would not even permit them to resign or to ask for a transfer. Any such move tended to damage the court's valuable Confucian image. On sensitive issues the emperors were likely to invoke their absolutist power, but this was usually handled gently, by quietly advising the opinion-officials to refrain from commenting on the issues again.

Under this bureaucratized manipulation by the court, the institution of opinion-officials degenerated. Often the emperors appointed their own friends to such posts; but just as often, when the emperors hinted that they were displeased with certain ministers, the opinion-officials dutifully responded with unfavourable evidence, thus furnishing the court with grounds for dismissals. Such Imperial manipulations served manifold purposes: safeguarding absolutist

power and its delegation to various individuals, disguising absolutism, and keeping the bureaucracy in balance.

**The chief councillors.** The later Nan Sung emperors preferred not to take on the awesome burden of managing the huge and complex bureaucracy. Most of them were concerned chiefly with security and the status quo.

Delegation of power in the Nan Sung

The Nan Sung court delegated a tremendous amount of power and thus had a series of dominant chief councillors; none of them, however, ever was a potential usurper. No bureaucrat during the Sung era had a political base, a hereditary hold, or a personal following in any geographic area. In addition, the size of the bureaucracy and fluidity of its composition precluded anyone from controlling it. The tenure of chief councillor was essentially dependent upon the sanction of the emperor. At times even the chief councillor had to reaffirm his loyalty along with other bureaucrats. Loyalty in absolutist terms being another name for submission, the court, bureaucratized as it was, retained its supreme position beyond challenge.

Nevertheless, the history of Nan Sung politics had much to do with powerful chief councillors, increasingly so as time went on. Kao-tsung at first had a rapid succession of ranking ministers, but none of them measured up to the difficult task at hand: seeking external security by maintaining peace with the northern empire, and internal security by undermining the power of leading generals. Only the chief councillor Ch'in Kuei did both; moreover, he increased tax revenues, strengthening the fiscal base of the court and enriching the private Imperial treasury. For these merits, he was given full support to impose tight control over the bureaucracy as long as he lived. Powerful as he was, he avoided doing anything that might arouse Imperial suspicion. He had many dissident scholar-officials banished from court, but only with Imperial sanction. He accommodated many bureaucrats, even those who neither opposed nor followed him, but he made many of them jealous of his great power and of the rapid promotions he gave to his son and grandson. Ch'in Kuei failed, however, to properly assess the wiles of his bureaucratized master, who turned out to be the more masterful politician. Upon Ch'in Kuei's death, the Emperor shifted all blame to him and recalled from banishment some of his opponents, thus restoring in time a balance in the bureaucracy.

After his voluntary abdication, Kao-tsung retained his power by using Hsiao-tsung more or less as a chief councillor. Hsiao-tsung subsequently failed to find a firm hand among his successive ministers, and the great burden on himself was probably one reason that he chose to abdicate. His son, Kuang-tsung, was mentally disturbed, unresponsive to bureaucratic consensus, and pathetically dominated by his consort. He turned against Hsiao-tsung and even refused to perform state funeral rites when the retired emperor died—an unprecedented default that shocked the court. The solution was equally unprecedented; the Empress Dowager, the palace personnel, and the ranking ministers agreed to force his abdication and oversee the accession of Ning-tsung. Through the crisis, Han T'o-chou, mentioned above in connection with the renewed war against the Juchen, moved rapidly into power. Related originally to the Empress Dowager and again to a new consort, he received deferential treatment from Ning-tsung. He was made chief councillor but found it hard to control many bureaucrats who objected to his lack of scholarly qualifications, questioned his political ability, and criticized his nepotistic appointments. Reacting to the hostility, he made first a crucial mistake and then a fatal one. First, he banned a particular school of Confucian idealists, led by Chu Hsi (see below *The rise of Neo-Confucianism*). This proved unpopular, even among neutral scholar-officials. After he rescinded the ban, he attempted to recruit support and to reunite the bureaucracy by initiating the war. After its defeat in the war, the Sung sacrificed him in its search for peace.

Rise of Han T'o-chou

Shih Mi-yüan emerged as the dominant chief councillor. Coming from a bureaucratic family background, he understood the gentle approach and the importance of accommodating various kinds of bureaucrats in order to achieve a political balance. Promoting on merit and refraining from nepotism, he restored stability. He recognized that the ideological prestige the followers of Chu Hsi had won had become a political factor, and he appointed some of their prominent leaders to highly respectable posts but without giving them real power. Like the emperors he served, he wanted to have both authority and a good political image. Ning-tsung had no son, and the chief councillor helped him adopt two heirs. When the Emperor died without designating an heir apparent, Shih Mi-yüan arbitrarily decided in favour of the younger one, which was contrary to the normal order of succession but had the backing of palace-connected personnel.

Both Li-tsung (reigned 1224/25–1264/65) and his successor Tu-tsung (reigned 1264/65–1274) indulged excessively in pleasure, though much of it was carefully concealed from the public. Shortly after the death of Shih Mi-yüan, the role of chief councillor went to Chia Ssu-tao. Though denounced in history, he actually deserves much credit. He dismissed many incompetents from the palace, the court, the bureaucracy, and the army. He curbed excessive corruption by instituting minor administrative reforms. His strict accounting made the generals personally liable for misappropriation of funds. A system of public fields was introduced, cutting into the concentration of landownership by requisitioning at a low price one-third of large estates beyond certain sizes and using the income for army expenditures when the government faced external danger and fiscal deficit. These measures, however, hurt the influential elements of the ruling class, making Chia unpopular. He too had failed to practice the gentle approach. He was denounced by those who had defected to the enemy and later reconciled their guilt by placing the blame on him.

Accomplishments of Chia Ssu-tao

Except in name, the several dominant chief councillors were nearly actual rulers by proxy. They ran the civil administration, supervised both state finance and military affairs, and controlled most scholar-officials by some varying combination of gentle accommodation and high-handed pressure. The emperors, however, kept their separate Imperial treasury—from which the government in deficit had to borrow funds—and their private intelligence systems to check upon the chief councillors. Moreover, potential competitors always existed in the bureaucracy, ready to criticize the chief councillors whenever state affairs went badly enough to displease or disturb the emperors. The chief councillors had enormous power only by virtue of the Imperial trust, and that lasted only as long as things went tolerably well.

**The bureaucratic style.** Regular posts in the Nan Sung civil service numbered about 20,000 without counting numerous sinecures, temporary commissions, and a slightly larger number of military officers. Besides eliminating most patronage privileges, by which high officials were entitled to obtain an official title for a son or other family member, the court occasionally considered a general reduction in the size of the bureaucracy. But vested interests always opposed it. Those who entered government service seldom dropped out or were thrown out. Meanwhile, new candidates waiting for offices came in waves from state examinations, from extra examinations on special occasions, from graduation from the National Academy, or from special recommendations and unusual sponsorship; others gained official titles because their families contributed to famine relief or to military expenditures. Thus, the ever-increasing supply of candidates far exceeded the vacancies.

According to Confucian theory, any prosperity that made possible more books in print, schools, and availability of a better educated elite was all for the good. But the original Confucian ideal intended to have the elite serve the society in general and the community in particular rather than flood the bureaucracy. Rising educational standards made the competition at examinations harder and perhaps raised the average quality of degree holders.

Families with members in the bureaucracy responded in part by successfully increasing the importance of other avenues of entrance into government service, especially the "protection" privilege that allowed high officials to secure official rank for their protégés (usually junior family members). People outside the civil service responded by altering their goals and values and by reducing the stress

on the importance of entering the bureaucracy. It is not accidental that this era saw the spread of Neo-Confucian academies that emphasized moral self-development, not success in examinations, as the proper goal of education.

Changes in the legal system

During the Sung period there was increased emphasis on morals and ethics and a continuous development of the law. The early Sung had adopted a legal code almost wholly traceable to an earlier T'ang code, but Sung circumstances differed from those of the T'ang. As a result, there was a huge output of legislation in the form of Imperial edicts and approved memorials that took precedence over the newly adopted code and soon largely displaced it in many areas of law. Sung legal bureaucrats periodically compiled and edited the results of this outpouring of new laws. The new rules not only altered the content of the (largely criminal) sphere covered by the code but also legislated in the areas of administrative, commercial, property, sumptuary, and ritual law. There were literally hundreds of compilations of various sorts of laws.

Perhaps as a result of the growth of this legal tangle from the late Pei Sung onward, magistrates made increasing use of precedents, decisions by the central legal authorities on individual cases, in reaching legal decisions. The government sought to help its officials by instituting a variety of devices to encourage officials and prospective officials to learn the law and to certify that those in office did have some familiarity with things legal. There was an increase in the writing and publication of other sorts of works concerned with the law, including casebooks and the world's oldest extant book on forensic medicine. Despite the appearance of such works, which were intended to help them, officials were under strong pressure to rule in a conservative way and to avoid rocking the boat.

Many scholar-officials sought simply to keep things quiet and maintain the appearance that there was no serious trouble. The bureaucratic style was to follow the accustomed ways in accordance with proper procedure, find expedient solutions based upon certain principles in spirit, make reasonable compromises after due consideration of all sides, and achieve smooth reconciliations of divergent views. To protect one's own career record it was essential to engage in time-consuming consultations with all appropriate offices and to report to all concerned authorities so that everyone else would have a share of responsibility. Anyone who criticized the bureaucratic style would be going against the prevalent mode of operation, namely, mutual accommodation. Even the emperor adopted the bureaucratic style.

The picture was not entirely dark. Evasions and deviations notwithstanding, the letter of the laws and the formalities of procedures had to be fulfilled. Definite limits were set on official negligence and misconduct. For example, suppressing evidence or distorting information were punishable offenses. Minor juggling of accounts went on, but outright embezzlement was never permissible. Expensive gifts were customary and even expected, but an undisguised bribe was unacceptable. The refined art of the bureaucratic style was not sophistry and hypocrisy alone; it required a circumspect adherence to the commonly accepted substandard norms, without which the maintenance of government would have been impossible.

**The clerical staff.** The norms for the clerks were even lower, especially in local government. Some 300 clerks in a large prefecture or nearly 100 in a small one were placed under the supervision of a few officials. The clerks had numerous dealings with various other elements in the community, whereas the officials, being outsiders, rarely had direct contacts. Holding practically lifelong tenure after benefiting from the cumulative experience of their fathers and uncles before them, the clerks knew how to operate the local administrative machinery far better than did the officials, who served only brief terms before moving elsewhere. Clerks often received inadequate salaries and were expected to support themselves with "gifts" from those needing their services. Under honest, strict, and hardworking magistrates, the clerks would recoil, but only briefly, because such magistrates would soon either gain promotion for their remarkable reputations, or their strict insistence on clean government would become intolerable to their superiors, colleagues, subordinates, and influential elements in the community who had connections with high circles. Though all bureaucrats complained of clerical abuses, many connived with the clerks and none had a viable alternative to the existing situation. One significant suggestion was to replace the clerks with the oversupply of examination candidates and degree holders, who presumably had more moral scruples. But this solution had no chance of being considered because it implied a downgrading of the status of those who considered themselves to be either potential or actual members of the ruling class.

Clerical abuses

The law did place definite limits on clerical misbehaviour. But when a clerk was caught in his wrongdoing, he knew enough to save himself—taking flight before arrest, getting a similar job elsewhere under a different name, defending himself through time-consuming procedures, appealing for leniency in sentencing, requesting a review, or applying for clemency on the occasion of Imperial celebrations. What prevented clerical abuses from getting worse was not so much official enforcement of legal limits as it was the social convention in the community. For themselves as well as for their descendants, the clerks could ill afford to overstep the socially acceptable limits.

The net result of a large bureaucracy and its supporting clerical staff, accommodating one another in various defaults, malfunctions, and misconduct within loose limits, was a declining tax yield, tax evasion by those who befriended colluding officials and clerks, and an undue shift of the tax burden onto those least able to pay.

**The rise of Neo-Confucianism.** The rise of the particular school of Neo-Confucianism led by Chu Hsi takes on special meaning in this context. The Neo-Confucian upsurge beginning in the late T'ang embraced many exciting extensions of the Classical vision. Noteworthy during the Pei Sung was the emergence of a new Confucian metaphysics that was influenced by Buddhism and that borrowed freely from Taoist terminology while rejecting both religions. Of relevance to Nan Sung political and social conditions was its continuous growth into a well-integrated philosophical system that synthesized metaphysics, ethics, social ideals, political aspirations, individual discipline, and self-cultivation.

Chu Hsi and his followers

The best thinkers of the early Nan Sung were disillusioned by the realization that previous Neo-Confucian attempts had failed. Reforms that had sought to apply statecraft had ended in abuses and controversies. The spread of education had not coincided with an uplifting of moral standards. The loss of the central plains was a great cultural shock, but to talk of recovering the lost territory was useless unless it was preceded by a rediscovery of the true meaning of Confucianism. To Chu Hsi and his followers, a state permeated by true Confucian practices would be so internally strong and would have such an attraction for outsiders that the retaking of the north would require only a minimal effort; a state lacking true Confucian practices would be so internally weak and unattractive that retaking the lost territories would be quite impossible.

Moreover, threatened by the Juchen adoption of the same heritage, the Sung felt driven to make an exclusive claim to both legitimacy and orthodoxy. Such a claim required that the new departures be interpreted as reaffirmation of ancient ideals. Thus, the intellectual trend that developed under Chu Hsi's leadership was at first referred to as Tao Hsüeh (the School of True Way) and later as Li Hsüeh (the School of Universal Principles). Education, to the thinkers of this school, meant a far deeper self-cultivation of moral consciousness, the ultimate extent of which was the inner experience of feeling at one with universal principles. These men, who might be described as transcendental moralists in Confucianism, also made a commitment to reconstruct a moral society—to them the only conceivable foundation for good government. With missionary-like zeal, they engaged in propagation of this true way and formed moral-intellectual fellowships. Chu Hsi, the great synthesizer, ranked the Classics in a step-by-step curriculum, interpreted his foremost choices, collectively known as the *Four Books,* summed up a monumental history in a short version full of moralistic judgments, prepared other extensive writings and sayings

of his own, and opened the way for an elementary catechism, entitled the "Three Word Classics," that conveyed the entire value system of this school in simple language for what approximated mass education.

Many idealistic scholars flocked to Chu Hsi, his associates, and his disciples. Frustrated and alienated by the prevalent conditions and demoralizing low standards, these intellectuals assumed a peculiar archaic and semireligious life-style. Prominent in scholarship, educational activities, and social leadership and filling some relatively minor government posts, they asserted their exclusive ideological authority with an air of superiority, much to the displeasure of many conventional Confucians. Though they were not keen about politics, the prestige they acquired was an implicit threat to those in power. The chief councillor Han T'o-chou was particularly alarmed when he found some of his political adversaries sympathetic to and even supporting this particular school. A number of other bureaucrats at various ranks shared Han's alarm; one after another, they accused this school of being similar to a subversive religious sect, calling it a threat to state security and attacking its alleged disrespect for the court. The school was proscribed as false learning and un-Confucian. Several dozen of its leaders, including Chu Hsi, were banished, some to distant places. Thenceforth, all state examination candidates had to declare that they had no connection with the school.

The government's response to Chu Hsi

Most historical accounts follow the view that the controversy was another example of factional strife, but that was not the case. The attackers were not a cohesive group, except for their common resentment toward this school, nor was the school itself an active group in politics. The conflict was in fact one between two polarized levels—political power and ideological authority. The nature of the Confucian state required that the two should converge if not coincide.

The persecution boomeranged by making heroes out of its victims and arousing sympathy among neutral scholar-officials. Realizing his mistake a few years later, Han lifted the ban. Most historical accounts leave an erroneous impression that, once the ban was removed, the Chu Hsi school of Neo-Confucianism by its preeminence soon gained wide acceptance, which almost automatically raised it to the coveted status of official orthodoxy. But in reality the rise to orthodoxy was slow and achieved by political manipulation, occasioned by an internal crisis of Imperial succession and then by the external Mongol threat. Shih Mi-yüan, the chief councillor who made Li-tsung emperor, created circumstances that forced the elder heir to commit suicide. This was damaging to the image of the court and to that of Shih himself. Mending political fences, he placed a few of the school's veteran leaders in prestigious positions in order to redress the balance of the bureaucracy. In 1233, the year before the Mongol conquest of Juchen, the Mongols honoured Confucius and rebuilt his temple in Peking. In 1237 their emerging nomadic empire, already occupying a large portion of North China, reinstituted a civil service examination, thus claiming that it, too, was a Confucian state. Threatened both militarily and culturally, the Nan Sung made Chu Hsi's commentaries official, his school the state orthodoxy, and its claim the accepted version—that the true way of Confucius had been lost for more than 1,000 years and that the line of transmission was not resumed until, inspired by the early Pei Sung masters, Chu Hsi reestablished it. This implied that whatever Confucianism the Mongols took over was but a pale imitation and without legitimacy.

**Internal solidarity during the decline of the Nan Sung.** Honouring the Chu Hsi school did not reinvigorate the Nan Sung administration, but the military, despite some weaknesses, maintained an effective defense against the Mongols for four decades—the longest stand against Mongol invasions anywhere. The final Sung defeat came in part because the Mongol forces, frustrated for many years in their attempts to break the main Sung line of resistance, drove through territories to the west and outflanked the Sung defenders. The Sung capital, Lin-an, finally fell in 1276 without much fighting, after high-ranking officials and officers had fled. The empire finally came to an end in 1279, after its last fleet had been destroyed near Canton, when a loyal minister with the boy pretender to the throne committed suicide by jumping into the sea.

Later Chinese historians attempted to explain the fall of the Nan Sung as the result of internal decay and abuses, and so they stressed the problems of heavy taxation, inflation of paper currency, bureaucratic laxity, and clerical abuses. The absence of any large-scale uprisings among the peasantry, however, suggests that they overstated the seriousness of such problems. To explain this lack of major uprisings, most historical accounts point to Chinese patriotism because the war against the Mongols was for cultural rather than merely dynastic survival. Though partly true, this was not the only reason. The impressive internal solidarity involved many other factors: (1) the government mobilized the resources of the wealthiest region, that of the lower Yangtze, without overburdening other parts; (2) the tax burden and the emergency requisitions fell mostly upon the prosperous urban sectors rather than on rural areas, the backbone of the empire; (3) scholar-officials in many areas, in spite of their shortcomings, were sophisticated in the art of administration, moving quickly to put down small uprisings before they got larger, or offering accommodative terms to induce some rebel leaders to come over while dividing the rest. Finally, the Neo-Confucian values had pervaded the country through more books, more schooling, and more efforts by Neo-Confucians to promote moral standards, community solidarity, and welfare activities, and through widespread Neo-Confucian roots planted at the local levels by half-literate storytellers, makeshift theatres, and traveling companies in various performing arts.

Factors in Nan Sung solidarity

The examination system itself played a major role in the Confucianization of Chinese society. Only a small percentage of the candidates actually passed the degree examinations and entered the civil service. The vast majority, thoroughly imbued with Confucian studies, returned to the larger society, often to serve as teachers to the next generation. Furthermore, the examination system reinforced the deeply Confucian character of the curriculum, from the lowest level of primary education to the highest level in the academies. Children began imbibing Confucian moral precepts when they began to read. These precepts stressed loyalty, and that in turn probably helped bolster the strength of the dynasty in the face of foreign invasion and helped limit internal disloyalty.

## SUNG CULTURE

The Sung was an era of great change in most facets of Chinese life. Some of these developments were the outgrowths of earlier patterns, while others were largely born under that dynasty. These developments often related to or were made possible by major changes in Chinese economic life.

An agricultural revolution produced plentiful supplies for a population of more than 100,000,000—by far the largest in the world at the time. Acreages under cultivation multiplied in all directions, stretching across sandy lands, climbing uphill, and pushing back water edges. A variety of early ripening rice, imported during the 11th century from Champa in modern Kampuchea, shortened farming time to fewer than 100 days, making two crops a year the norm and three crops possible in the warm south. Among other new crops the most important was cotton, which provided clothing for rich and poor alike; silk and hemp were also important. Improved tools, new implements, and mechanical devices that raised manpower efficiency were widely used and found their way into guidebooks used by the literate community leaders. Mineral productivity of such substances as gold, silver, lead, and tin also increased. Consumption of iron and coal grew at a faster rate from 850 to 1050 than it had in England during the first two centuries of the Industrial Revolution. The Chinese, however, never combined the two resources to generate power mechanically.

The agricultural revolution of the Sung

Manufacturing made tremendous headway within the skill-intensive pattern but with the aid of new devices, better processing, a beginning of division of labour, and expertise. Chinese porcelain attained international fame. Though information on ordinary handicrafts was available

in handbooks and encyclopaedias, advanced skills were guarded as trade secrets. Specialization in production and regional trade stimulated mutual growth.

Improved transportation

Transportation facilities improved, allowing production away from the sources of supplies and making products available to distant parts. The state maintained highways, with staffed stations, for official travel and a courier service network, the latter being an index of centralized government control. Along the highways and branching byways stood private hostels and inns frequented by private traders. Rivers carried tribute vessels and barges, private shipping, transfer crafts, fishing boats, and pleasure yachts. Large ships with multiple decks were propelled by fast-moving wheels paddled by manpower; many sailed on the high seas, aided by accurate compasses, charts, and instruments as well as by experience in distant navigation. The expanding sea trade, apart from that with Japan and Korea, moved southward and linked up with merchants from Persia and Arabia. Some Chinese merchants began to settle in Southeast Asia. For the first time in history, Chinese naval forces assumed a vital military role, though China had not become a sea power.

An advanced money economy was everywhere in evidence. Many cultivated lands produced cash crops. By 1065 the Pei Sung government was taking in annual cash tax payments that were 20 times what the T'ang had received in 749. The income of the Nan Sung consisted of more cash revenues than grain and textile receipts. The economy had progressed to such a state that it needed more means of exchange. Merchants used drafts called *fei-ch'ien* ("flying money") and certificates of deposits made elsewhere. State monopoly agencies in salt and tea followed with their respective certificates, which were as good as money. The government first permitted printed paper money for limited regional circulation and then authorized it as nationwide legal tender. (China was the first country to do so.)

Busy transactions approached a commercial revolution, carried on by rapid calculations on the abacus, a specialized service skill that remained unmatched until the appearance of adding machines and computers. Cities changed: the T'ang pattern of walled-in blocks, each for a particular trade, broke down; stores appeared in various parts of cities, and trade guilds proliferated. Though official documents and scholarly essays adopted a downgrading tone toward commercial activities, Sung China became a society of wholesalers, shippers, storage keepers, brokers, traveling salesmen, retail shopkeepers, and peddlers. Urban life reached a new intensity. The populations of several metropolitan areas approached 1,000,000.

Changes in urban life

Crowding was serious. Houses usually had narrow frontages. Fires were frequent and disastrous. Neighbourhood fire squads, with water containers at hand, could not prevent destruction, and some fires lasted several days. Nonetheless, prosperity was the keynote of urban life. Teahouses, wine shops, exquisite cuisines, and catering services for private parties existed in multitude and variety. Pleasure grounds provided daily amusement and festival merriment with acrobats, jugglers, wrestlers, sword swallowers, snake charmers, fireworks, gambling, performing arts of all sorts, puppet shows, storytellers, singing girls, and professionally trained courtesans. Upper-class families enjoyed higher culture, with such diversions as music, pets, intricate games, hobbies, calligraphy, painting, and poetry. Noticeably declining were hunting, horseback riding, and polo. Gentility displaced sportsmanship. The prosperous cities also provided easy prey for pickpockets and professional thieves. Inasmuch as pauperism appeared in cities, parallel to rural underemployment and unemployment, the government undertook relief and such welfare measures as orphanages, nursing homes for the aged poor, charitable graveyards, and state pharmacies.

Expansion of knowledge

Knowledge expanded because of specialization. Medicine embraced such skills as acupuncture, obstetrics, dentistry, laryngology, ophthalmology, and treatment of rheumatism and paralysis. The demand for improvement in technology, aided by certain concerns of the Neo-Confucian philosophy, helped to promote numerous investigations approaching science. Literacy spread with printing, which evolved from rubbing through block printing to the use of movable type that facilitated a large quantity of production at reduced cost. A large number of scholars achieved high standing through Classical studies, newly developed archaeology, philosophical interpretations, statecraft ideas, classical forms of poetry, an evolving lyric poetry called *tz'u,* which had its origin in singing, and written versions of popular songs, called *san-ch'u.* Of greatest influence on scholar-officials in succeeding generations was a masterly prose style that was original and creative but that was always used in the name of reviving ancient models. Diversified and specialized developments widened knowledge so much that scholars compiled voluminous histories, collected works, comprehensive handbooks, compendiums, and encyclopaedias. Fine arts also reached new heights.

The term early modern has often been applied in describing Sung culture because it not only advanced beyond the earlier pattern in China and far ahead of the rest of the world at the time but it also had many startlingly new features that approximated later developments in western Europe. This characterization, though helpful to highlight and appreciate the Sung progress, tends to be misleading. The so-called early modern stage did not pave the way for more modernity later. On the contrary, the Sung pattern attained cultural stability, giving rise to the myth of an unchanging China.

These conflicting images stemmed from the cultural and regional diversity of the Sung, in which modern-style advances existed alongside continuing older practices. In some areas such as the delta lands immediately south of the Yangtze River, sizable estates grew up with a complicated social pattern characterized by tenant farming. Elsewhere, in less well-developed areas, there was a greater proportion of owner-farmers, while in still other regions, the landlords tried to bind the tillers to the soil. The same confusion was reflected in the status of women. During the Sung the notorious practice of foot binding first became common, clearly marking a fall in the status of women, but there is evidence that during the Nan Sung (unlike any other Chinese dynasty) daughters as well as sons could inherit property in their own names. Furthermore, Sung families tried in various ways to strengthen the ties created by the marriages of their daughters to other families.

The extraordinarily rapid pace of economic and technological change that marked the Pei Sung seems to have slowed during the Nan Sung. For reasons that are not wholly clear, Chinese society did not break through its inherited patterns in any radically new ways. It may be that, with an abundance of inexpensive labour, economic rationality moved men to produce through increased amounts of labour rather than through innovation or capital investment. This disincentive to investment helped create a relatively stable economic and technological pattern that remained with little change for centuries thereafter. Despite this slowing of economic and technological development, however, the Sung did give birth to changes. It not only saw the creation of a new Confucian synthesis but also helped create the devices that spread the new ideas among the people at large. The urban and urbanized culture that arose in the Sung was retained and developed in succeeding dynasties, when the early modern (or neotraditional) pattern created in the Sung provided both the model for and the basis of the gradual transformation of some aspects of Chinese life that belied the image of China as unchanging. (J.T.C.L./B.E.McK.)

## The Yüan, or Mongol, dynasty

### THE MONGOL CONQUEST OF CHINA

Genghis Khan rose to supremacy over the Mongol tribes in the steppe in 1206, and within a few years he attempted to conquer northern China. By securing the allegiance of the Tangut state of Hsi Hsia in what are now Kansu and northeastern Tibet (1209), he disposed of a potential enemy and prepared the ground for an attack against the Chin state of the Juchen in northern China. At that time the situation of Chin was precarious: the Juchen were exhausted by a costly war (1206–08) against their hereditary enemies, the Nan Sung. Discontent among the

Preparations for war against China

non-Juchen elements of the Chin population (Chinese and Khitan) had increased, and not a few Chinese and Khitan nobles defected to the Mongol side. Genghis Khan, in his preparation for the campaign against Chin, could therefore rely on foreign advisers who were familiar with the territory and the conditions of the Chin state.

**Invasion of the Chin.** The Mongol armies started their attack in 1211, invading from the north in three groups; Genghis Khan led the centre group himself. For several years they pillaged the country; finally, in 1214 they concentrated upon the central capital of the Chin, Chung-tu (modern Peking). Its fortifications proved difficult to overcome, so the Mongols concluded a peace and withdrew. Shortly afterward the Chin emperor moved to the southern capital at Pien (modern K'ai-feng). Genghis Khan considered this a breach of the armistice; a renewed attack brought large parts of northern China under Mongol control and finally resulted in 1215 in the capture of Chung-tu (renamed Ta-tu in 1272). The Mongols had had little or no experience in siege craft and warfare in densely populated areas; their strength had been chiefly in cavalry attacks. The assistance of defectors from the Chin state probably contributed to this early Mongol success. In subsequent campaigns the Mongols relied even more upon the sophisticated skills and strategies of the increased number of Chinese under their control.

After 1215 the Chin were reduced to a small buffer state between the Mongols in the north and Sung China in the south, and their extinction was but a matter of time. The Mongol campaigns against Hsi Hsia in 1226–27 and the death of Genghis Khan in 1227 brought a brief respite for Chin, but the Mongols resumed their attacks in 1230.

The Sung Chinese, seeing a chance to regain some of the territories they had lost to the Juchen in the 12th century, formed an alliance with the Mongols and besieged Pien in 1232. The Emperor of Chin left Pien in 1233, just before the city fell, and took up his last residence in Ts'ai Prefecture (Honan), but this refuge was also doomed. In 1234 the Emperor committed suicide, and organized resistance ceased. The southern border of the former Chin state—the Huai River—now became the border of the Mongol dominions in northern China.

**Invasion of the Sung.** During the next decades there was an uneasy coexistence between the Mongols in northern China and the Sung state in the south. The Mongols resumed their advance in 1250 under the grand khan Möngke and his brother Kublai Khan—grandsons of Genghis Khan. Their armies outflanked the main Sung defenses on the Yangtze River and penetrated deeply into southwestern China, conquered the independent Tai state of Nan Chao (in what is now Yunnan), and even reached present-day northern Vietnam. Möngke died in 1259 while leading an army to capture a Sung fortress in Szechwan, and Kublai succeeded him. Kublai sent an ambassador, Hao Ching, to the Sung court with an offer to establish peaceful coexistence. Hao did not reach the Sung capital of Lin-an (now Hang-chou), however, but was interned at the border and regarded as a simple spy. The Sung chancellor, Chia Ssu-tao, considered the Sung position strong enough to risk this affront against Kublai; he thus ignored the chance offered by Kublai and instead tried to strengthen the military preparations against a possible Mongol attack. He secured military provisions by a land reform that included the confiscation of land from large owners, but this alienated the greater part of the landlord and official class. The Sung generals, whom Chia distrusted, also had grievances, which may explain why a number of them later surrendered to the Mongols without fighting.

Mongol penetration into southwestern China

From 1267 onward, the Mongols, this time assisted by numerous Chinese auxiliary troops and technical specialists, attacked on several fronts. The prefectural town of Hsiang-yang (modern Hsiang-fan) on the Han River was a key fortress, blocking the access to the Yangtze River. The Mongols besieged Hsiang-yang for five years (1268–73). The Chinese commander finally surrendered in 1273, after he had obtained a solemn promise from the Mongols to spare the population, and he took office with his former enemies.

Kublai Khan's warning to his forces not to engage in indiscriminate slaughter seems to have been heeded to a certain extent. Several prefectures on the Yangtze River surrendered; others were taken after brief fighting. In January 1276, Mongol troops reached Lin-an. Last-minute attempts by the Sung court to conclude a peace failed, and the Mongol armies took Lin-an in February. The reigning Sung empress dowager and the nominal emperor—a boy—were taken to Ta-tu and granted an audience by Kublai Khan.

National resistance in the Sung state continued, however, and loyalists retreated with two Imperial princes into the southern province of Fukien and from there to the region of Canton. In 1277 the last remnants of the court left Canton and eventually fled the mainland by boat. A faithful minister drowned himself and the last surviving Imperial prince in the ocean in March 1279. When organized resistance ceased soon afterward, foreign invaders controlled the whole Chinese empire for the first time in history.

## CHINA UNDER THE MONGOLS

**Mongol government and administration.** After their initial successes in northern China in 1211–15, the Mongols faced the problem of how to rule and extract material benefits from a largely sedentary population. They were assisted by Khitan and Chinese and even Juchen renegades; these defectors were treated as "companions" (*nökör*) of the Mongols and were given positions similar to the higher ranks of the steppe aristocracy—their privileges included the administration and exploitation of fiefs considered as their private domain.

Mongol use of defectors

*Early Mongol rule.* The government system during the early years of the Mongol conquest was a synthesis of Mongol military administration and a gradual return to Chinese traditions in those domains ruled by former subjects of the Chin state. The most important office or function in Mongol administration was that of the *darughatchi* (or seal bearer), whose powers were at first all-inclusive; only gradually were subfunctions entrusted to specialized officials in accordance with Chinese bureaucratic tradition. This refeudalization of northern China along Mongol lines with a slight understructure of Chinese-type bureaucrats lasted for many years.

The central administration of Mongol China was largely the creation of Yeh-lü Ch'u-ts'ai, originally a Chin state official of Khitan extraction who had acquired a profound Chinese scholarship and had become one of Genghis Khan's trusted advisers. Yeh-lü continued to serve under Ögödei, who became grand khan in 1229, and persuaded him to establish a formal bureaucracy and to replace indiscriminate levies with a rationalized taxation system along Chinese lines. An important part of Yeh-lü's reforms was the creation of the Central Secretariat (Chung-shu sheng), which centralized the civilian administration and achieved some continuity. The territory was divided into provinces, and the provincial administrations were responsible for regularized taxation. The people had to pay a land tax and a poll tax, either in kind (textiles and grain) or in silver. Merchants had to pay a sales tax. Monopolies on wine, vinegar, salt, and mining products were also introduced. All this enabled the treasuries of the Mongol court to accumulate considerable wealth.

In spite of the success of his economic policy, Yeh-lü's influence decreased during his later years. One reason was bitter opposition from the Mongol feudatories and from those Chinese, Juchen, and Khitan nobles who were used to ruling independently in their appanages, which they exploited at will. Also, Ögödei himself apparently lost interest in the internal conditions of the Mongol dominion in China. During the 1230s Muslims from the Middle East had already begun to fill the higher positions at the Mongol court, and their ruthless exploitation of the Chinese created widespread resentment of Mongol rule. A relapse into feudal anarchism seemed inevitable, and Yeh-lü's reforms fell into temporary abeyance. China was ruled more or less like a colony by the foreigners and their allies.

*Changes under Kublai Khan and his successors.* Kublai Khan's ascendancy in 1260 marked a definite change in Mongol government practice; Kublai moved the seat

of Mongol government from Karakorum in Mongolia to Shang-tu (Upper Capital) near modern To-lun in Inner Mongolia. In 1267 the official capital was transferred to Chung-tu, where Kublai ordered the construction of a new walled city, replete with grand palaces and official quarters, that was renamed Ta-tu ("Great Capital") before its completion. Under its Turkicized name, Cambaluc (Khanbaliq, "The Khan's Town"), the capital became known throughout Asia and even Europe. But, true to nomad traditions, the Mongol court continued to move between these two residences—Shang-tu in summer and Ta-tu in winter. With the establishment of Ta-tu as the seat of the central bureaucracy, Mongolia and Karakorum no longer remained the centre of the Mongol Empire. Mongolia began to fall back to the status of a northern borderland, where a nomadic way of life continued and where Mongol grandees, dissatisfied with the growing Sinicization of the court, repeatedly engaged in rebellions.

Kublai, who even prior to 1260 had surrounded himself with Chinese advisers such as the eminent Buddho-Taoist Liu Ping-chung and several former Chin scholar-officials, was still the nominal overlord of the other Mongol dominions (*ulus*) in Asia. By then, however, his Chinese entourage had persuaded him to accept the role of a traditional Chinese emperor. A decisive step was taken in 1271 when the Chinese dominion was given a Chinese dynastic name—Ta Yüan, the "Great Origin." Before this the Chinese name for the Mongol state was Ta Ch'ao ("Great Dynasty"), introduced *c.* 1217. It was a translation of the Mongol name Yeke Mongghol Ulus ("Great Mongol Nation") adopted by Genghis Khan *c.* 1206. The new name, however, was a departure from Chinese traditions. All earlier Chinese dynasties were named after ancient feudal states or geographic terms; even the Khitan and the Juchen had followed this tradition by naming their states Liao (after the Liao Ho in Manchuria) and Chin ("Gold," after a river in Manchuria that had a Juchen name with that meaning). Yüan was the first nongeographic name of a Chinese dynasty since Wang Mang established the Hsin dynasty (AD 9–25).

Foundation of the Yüan dynasty

During the 1260s the central bureaucracy and the local administration of the Chinese empire were remodeled on Chinese lines, with certain alterations introduced by the Chin, or Juchen, state. The Central Secretariat remained the most important civilian authority, with specialized agencies such as the traditional six ministries of finance, war, officials, rites, punishments, and public works. The Shu-mi yüan (Military Council) was another institution inherited from previous dynasties. A Yü-shih t'ai (Censorate) was originally created for remonstrations against the emperor and criticism of policies, but it became more and more an instrument of the court itself and a tool to eliminate other members of the bureaucracy. In the main, the territorial divisions followed Chinese models, but the degree of local independence was much smaller than it had been under the Sung; the provincial administrations were actually branches of the Central Secretariat. The structures of the various provincial administrations throughout China were smaller replicas of the Central Secretariat. According to Chinese sources, in 1260–61 the lower echelons in the Central Secretariat were mostly Chinese; the high offices, however, even if they had traditional Chinese names, were reserved for non-Chinese. Surprisingly, Kublai Khan had few Mongols in high administrative positions; apparently suspicious of some of his tribal leaders, he preferred absolute foreigners. The military sphere was affected least by the attempts to achieve a synthesis between Chinese and native ways of life; there the Mongol aristocracy remained supreme.

Too many antagonistic social and ethnic groups existed within the Yüan government to secure a stable rule. The traditional Chinese value system had largely disappeared, and no political ethics had replaced it. While personalized loyalty focused on the ruler, the companionship of *nökör* relations was not enough to amalgamate the heterogeneous ruling group into a stable body. This unbalanced system of government could function only under a strong ruler; under a weak or incompetent emperor disintegration was certain, and a decline in efficiency resulted.

Problems of non-assimilation

The former scholar-officials of China remained to a great extent outside the governmental and administrative structure; only minor positions were open to them. The Mongols never made full use of the administrative potential of the scholar-officials, fearing their competence and abilities. The ruling foreign minority in China was more an elite of the colonialist type than a part of the Chinese social system.

The unwillingness of the Mongols to assimilate with the Chinese is shown by their attempts to cement the inequalities of their rule. After the Sung Empire had been conquered, the population of China was divided into four classes. The first class was the Mongols themselves, a tiny but privileged minority. Next came the *se-mu jen* ("persons with special status"), such confederates of the Mongols as Turks or Middle Eastern Muslims. The third group was called the *han-jen* (a term that generally means Chinese but that was used to designate the inhabitants of only northern China); this class included the Chinese and other ethnic groups living in the former Chin state, as well as Hsi Hsia, Juchen, Khitan, Koreans, Po-hai, and Tangut, who could be employed in some functions and who also formed military units under Mongol leadership. The last group was the *man-tzu,* an abusive term in Chinese, meaning "southern barbarian," which designated the former subjects of Sung China, about three-fourths of the Chinese empire. The lowest stratum in Yüan China was the slaves, whose numbers were quite considerable. The slave status was hereditary, and only under certain conditions could a slave be freed.

More than 80 percent of the taxpayers came from the *man-tzu* group, which was generally barred from holding higher office (only rarely would one of them rise to some prominence). The Mongols and the *se-mu jen* were tax-exempt and enjoyed the protection of the law to a higher degree than did the *han-jen* and *man-tzu.*

The formal distinction between various ethnic groups and the corresponding graded status was no Mongol invention but a social differentiation inherited from the Chin state. In the same way, many institutions were taken over from the Chin. Law in Yüan China was based partly on the legislation of the Chin and partly on traditional Chinese law; but Mongol legal practices and institutions also played a great role, particularly in penal law. The Yüan legal code has been preserved in the dynastic history, *Yüan shih,* and other sources. In addition, many rules, ordinances, and decisions of individual cases are collected in such compilations as *Yüan tien-chang,* which throw much light not only on the legal system but also on social conditions in general.

Mongol and Chinese dualism was also reflected in the problem of administrative documents and languages. Few of the ruling Mongols, even in the later years of the Yüan, knew Chinese, and the number who mastered the Chinese script was still smaller. On the other hand, very few Chinese bothered to learn the language of the conquerors. Administration and jurisdiction, therefore, had to rely largely on interpreters and translators. Mongol was the primary language; most decisions, ordinances, and decrees were originally drafted in Mongol, and a Chinese interlineary version was added. This Chinese version was in the colloquial language instead of the formal documentary style, and it followed the Mongol word order so that these Chinese versions must have seemed barbaric to the native literati. Many of these Chinese versions have survived in such collections as *Yüan tien-chang.*

*Economy.* The Mongol conquest of the Sung Empire had for the first time since the end of T'ang reunified all China. Sung China had traded with its neighbours, the Liao and Chin, but trade had been strictly controlled and limited to authorized border markets. The Mongol conquest therefore reintegrated China's economy. The Mongol administration, in its desire to utilize the resources of the former Sung territory, the most prosperous part of China, tried to promote internal trade and aimed at a fuller integration of north and south. The region around the capital was dependent on grain transports from the south, and large quantities of food and textiles were needed to keep the Mongol garrisons. The Grand Canal, which

Reintegration of internal trade

had linked the river systems of the Yangtze, the Huai, and the Huang since the early 7th century, was repaired and extended to Ta-tu in 1292–93 with the use of corvée under the supervision of a distinguished Chinese astronomer and hydraulic engineer, Kuo Shou-ching—an action entirely within Chinese tradition. This was preceded, however, by another measure in the field of economic communications that was unorthodox in Chinese eyes: in about 1280, concessions for grain transport overseas were granted to some private Chinese entrepreneurs from the southeastern coastal region (some Chinese government officials were traditionally antagonistic toward private trade and enterprise, an attitude that the ruling Mongols did not share). These private shipowners transported in their fleets grain from the lower Yangtze region to northern Chinese harbours and from there to the capital. Early in the 14th century, however, these private fleet owners, who had made huge fortunes, were accused of treason and piracy, and the whole action was abolished. The Mongol government never replaced them with government fleets.

Another factor that contributed to the flourishing internal trade in China was the unification of currency. The Sung and Chin had issued paper money but only in addition to bronze coins, which had remained the basic legal tender. The Yüan government was the first to make paper money the only legal currency throughout the empire (1260). This facilitated financial transactions in the private sector as well as in the state treasuries. As long as the economy as such remained productive, the reliance on paper money as the basic currency had no detrimental effects. Only when the economy began to disintegrate under the last Mongol ruler did the paper money become gradually valueless and inflation set in. One reason for the paper currency might have been that much bronze and copper was used for the Buddhist cult and its statues, another that metal ores in China proper were insufficient to supply enough coins for some 80,000,000 people.

**Religious and intellectual life.** The Mongols did not try to impose their own religion (a cult of Heaven, the forces of nature, and shamanistic practices) on their subjects. This gave comparative freedom to the existing religions in China. These included what the Mongol rulers considered to be the *san chiao* ("three teachings"): Taoism, Buddhism, and Confucianism. Both Taoism and Buddhism retained their distinctive identities and organizations; although they often rivaled each other, they were not mutually exclusive. The Neo-Confucianism of the Chu Hsi school enjoyed orthodox status after the 1310s, but adherents of the three teachings interacted philosophically and intellectually in a way that popularized the "amalgamation" of the three schools among the common people and literati, if not the foreign residents, of China.

Mongol religious policies

*Taoism.* Under the Chin dynasty several popular Taoist sects had flourished in northern China, and Genghis Khan had apparently been impressed by the Taoist patriarch Ch'ang-ch'un. In 1223 Genghis Khan granted to Ch'ang-ch'un and his followers full exemption from taxes and other duties demanded by the government; this was the first of a series of edicts granting special privileges to the clergy of the various religions in China.

For some time it seemed as if Chinese Taoism would win favour with the Mongol rulers at the expense of Chinese Buddhism. The Buddhists, however, also profited from the open-minded attitude at the court; they tried to win influence within the Imperial family, prompted by the fact that many Buddhist institutions had been occupied by the Taoists, who relied on Mongol favour. Under the grand khan Möngke, several discussions were held between the Taoist and Buddhist clergy (1255–58), ending in a ruling that the former Buddhist temples should be returned to their original purpose. Imperial orders also outlawed some apocryphal Taoist texts, in which Buddhism was presented as a branch of Taoism and the Buddha as a reincarnation of Lao-tzu, the founder of Taoism. But Taoism as such continued to exist under the Yüan, and the fiscal privileges originally granted to the Taoist followers of Ch'ang-ch'un were extended on principle to all clergies.

*Buddhism.* The spokesmen of Chinese Buddhism under the early Mongol rulers came from the Ch'an sect (meditation Buddhism). Their high intellectuality and refined aestheticism, however, did not appeal to the Mongols, who felt more attracted by the mixture of magic practices, rather nebulous metaphysics, and impressive symbolism in the visual arts of Tibetan Buddhism. Kublai Khan appointed a young Tibetan lama known by the honorific name of 'Phags-pa as Imperial teacher (*ti-shih*); 'Phags-pa became the head of the Buddhist church in all Mongol dominions, including China. A special government agency to deal with Buddhism was established in 1264 and served as a sort of bureau for the Imperial teacher; it was in charge not only of Buddhist affairs in general but also of Tibetan affairs, although Tibet remained outside the administration of China proper, and no Mongol garrisons were ever established in Tibet. Tibetan politicians had thus succeeded in winning over the Mongol court and in retaining a more than nominal independence.

After the conquest of Sung China, a special agency for the supervision of Buddhism in southern China was established and placed under the control of another Tibetan lama. There thus existed two supervisory offices for Buddhism—one in Ta-tu for northern China and Tibet and one in Lin-an for southern China. The southern office caused great resentment among Chinese Buddhists and the population at large by its brutal and avaricious procedures, seizures of property, and extortions from the population. Throughout the Yüan dynasty, complaints continued against the arrogant behaviour of Tibetan lamas. (Under the last emperor, Togon-temür, Tibetan clerics introduced the court to sexual rites calling for intercourse with consecrated females—practices not unfamiliar in Indian and Tibetan cultures but shocking to the Chinese elite.)

Complaints against Tibetan lamas

Although Buddhism had won a victory among the ruling minority of China, it was a foreign rather than a Chinese Buddhism. The national varieties of Buddhism, especially Ch'an Buddhism, continued to exist, and monasteries in southern China sometimes became islands of traditional civilization where monks and lay Buddhists alike cultivated poetry, painting, and all the intellectual pastimes of the Chinese literati class; but, on the whole, Chinese Buddhism suffered from the general conditions in the Yüan Empire. The exemption from taxes and corvée attracted many persons to monastic life for purely utilitarian reasons, and, the more society disintegrated, the more people sought refuge behind the monastery walls. In about 1300 the number of monks throughout China was estimated at 500,000, and it must have grown during the last decades of Mongol rule. Monks played a great role in the rebellions to which the Yüan Empire eventually succumbed; also, the first Ming emperor had been a monk for some time.

*Foreign religions.* Tibetan Buddhism always remained outside Chinese civilization, as did other imported religions. A certain number of Muslims came to China, all from the Middle East or from Central Asia. The Turkic Öngüt tribe was largely Nestorian Christian. Many tombstones have been preserved with a bilingual Turkic and Chinese inscription, but none of these believers seems to have been Chinese by origin; a census taken about 1300 in Chen-chiang (in the present-day province of Kiangsu) lists the Nestorians together with foreign nationalities. The number of Nestorian Christians in China was so great that in 1289 a special agency for their supervision was established in Ta-tu. Manichaeism, which had spread to China under the T'ang, became extinct under the Yüan as an organized religion; but some Manichaean communities were probably absorbed by messianic Buddhist sects such as the White Lotus sect, a group that attracted many followers among the Chinese lower classes.

*Confucianism.* Confucianism was perceived by the Mongols as a Chinese religion, and it had mixed fortunes under their rule. The teachings of the Neo-Confucian school of Chu Hsi from the Sung period were introduced to the Mongol court at Chung-tu in the late 1230s but were confined to limited circles there and in North China. Confucian scholars enjoyed the benefits extended to the clergy of all religions, but they were dealt a strong blow by the discontinuation of the literary examinations following the Mongol conquest. For many centuries the examinations, based on Confucian texts, had been the basis for the selec-

tion of officials and for their privileged position within the state and society. After Kublai's accession, Confucianism had a more cordial reception at the Mongol court through the efforts of Chinese advisers like Liu Ping-chung and the great Confucian master Hsü Heng. Under their stewardship a certain Confucianization took place in government and education. Chinese rituals were performed for a while in the dynastic temple (*t'ai-miao*), erected in Chung-tu in 1263. State sacrifices were offered to Confucius, and the study of the Classics was encouraged. Many of the rites observed at the court that were either Tibetan Buddhist or inherited from the Mongol nomadic past were continued, however. The emperor Buyantu (ruled 1311–20), one of the most Sinicized Mongol rulers, reintroduced the examination system in 1313, but it remains doubtful how well the examinations functioned. They certainly did not guarantee an official career, as those under the Sung and, to a certain extent, under the Chin had done.

Examination system reintroduced

The system of the Yüan, as introduced in 1313, provided different types of curricula for Mongols, other foreigners (*se-mu jen*), and Chinese; also, the requirements were different—Chinese had to show their complete mastery of the curriculum, whereas Mongols and other foreigners had to give only a mediocre performance. This inequality was even formalized for the candidates who were to be admitted to the state academy (*kuo-tzu chien*). The first examinations were held in the presence of the Emperor in 1315, and, of the 300 persons granted the title of doctor (*chin-shih*), 75 were Mongols, 75 were other foreigners, 75 were northern Chinese (*han-jen*), and 75 came from southern China; they all received official positions within the bureaucracy, Mongols the higher and Chinese the lesser posts. The positions of power within the hierarchy remained in the hands of the Mongols and other foreigners.

Under Buyantu, for the first time, the interpretation and commentaries of the Neo-Confucian school were made obligatory. This cemented Neo-Confucian ideology not only among the Chinese literati who wished to pass an examination but also for future generations. Chinese Confucian orthodoxy from the 14th to the 19th century therefore rested largely on the foundations it had received under the Yüan. In spite of all this, Classical scholarship under the Yüan did not produce a single remarkable work but struggled under an adverse political and intellectual climate. Striving to preserve their sacred tradition, the Confucian scholars were content with expounding the doctrines laid down by the Sung philosophers, seeking to harmonize the different philosophical issues and points of view rather than exploring new horizons.

*Literature.* Chinese literature also shows conservative tendencies. Poetry composition remained a favourite pastime of the educated class, including the Sinicized scholars of Mongol, Central Asian, and western Asian origins, but no great works or stylistic innovations were created. During the last chaotic decades of the Yüan, some notable poets emerged, such as the versatile Yang Wei-chen and the bold and unconventional Kao Ch'i. Many prose works dealing with contemporary events and persons were written under the Yüan, but these were notable for their content, not their literary merit. Surprisingly harsh criticism and satire against the Mongols and also undisguised Sung loyalism found open expression, presumably because the Mongols were uninterested in what the Chinese wrote in Chinese and, moreover, were mostly unable to read it. Some writers collected rare or interesting and piquant items and transmitted many aspects of Sung culture to future generations. The lament for the refinement and grandeur of the Sung is a constant theme in Yüan writings.

Criticism and satirization of the Mongols

During the early Yüan period the traditional Chinese official historiography was restored under the charge of the Hanlin Academy, which sponsored the compilation of the official dynastic histories of the Sung, Liao, and Chin states conquered by the Mongols and undertook the compilation of the reign chronicles (*shih-lu*) and other governmental compendiums. The major achievement of official historiography was the compilation (1329–33) of the *Ching-shih ta-tien,* a repository of 800 *chüan* (chapters) of official documents and laws, but the text is now lost. Private historiography, especially works on the events of the Sung, fared rather poorly under the Yüan because of the adverse political and intellectual climate. The most distinguished contribution was written by Ma Tuan-lin and titled *Wen-hsien t'ung-k'ao* ("Comprehensive Survey of Literary Remains"). It was an encyclopaedic documentary history of Chinese institutions from the earliest times down to the middle reign of the Nan Sung dynasty.

Popular literature

In urban society a literature in the vernacular language began to flourish, untrammeled by rigid norms of formalistic or ideological orthodoxy. Novels and stories were written for the amusement of a wide-reading public. And dramatic literature reached such a peak in Yüan China that later literary criticism regarded the Yüan as the classical age for operatic arias (*ch'ü,* a word that also means a full opera, with arias and chanted recitativo). The collection *Yüan ch'ü hsüan* ("Selection from Yüan Operas"), with 100 opera librettos, and the storyteller "prompt books" for such dramatized historical romances as *San-kuo* ("Three Kingdoms") give ample evidence for the creativity and vitality of Chinese dramatic literature. This phenomenon may perhaps be considered as evidence that under the Yüan a certain urbanization took place and something like a bourgeoisie emerged, because dramatic literature and colloquial novels found their clientele chiefly among the merchant and artisan classes.

Foreigners, chiefly of Turkic or Persian origin, also contributed to Chinese literature under the Yüan. They wrote poetry and painted in the Chinese way in order to distinguish themselves in fields where they could gain prestige among the educated Chinese. All the foreigners who wrote in Chinese seem to have avoided any reference to their foreign origin or creed. Nothing, in fact, could be more Chinese than their productions. Even foreigners who, like the Persians, came from a country with a considerable literary tradition of its own never attempted to introduce their native forms, subject matter, or religions. No literary symbiosis seemed possible, and, although China was exposed to more external influences under the Yüan than ever before, Chinese literature shows little effect from such contacts with the outside world. It is perhaps symptomatic that under the Yüan no literary works from other civilizations were translated into Chinese and that practically no translations of Chinese Classical and historical works into Mongol have survived. There seemed to be only the alternatives of complete rejection of Chinese civilization, as practiced by most Mongols, or wholesale absorption by Chinese culture. (H.Fr./H.Ch.)

*The arts.* Conservatism played a dominant role in the arts during the Mongol period. In sponsored arts such as sculpture and ceramics, the Mongols' desire to lay claim to the Chinese Imperial heritage was not complemented by any strong artistic vision of their own, and conservatism meant mere perpetuation. Sung, Liao, and Chin ceramic types were continued, often altered only by increased bulk, while the great artistic achievement of the era, blue-and-white ware, probably derived from non-Imperial sources. Government-sponsored Buddhist sculpture often attained high artistic standards, preserving the realism and powerful expression of T'ang and Sung traditions, while in the finest sculpture of the time, such as the reliefs at the Chüyung Gate north of Ta-tu (1342–45), this was combined with a flamboyant surface decor and striking dramatization better suited to foreign taste than to the increasingly restrained Chinese aesthetic.

Conservatism also tempered the private arts of calligraphy and painting: the scholar-amateurs who produced them felt impelled to preserve their heritage against a perceived barbarian threat. Conservatism, however, often took the form of a creative revival that combed the past for sources of inspiration and then artistically transformed them into a new idiom. In calligraphy, Chao Meng-fu gave new impetus to the 4th-century style of Wang Hsi-chih, which then became a standard for Chinese writing and book printing for centuries. In painting, Chao and his contemporary Ch'ien Hsüan helped to complete the development of a distinctively amateur style that ushered in a new phase in the history of Chinese painting. Their work did not continue that of the previous generation but ranged widely over the available past tradition, and past

Works of Chao Meng-fu

styles rather than observed objects became the subject of artistic interpretation. The naturalism of Sung painting gave way to calligraphically inspired abstractions. Paintings became closely linked in style to the written inscriptions that appeared upon them with increasing frequency and prominence. Skillful professional techniques and overt visual attractiveness were avoided, replaced by deliberate awkwardness and an intellectualized flavour. Their works were done for private purposes, often displaying or concealing personal and political motives, to be understood only by fellow literati through the subtle allusions of their subject matter, stylistic references, or inscriptions.

Naturalistic painting styles also continued in popularity throughout the first two-thirds of the period, painted by such important artists as Li K'an and Jen Jen-fa. Perpetuating northern traditions of the T'ang and Sung periods, these styles were practiced chiefly by scholar-officials associated with the court at the capital. Several members of the Mongol royal family became major patrons or collectors of such conservative styles, although Imperial patronage remained slight in comparison with earlier periods.

In the latter third of the dynasty, with a sharp decline in the practice of painting by scholar-officials and northerners, Yüan painting was increasingly represented by the innovative approach of Chao Meng-fu as practiced by reclusive scholars from the Su-chou–Wu-hsing area. Four of these—the landscape painters Huang Kung-wang, Wu Chen, Ni Tsan, and Wang Meng—transformed and blended certain elements from the past into highly personal, easily recognizable styles and later came to be known as the Four Masters of the Yüan dynasty. In the early Ming period the Hung-wu emperor decimated the Su-chou literati and with it Su-chou painting, but by the end of the 15th century Su-chou artists came once again to dominate Chinese painting, and the styles of the Four Masters became the most influential of all painting models in later Chinese history. (Je.Si.)

The Four Masters of the Yüan dynasty

**Yüan China and the West.** As mentioned above, Mongol rulers favoured trade in all their dominions. In China, too, they eliminated state trade controls that had existed under the Sung and Chin, so that internal and external trade reached unprecedented proportions. It seems, however, that China's transcontinental trade with the Middle East and Europe was in the hands of non-Chinese (Persians, Arabs, Syrians). Silk, the Chinese export commodity par excellence, reached the Middle East and even Europe via the caravan routes across Asia; Chinese ceramics were also exported, chiefly into the Islāmic countries. The Asian countries concentrated their European trade largely with the Italian republics (*e.g.,* Genoa, Venice). To the Italians, trade with the East was so important that the *Practica della mercatura,* a handbook on foreign trade, included the description of trade routes to China.

Contacts with the Middle East and Europe

Direct contacts between China and Europe were insignificant, however, in spite of China's being a part of an empire stretching from Ta-tu to southern Russia. Chinese historical and geographic literature had little to say about the European parts of the Mongol Empire; in the official dynastic history of the Yüan, references to foreign countries are limited to countries such as Korea, Japan, Nam Viet, Burma, and Champa, with which China had carried on trade or tributary relations for centuries, and there are some scattered data on Russia. For some time a Russian guards regiment existed in Ta-tu, and some Russian soldiers were settled in military colonies in eastern Manchuria. As a whole, however, the civilizations of Europe and China did not meet, although contacts were made easy; Europe remained for the Chinese a vague region somewhere "beyond the Uighur."

More important were the contributions from the Muslim countries of the Middle East, chiefly in the fields of science and technology. During the reign of Kublai Khan, Arab-Persian astronomy and astronomical instruments were introduced into China, and the Chinese astronomer Kuo Shou-ching operated an observatory. Nevertheless, the basic conceptions of astronomy remained Chinese, and no attempt was made to adopt the Middle Eastern mathematical and theoretical framework. Similarly, Middle Eastern physicians and surgeons practiced successfully in China, but Chinese medical theory remained uninfluenced by Western practices. In geography a Chinese world map of the 14th century incorporates Arabic geographical knowledge into the Chinese worldview. It shows not only China and the adjacent countries but also the Middle East, Europe, and Africa; the African continent is already given in its actual triangular shape. But this knowledge probably never spread beyond a limited circle of professional geographers, and it is certain that the Sinocentric world conception continued unchallenged under the Yüan dynasty; no curiosity of what lay beyond the Chinese borders was aroused. For the countries to be reached by sea (such as Southeast Asian countries and India), Chinese works of the Yüan offer only a poor extract from the Sung work *Chu-fan chih* ("Description of the Barbarians") of *c.* 1225.

The situation is different regarding European knowledge of China. The Mongol advance into eastern Europe had given Europeans an acute awareness that actual people lived in regions hitherto shrouded in vague folkloric legends and myths; the Islāmic world had become a reality to Europeans with the first Crusades. It was, therefore, only natural that the Roman Catholic Church looked for potential converts among the obviously non-Muslim people of Asia. After Franciscan envoys brought back information on what was known as Cathay (China) in the mid-13th century, Pope Nicholas IV, a former Franciscan, dispatched a Franciscan mission to the court of the Grand Khan in Ta-tu (known in Europe as Cambaluc). The missionaries formed the nucleus of a Catholic hierarchy on Chinese soil: Cambaluc became the seat of an archbishopric, and in 1323 a bishopric was established in Ch'üan-chou. A famous Franciscan missionary was Odoric of Pordenone, who traveled in China in the 1320s; his reports—together with letters written by other Catholic missionaries—brought firsthand information on China to medieval Europe and today throw some light on the earliest missionary work in China. The Franciscan mission, which had to compete with the Nestorian clergy, was carried on more by the foreigners in China than by the Chinese themselves. The friars preached in Tatar, which means either Mongol or Turkic, and apparently won no Chinese converts. Significantly, no Chinese source mentions the activities of these missionaries; the Chinese probably regarded the Franciscans as one of the many strange, foreign sects, perhaps an outlandish variety of Buddhism. Archaeological evidence of the presence of Europeans and of Roman Catholicism has been discovered only in modern times, in Yang-chou (in present-day Kiangsu), on a Latin tombstone dated 1342 and recording the death of an Italian lady whose name suggests some relation to a Venetian family engaged in trade with Asia.

European knowledge of China

Only the last direct contact between the papal see and Yüan China can be corroborated by both Western and Chinese sources. In 1336 a group of Alani Christians in Ta-tu sent a letter to Pope Benedict XII, who sent John of Marignola with a mission to the Mongol court. The mission reached the summer capital, Shang-tu, in 1342. Chinese sources have recorded the date of its audience as Aug. 19, 1342. The country from where the envoys came is given by the Chinese source as Fu-lang, a Chinese version of the name Farang (Franks), which was used as a general term for Europeans in the Middle East. The arrival of envoys from what must have seemed the end of the world so impressed the court that an artist was commissioned to paint a portrait of the battle horse that Marignola had brought as a present; this portrait was still extant in the 18th century but is now lost. Chinese literati wrote many eulogies on the portrait of the horse; the country of Fu-lang, however, did not interest the Chinese poets, and the whole embassy of Marignola is invariably described in terms that point to an unbroken Sinocentric attitude. Thus, the contact between the pontiff and the Mongol court remained without further consequence. The end of Mongol rule over China and the strong nationalism of the Ming dynasty also doomed the Catholic missions of the 14th century. The reports of Marco Polo, on the other hand, inaugurated for Europe the era of discoveries and created a new vision of the world, with China as a part.

Although China as a separate cultural entity was only

Chinese influence in Asia

realized dimly and gradually in the European West, Chinese influences spread under the Yüan dynasty to other parts of Asia. Chinese medical treatises were translated into Persian, and Persian miniature painting in the 13th and 14th centuries shows many influences of Chinese art. Chinese-type administration and chancellery practices were adopted by various Mongol dominions in Central Asia and the Middle East. It has even been suggested that the invention of gunpowder and of printing in Europe was due to a sort of stimulus diffusion from China, although a direct influence from China cannot be proved.

Chinese civilization itself remained very much what it had been before the Yüan dynasty, with a certain cultural isolationism a distinctive element. Neither the self-image of the Chinese nor China's position in the world changed very drastically. The change and challenges to which China was exposed under the Yüan, however, can explain many of the characteristic traits of Ming history.

**The end of Mongol rule.** The basic dilemma of Mongol rule in China—the Mongols' inability to achieve a durable identification with Chinese civilian institutions and to modify the military and colonialist character of their rule—became more apparent under Kublai's successors and reached a maximum under Togon-temür, the last Yüan ruler. Togon-temür was not unfriendly toward Chinese civilization, but this could not alter the contempt of many leading Mongols for Chinese civilian institutions. For centuries China had known clique factionalism at court, but this was mostly fought with political means; Mongol factionalism usually resorted to military power. Militarization gradually spread from the Mongol ruling class into Chinese society, and not a few dissatisfied Chinese leaders established regional power based on local soldiery. The central administration, headed by a weak emperor, proved incapable of preserving its supremacy.

Thus, the military character of Mongol rule paved the way for the success of Chinese rebels, some of whom came from the upper class, while others were messianic sectarians who found followers among the exploited peasantry. The Mongol court and the provincial administrations could still rely on a number of faithful officials and soldiers, and so the progress of the rebel movement in the 1350s and 1360s remained slow. But the rebel armies who had chosen what is now Nanking as their base took Ta-tu in 1368; the Mongol emperor fled, followed by the remnants of his overthrown government.

Fall of Ta-tu

The Mongols remained a strong potential enemy of China for the next century, and the Genghis Khan clan in Mongolia continued to regard itself as the legitimate ruler of China. The century of Mongol rule had some undesirable effects upon the government of China—Imperial absolutism and a certain brutalization of authoritarian rule, inherited from the Yüan, were features of the succeeding Ming government. Yet, Mongol rule lifted some of the traditional ideological and political constraints on Chinese society. The Confucian hierarchical order was not rigidly enforced as it had been under the T'ang and Sung, and the Mongols thereby facilitated the upward mobility of some social classes, such as the merchants, and encouraged extensive growth of popular culture, which had been traditionally downgraded by the literati. (H.Fr./H.Ch.)

## The Ming dynasty

### POLITICAL HISTORY

Ineptitude on the throne, bureaucratic factionalism at court, rivalries among Mongol generals, and ineffective supervision and coordination of provincial and local administration had gravely weakened the Yüan government by the 1340s. And in 1351 disastrous flooding of the Huang and Huai river basins aroused hundreds of thousands of long-oppressed Chinese peasants into open rebellion in northern Anhwei, southern Honan, and northern Hupeh provinces. Capitalizing on the breakdown of Yüan control, rebel movements spread rapidly and widely, especially throughout central China. By the mid-1360s large regional states had been created that openly flouted Yüan authority: Sung in the Huai Basin, under the nominal leadership of a mixed Manichaean-Buddhist secret-society leader named Han Lin-erh; Han in the central Yangtze Valley, under a onetime fisherman named Ch'en Yu-liang; Hsia in Szechwan, under an erstwhile general of the rebel Han regime named Ming Yü-chen; and Wu in the rich Yangtze Delta area, under a former Grand Canal boatman named Chang Shih-ch'eng. A onetime salt trader and smuggler named Fang Kuo-chen had simultaneously established an autonomous coastal satrapy in Chekiang. While Yüan chieftains contended with one another for dominance at the capital, Ta-tu (present-day Peking), and in the North China Plain, these rebel states to the south wrangled for survival and supremacy. Out of this turmoil emerged a new native dynasty called Ming (1368–1644).

**The dynasty's founder.** Chu Yüan-chang, founder of the new dynasty, came from a family of physiognomists from K'ai-feng who in Yüan times had deteriorated into itinerant tenant farmers in northern Anhwei Province. Orphaned by famine and plague in 1344, young Chu was taken into a small Buddhist monastery near Feng-yang city as a lay novice. For more than three years he wandered as a mendicant through the Huai Basin; then he studied for the Buddhist priesthood in his monastery. In 1352, after floods, rebellions, and Yüan campaigns against bandits had devastated and intimidated the whole region, Chu was persuaded to join a Feng-yang branch of Han Lin-erh's uprising. He quickly made himself the most successful general on the southern front of the rebel Sung regime, and in 1356 he captured and set up his headquarters in Nanking, a populous and strategically located city on the Yangtze River. There he began assembling a rudimentary government and greatly strengthened his military power. Between 1360 and 1367, still nominally championing the cause of the Sung regime, his armies gained control of the vast central and eastern stretches of the Yangtze Valley, absorbing first the Han domain to the west of Nanking and then the Wu domain to the east. He also captured the Chekiang coastal satrap, Fang Kuo-chen. Chu then announced his intention of liberating all China from Mongol rule and proclaimed a new dynasty effective with the beginning of 1368. The dynastic name Ming, meaning "Brightness," reflects the Manichaean influence in the Sung-revivalist Han Lin-erh regime under which Chu had achieved prominence. Chu came to be known by his reign name, the Hung-wu ("Vastly Martial") emperor.

Background of the founding emperor

Vigorous campaigning in 1368 drove the Mongols out of Shantung, Honan, and Shansi provinces and from Ta-tu itself, which was occupied by Ming forces on September 14, and simultaneously extended Ming authority through Fukien and Hunan into Kwangtung and Kwangsi provinces on the south coast. In 1369–70 Ming control was established in Shensi, Kansu, and Inner Mongolia; and continued campaigning against the Mongols thereafter extended northwestward to Ha-mi (1388), northeastward to the Sungari River in Manchuria (1387), and northward into Outer Mongolia beyond Karakorum, almost to Lake Baikal (1387–88). In operations to the west and southwest, Ming forces destroyed the rebel Hsia regime in Szechwan in 1371, wiped out major Mongol and aboriginal resistance in Kweichow and Yunnan in 1381–82, and pacified aboriginal tribesmen on the Sino-Burmese border in 1398. Thus, by the end of the Hung-wu emperor's 30-year reign in 1398, his new dynasty controlled the whole of modern China proper and dominated the northern frontier regions, from Ha-mi through Inner Mongolia and into northern Manchuria.

**The dynastic succession.** The Ming dynasty, which encompassed the reigns of 16 emperors, proved to be one of the stablest and longest dynasties of Chinese history. Rulers of Korea, Mongolia, East Turkistan, Burma, Siam, and Nam Viet regularly acknowledged Ming overlordship, and at times tribute was received from as far away as Japan, Java and Sumatra, Ceylon and South India, the East African coast, the Persian Gulf region, and Samarkand. Modern Chinese honour the Ming emperors especially for having restored China's international power and prestige, which had been in decline since the 8th century. The Ming emperors probably exercised more far-reaching influence in East Asia than any other native rulers of China, and their attitude toward the representa-

tives of Portugal, Spain, Russia, Britain, and Holland who appeared in China before the end of their dynasty was a condescending one.

Ming reign names

For the first time in Chinese history the Ming rulers regularly adopted only one reign name (*nien-hao*) each; the sole exception was the sixth emperor, who had two reigns separated by an interval of eight years. Because of this reign-name practice, which was perpetuated under the succeeding Ch'ing dynasty, modern writers, confusingly but correctly, refer to the Wan-li emperor, for example, by his personal name, Chu I-chün, or as Ming Shen-tsung, or sometimes, incorrectly but conveniently, simply as Wan-li, as if the reign name were a personal name.

The Ming dynasty's founder, the Hung-wu emperor, was one of the strongest and most colourful personalities of Chinese history. His long reign established the governmental structure, policies, and tone that characterized the whole dynasty. After his death in 1398 his grandson and successor, the Chien-wen emperor, trying to assert control over his powerful uncles, provoked a rebellion on the part of the Prince of Yen and was overwhelmed in 1402. The Prince of Yen took the throne as the Yung-lo emperor (reigned 1402–24) and proved to be vigorous and aggressive. He subjugated Nam Viet, personally campaigned against the reorganizing Mongols in the north, and sent large naval expeditions overseas, chiefly under the eunuch admiral Cheng Ho, to demand tribute from rulers as far away as Africa. He also returned the empire's capital to Peking, giving that city its modern name.

For a century after the Yung-lo emperor, the empire enjoyed stability, tranquillity, and prosperity. But state administration began to suffer from exploitative domination of weak emperors by favoured eunuchs: Wang Chen in the 1440s, Wang Chih in the 1470s and 1480s, and Liu Chin from 1505 to 1510. The Hung-hsi (reigned 1424–25), Hsüan-te (1425–35), and Hung-chih (1487–1505) emperors were, nevertheless, able and conscientious rulers in the Confucian mode. The only serious disruption of the peace occurred in 1449 when the eunuch Wang Chen led the Cheng-t'ung emperor (first reign 1435–49) into a

Conflict with the western Mongols

disastrous military campaign against the Oyrat (western Mongols). The Oyrat leader Esen Taiji ambushed the Imperial army, captured the Emperor, and besieged Peking. The Ming defense minister, Yü Ch'ien, forced Esen to withdraw unsatisfied and for eight years dominated the government with emergency powers. When the interim Ching-t'ai emperor (reigned 1449–57) fell ill in 1457, the Cheng-t'ung emperor, having been released by the Mongols in 1450, resumed the throne as the T'ien-shun emperor (1457–64). Yü Ch'ien was executed as a traitor.

The Cheng-te (reigned 1505–21) and Chia-ching (1521–1566/67) emperors were among the less esteemed Ming emperors. The former was an adventure-loving carouser, the latter a lavish patron of Taoist alchemists. For one period of 20 years, during the regime of an unpopular grand secretary named Yen Sung, the Chia-ching emperor withdrew almost entirely from governmental cares. Both emperors cruelly humiliated and punished hundreds of officials for their temerity in remonstrating.

China's long peace ended in the Chia-ching emperor's reign. The Oyrat, under the vigorous new leadership of Altan Khan, were a constant nuisance on the northern frontier from 1542 on; in 1550 Altan Khan raided the suburbs of Peking itself. During the same era Japan-based sea raiders repeatedly plundered China's southeastern coast. Such sea raiders, a problem in Yüan times and from the earliest Ming years, had been suppressed in the reign of the Yung-lo emperor, when the Ashikaga shogunate offered nominal submission to China in exchange for generous trading privileges. But eventually changes in the official trade system provoked new discontent along the coast, and during the 1550s corsair fleets looted the Shanghai–Ning-po region almost annually, sometimes sending raiding parties far inland to terrorize cities and villages throughout the whole Yangtze Delta. Although coastal raiding was not totally suppressed, it was brought under control in the 1560s. Also in the 1560s Altan Khan was repeatedly defeated, so that he made peace in 1571. For the next decade, during the last years of the Lung-ch'ing emperor (reigned 1566/67–1572) and the early years of the Chia-ching emperor (1572–1620), there was a high level of governmental stability. The court was dominated by the outstanding grand secretary of Ming history, Chang Chü-cheng, and capable generals such as Ch'i Chi-kuang restored and maintained effective military defenses.

In 1592, when Japanese forces under Toyotomi Hideyoshi invaded Korea, Ming China was still strong and responsive enough to campaign effectively in support of its tributary neighbour. But the Korean war dragged on indecisively until 1598, when Hideyoshi died and the Japanese withdrew. It made heavy demands on Ming resources and apparently precipitated a military decline in China.

The reign of the Wan-li emperor was a turning point of Ming history in other regards as well. Partisan wrangling among civil officials had flared up in the 1450s in reaction to Yü Ch'ien's dominance and again in the 1520s during a prolonged "rites controversy" provoked by the Chia-ching emperor on his accession; after Chang Chü-cheng's death in 1582, it became the normal condition of court life. Through the remainder of the Wan-li emperor's long reign a series of increasingly vicious partisan controversies absorbed the energies of officialdom, while the harassed emperor abandoned more and more of his responsibilities to eunuchs. The decline of bureaucratic discipline and morale continued under the T'ai-ch'ang emperor, whose sudden death after a reign of only one month in 1620 fueled new conflicts. The T'ien-chi emperor (reigned 1620–27) was too young and indecisive to provide needed leadership. In 1624 he finally gave almost totalitarian powers to his favourite, Wei Chung-hsien, the most notorious eunuch of Chinese history. Wei brutally purged hundreds of officials, chiefly those associated with a reformist clique called the Tung-lin party, and staffed the government with sycophants.

The threat from the Manchu

A new threat had meantime appeared on the northern frontier. The Manchu, quiet occupants of far eastern Manchuria from the beginning of the dynasty, were aroused in 1583 by an ambitious young leader named Nurhachi. During the Wan-li emperor's latter years they steadily encroached on central Manchuria. In 1616 Nurhachi proclaimed a new dynasty, and overwhelming victories over Ming forces in 1619 and 1621 gave him control of the whole northeastern segment of the Ming Empire, down to the Great Wall at Shan-hai-kuan.

The Ch'ung-chen emperor (reigned 1627–44) tried to revitalize the deteriorating Ming government. He banished Wei Chung-hsien but could not quell the partisan strife that was paralyzing the bureaucracy. The Manchu repeatedly raided within the Great Wall, even threatening Peking in 1629 and 1638. Taxes and conscriptions became more and more oppressive to the Chinese population, and banditry and rebellions spread in the interior. The Ming government became completely demoralized. Finally, a domestic rebel named Li Tzu-ch'eng captured the capital in April 1644, and the Ch'ung-chen emperor committed suicide. The Ming commander at Shan-hai-kuan accepted Manchu help in an effort to punish Li Tzu-ch'eng and restore the dynasty, only to have the Manchu seize the throne for themselves.

Ming loyalists ineffectively resisted the Manchu Ch'ing dynasty from various refuges in the south for a generation. Their so-called Nan (Southern) Ming dynasty principally included the Prince of Fu (Chu Yu-sung, reign name Hung-kuang), the Prince of T'ang (Chu Yü-chien, reign name Lung-wu), the Prince of Lu (Chu I-hai, no reign name), and the Prince of Kuei (Chu Yu-lang, reign name Yung-li). The loyalist coastal raider Cheng Ch'eng-kung (Koxinga) and his heirs held out on Taiwan until 1683.

## GOVERNMENT AND ADMINISTRATION

**Local government.** The Ming state system was built upon a foundation of institutions inherited from the T'ang and Sung dynasties and modified by the intervening dynasties of conquest from the north, especially the Yüan. The distinctive new patterns of social and administrative organization that emerged in Ming times persisted, in their essential features, through the Ch'ing dynasty into the 20th century.

Social organization

At local and regional levels the traditional modes and personnel of government were perpetuated in ad hoc fashion in the earliest Ming years, but as the new empire became consolidated and stabilized, highly refined control structures were imposed that—in theory and probably also in reality—eventually subjugated all Chinese to the throne to an unprecedented and totalitarian degree. The Ming law code, promulgated in final form in 1397, reinforced the traditional authority and responsibility of the paterfamilias, considered the basis of all social order. Each family was classified according to hereditary status, the chief categories being civilian, military, and artisan; and neighbouring families of the same category were organized into groups, for purposes of self-government and mutual help and surveillance. Civilians were grouped in "tithings" of 10 families, and these, in turn, were grouped in "communities" totaling 100 families, plus 10 additional prosperous households, which in annual rotation provided community chiefs who were intermediaries between the citizenry at large and the formal agencies of government. This system of social organization, called *li-chia* (later replaced by or coexistent with a local defense system called *pao-chia*), served to stabilize, regulate, and indoctrinate the populace under relatively loose formal state supervision.

As in earlier times, formal state authority at the lowest level was represented by court-appointed magistrates of districts (*hsien*), and each cluster of neighbouring districts was subordinate to a supervisory prefecture (*fu*) normally governed from and dominated by a large city. Government at the modern provincial (*sheng*) level, after beginnings in Yüan times, was now regularized as an intermediary between the prefectures and the central government. There were 13 Ming provinces, each as extensive and populous as modern European states: Shantung, Honan, Shansi, Shensi (incorporating modern Kansu), Szechwan, Hukuang (comprising modern Hupeh and Hunan), Kiangsi, Chekiang, Fukien, Kwangtung, Kwangsi, Kweichow, and Yunnan. Nam Viet was a 14th province from 1407 to 1428. The large regions dominated by the great cities Peking (in modern Hopeh) and Nanking (in modern Kiangsu and Anhwei) were not subordinated to provincial-level governments but for administrative supervision were "directly attached" (*chih-li*) to the capital establishments in those cities; they are normally referred to as the northern and southern metropolitan areas (Pei Chih-li and Nan Chih-li). Nanking was the Ming capital through 1420. Thereafter Peking was the capital, but Nanking retained special status as auxiliary capital.

Ming provincial governments consisted of three coordinate agencies with specialized responsibilities for general administration, surveillance and judicial affairs, and military affairs. These were the channels for routine administrative contacts between local officials and the central government.

**Central government.** In its early form the Ming central government was dominated by a unitary Secretariat. The senior executive official of the Secretariat served the emperor as a chief counselor, or prime minister. Suspected treason on the part of the chief counselor Hu Wei-yung in 1380 caused the Hung-wu emperor to abolish all executive posts in the Secretariat, thus fragmenting general administration authority among the six functionally differentiated, formerly subordinate ministries of Personnel, Revenue, Rites, War, Justice, and Works. This effective abolition of the Secretariat left the emperor as the central government's sole coordinator of any significance, strengthened his control over the officialdom, and, in the view of many later scholars, gravely weakened the Ming state system.

Especially prominent among other agencies of the central government was a Censorate, which was charged with the dual functions of maintaining disciplinary surveillance over the whole of officialdom and remonstrating against unwise state policies and improprieties in the conduct of the emperor. Equally prominent were five chief military commissions, each assigned responsibility, jointly with the Ministry of War, for a geographically defined segment of the empire's military establishment. There was originally a unitary Chief Military Commission paralleling the Secretariat, but in the 1380s its authority was similarly fragmented. The hereditary soldiers, who were under the administrative jurisdiction of the chief military commissions, originated as members of the rebel armies that established the dynasty, as surrendering enemy soldiers, in some instances as conscripts, or as convicted criminals. They were organized and garrisoned, principally along the frontiers, near the capital, or in other strategic places, but also throughout the interior, in units called guards and battalions. Whenever possible, such units were assigned state-owned agricultural lands so that, by alternating military duties with farm labour, the soldiers could be self-supporting. The military families, in compensation for providing soldiers in perpetuity, enjoyed exemptions from labour services levied by the state on civilian families. Each guard unit reported to its Chief Military Commission at the capital through a provincial-level Regional Military Commission. Soldiers from local guards were sent in rotation to the capital for special training or to the Great Wall or another area of comparable military importance for active patrol and guard duty. At such times, as on large-scale campaigns, soldiers served under tactical commanders who were on ad hoc duty assignments, detached from their hereditary posts in guard garrisons or higher echelons of the military service.

Military organization

**Later innovations.** In the 15th century new institutions were gradually devised to provide needed coordination both in the central government and in regional administration. Later emperors found the Hung-wu emperor's system of highly centralized power and fragmented government structure inefficient and inconvenient. Litterateurs of the traditional and prestigious Hanlin Academy came to be assigned to the palace as secretarial assistants, and they quickly evolved into a stable Grand Secretariat (Nei-ko) through which emperors guided and responded to the ministries and other central government agencies. Similarly, the need for coordinating provincial-level affairs led to the delegation of high-ranking central government dignitaries to serve as regional commanders (*tsung-ping kuan*) and governor-like grand coordinators (*hsün-fu*) in the provinces. Finally, clusters of neighbouring provinces came under the supervisory control of still more prestigious central government officials, known as supreme commanders (*tsung-tu*), whose principal function was to coordinate military affairs in extended, multiprovince areas. As the dynasty grew older, as the population expanded, and as administration became increasingly complex, coordinators proliferated even at subprovincial levels in the form of circuit intendants (*tao-t'ai*), who were delegated from provincial agencies as functionally specialized intermediaries with prefectural administrations.

Examinations for civil service

To an extent unprecedented except possibly in Sung times, Ming government was dominated by nonhereditary civil service officials recruited on the basis of competitive written examinations. Hereditary military officers, although granted ranks and stipends higher than their civil service counterparts and eligible for noble titles rarely granted to civil officials, always found themselves subordinate to policymaking civil servants except in the very early years of the dynasty. Members of the Imperial clan, except in the earliest and latest years of the dynasty, were forbidden to take active part in administration, and the Ming practice of finding Imperial consorts in military families effectively denied Imperial in-laws access to positions of significant authority. High-ranking civil officials usually could place one son each in the civil service by hereditary right, and, beginning in 1450, wealthy civilians often were able to purchase nominal civil service status in government fund-raising drives. But those entering the service in such irregular ways rarely had notable, or even active, careers in government. In the early decades of the dynasty, before competitive examinations could provide sufficient numbers of trustworthy men for service, large numbers of officials were recruited directly from government schools or through recommendations by existing officials; and such recruits often rose to eminence. But after about 1400, persons entering the civil service by avenues other than examinations had little hope for successful careers.

In a departure from traditional practices but in accordance with the Yüan precedent, there was only one type

of examination given in Ming times. It required a general knowledge of the Classics and history and the ability to relate Classical precepts and historical precedents to general philosophical or specific political issues. As in Yüan times, interpretations of the Classics by the Chu Hsi school of Neo-Confucianism were prescribed. By the end of the Ming dynasty the writing of examination responses had become highly stylized and formalized in a pattern called "the eight-legged essay" (*pa-ku-wen*), which in subsequent centuries became notoriously repressive of creative thought and writing.

Local schools

Beginning in the Hung-wu emperor's reign, the government sponsored district-level schools, in which state-subsidized students prepared for the civil service examinations. Especially talented students could be promoted from such local schools into programs of advanced learning and probationary service at a National University in the capital. Especially after 1500 there was a proliferation of private academies in which scholars gathered to discuss philosophy and students were also prepared for the examinations. Education intendants from provincial headquarters annually toured all localities, examining candidates who presented themselves and certifying those of "promising talent" (*hsiu-ts'ai*) as being qualified to undertake week-long examination ordeals that were conducted every third year at the provincial capitals. Those who passed the provincial examinations (*chü-jen*) could be appointed directly to posts in the lower echelons of the civil service. They were also eligible to compete in triennial metropolitan examinations conducted at the national capital. Those who passed were given degrees often called doctorates (*chin-shih*) and promptly took an additional palace examination, nominally presided over by the emperor, on the basis of which they were ranked in order of excellence. They were registered as qualified officials by the Ministry of Personnel, which assigned them to active-duty posts as vacancies occurred. While on duty they were evaluated regularly by their administrative superiors and irregularly by touring inspectors from the Censorate. It was normally only after long experience and excellent records in low- and middle-grade posts, both in the provinces and in the capital, that an official might be nominated for high office and appointed by personal choice of the emperor.

Although acceptance into, and success in, the civil service were the most highly esteemed goals for all and were nominally determined solely by demonstrated scholastic and administrative abilities, other factors inevitably intruded to prevent the civil service system from being wholly "open." Differences in the economic status of families made for inequalities of educational opportunity and, consequently, inequalities of access to civil service careers. The sons of well-to-do families clearly had advantages, and men of the affluent and cultured southeastern region so threatened to monopolize scholastic competitions that regional quotas for passers of the metropolitan examinations were imposed by the government, beginning in 1397. Once in the service one's advancement or even survival often depended on shifting patterns of favouritism and factionalism. Modern scholarship strongly suggests, nevertheless, that "new blood" was constantly entering the Ming civil service, that influential families did not monopolize or dominate the service, and that men regularly rose from obscurity to posts of great esteem and power on the basis of merit. Social mobility, as reflected in the Ming civil service, was very possibly greater than in Sung times and was clearly greater than in the succeeding Ch'ing era.

Weaknesses of Ming government

The Ming pattern of government has generally been esteemed for its stability under civil service dominance, its creativity in devising new institutions to serve changing needs, and its suppression of separatist warlords on one hand and disruptive interference by Imperial clansmen and palace women on the other. It suffered, however, from sometimes vicious factionalism among officials, from recurrences of abusive influence on the part of palace eunuchs, and from defects in its establishment of hereditary soldiers. The military system not only failed to achieve self-support but stagnated steadily, so that from the mid-15th century onward it had to be supplemented by conscripts and, finally, all but replaced by mercenary recruits. Most notoriously, the Ming state system allowed emperors to behave capriciously and abusively toward their officials. Despite their high prestige, officials had to accept being ignored, humiliated, dismissed, or subjected to bodily punishment and to risk being cruelly executed (sometimes in large numbers), as suited the Imperial fancy. Power was concentrated in the hands of the Ming emperors to a degree that was probably unparalleled in any other long-lived dynasty of Chinese history, and the Ming emperors often exercised their vast powers in abusive fashion.

## FOREIGN RELATIONS

Whereas in Ming times the Chinese organized themselves along wholly bureaucratic and tightly centralized lines, the Ming emperors maintained China's traditional feudal-seeming relationships with foreign peoples. These included the aboriginal tribes of south and southwest China, who often rose in isolated rebellions but were gradually being assimilated. The Chinese took for granted that their emperor was everyone's overlord and that de facto (mostly hereditary) rulers of non-Chinese tribes, regions, and states were properly his feudatories. Foreign rulers were thus expected to honour and observe the Ming ritual calendar, to accept nominal appointments as members of the Ming nobility or military establishment, and, especially, to send periodic missions to the Ming capital to demonstrate fealty and present tribute of local commodities. Tributary envoys from continental neighbours were received and entertained by local and provincial governments in the frontier zones. Those from overseas were welcomed by special maritime trade supervisorates (*shih-po ssu,* often called offices of trading ships) at three key ports on the southeast and south coasts: Ning-po in Chekiang (for Japanese contacts), Ch'üan-chou in Fukien (for contacts with Taiwan and the Ryukyu Islands), and Canton in Kwangtung (for contacts with Southeast Asia). The frontier and coastal authorities forwarded foreign missions to the national capital, where the Ministry of Rites offered them hospitality and arranged for their audiences with the emperor. All envoys received valuable gifts in acknowledgement of the tribute they presented. They also were permitted to buy and sell private trade goods at specified, officially supervised markets, both in the capital and on the coasts and frontiers. Thus, copper coins and luxury goods (notably silks and porcelains) flowed out of China, and pepper, other spices, and similar rarities flowed in. On the western and northern frontiers the principal exchange was in Chinese tea and steppe horses. On balance, the combined tribute and trade activities were highly advantageous to foreigners—so much so that the Chinese early established limits for the size and cargoes of foreign missions and prescribed long intervals that must elapse between missions.

Aims of foreign policy

The principal aim of Ming foreign policy was political: to maintain China's security and, especially, to make certain the Mongols could not threaten China again. To this end the Hung-wu emperor repeatedly sent armies northward and northwestward to punish resurgent Mongol groups and prevent any reconsolidation of Mongol power. The Yung-lo emperor was even more zealous: he personally campaigned into the Gobi (desert) five times, and his transfer of the national capital from Nanking to Peking, completed in 1421 after long preparations, was largely a reflection of his concern about the frontier. His successors, though less zealous than he in this regard, were vigilant enough so that the Great Wall was restored and expanded to its present-day extent and dimensions. Frontier defense forces, aligned in nine defense commands stretching from Manchuria to Kansu, kept China free from Mongol incursions, except for occasional raiding forays such as those by Esen Taiji and Altan Khan.

The fact that the Mongols could not reunite themselves was a fortunate circumstance for Ming China. As early as the Yung-lo emperor's time the Mongols were divided into three groups that were often antagonistic to one another. They were the so-called western Mongols or Oyrat (also known as Eleuthes, Kalmycks, or Dzungar), the eastern Mongols or Tatars, and a group in the Ch'engte area known as the Urianghad tribes. The Urianghad tribes surrendered to the Hung-wu emperor and were in-

corporated into China's frontier defense system under a Chinese military headquarters. Because they served the Yung-lo emperor as a loyal rear guard during his seizure of the throne, he rewarded them with virtual autonomy, withdrawing the Chinese command post from their homeland beyond the Great Wall. Subsequently, the Hsüan-te emperor similarly withdrew the command post that the Hung-wu emperor had established at the Mongols' old extramural capital, Shang-tu. These withdrawals isolated Manchuria from China proper, terminated active Chinese military control in Inner Mongolia, and exposed the Peking area in particular to the possibility of probing raids from the nearby steppes. They reflected an essentially defensive Chinese posture in the north, which by late Ming times allowed the Oyrat to infiltrate and dominate Ha-mi and other parts of the northwestern frontier, and the Manchu to rise to power in the northeast.

As for foreign peoples other than the Mongols, the Ming attitude was on the whole unaggressive: so long as they were not disruptive, the Ming emperors left them to themselves. The Hung-wu emperor made this his explicit policy. Even though he threatened the Japanese with punitive expeditions if they persisted in marauding along China's coasts, he dealt with the problem by building strong fortresses and coastal-defense fleets that successfully repulsed the marauders. He did send an army to subdue T'u-lu-p'an in 1377 when the Turko-Mongol rulers of that oasis region rebelled and broke China's traditional transport routes to the west. But he refused to intervene in dynastic upheavals in Nam Viet and Korea (where the Koryŏ fell to the Yi), and he was unmoved by the rise of the Turko-Mongol empire of Timur in the far west at Samarkand, even though Timur murdered Chinese envoys and planned to campaign against China.

'ribute-ollecting oyages f the 'ung-lo mperor's eign

The Yung-lo emperor was much more aggressive. He sent the eunuch admiral Cheng Ho on tribute-collecting voyages into Southeast Asia, the Indian Ocean, the Persian Gulf, and as far as East Africa. On one early voyage Cheng Ho intervened in a civil war in Java and established a new king there; on another, he captured the hostile king of Ceylon and took him prisoner to China. The Yung-lo emperor also reacted to turbulence in Nam Viet by sending an expeditionary force that incorporated the area into the Ming domain as a province in 1407.

After the Yung-lo era the Ming government reverted to the founding emperor's unaggressive policy toward foreign states. Nam Viet was abandoned in 1428 after protracted guerrilla-style resistance had thoroughly undermined Chinese control there. A new civil war in Nam Viet provoked the Chinese, after long and agonized discussion, to prepare to intervene there again in 1540; but the offer of ritual submission by a usurper gave the Chinese an opportunity to avoid war, and they welcomed it. On only two other occasions were Ming military forces active outside China's borders: in 1445–46 when Chinese troops pursued a rebellious border chief into Burma despite Burmese resistance, and in 1592–98 when the Ming court undertook to help Chosŏn (Korea under the Yi dynasty) repulse Japanese invaders, in a long and costly effort.

In order to preserve the government's monopolistic control of foreign contacts and trade, and, at least in part, to keep the Chinese people from being contaminated by "barbarian" customs, the Ming rulers prohibited private dealings between Chinese and foreigners and forbade any private voyaging abroad. The rules were so strict as to disrupt even coastal fishing and trading, on which large populations in the south and southeast had traditionally depended for their livelihood. Such unrealistic prohibitions were unpopular and unenforceable, and from about the mid-15th century Chinese readily collaborated with foreign traders in widespread smuggling, for the most part officially condoned. By late Ming times, also, thousands of venturesome Chinese had migrated to become mercantile entrepreneurs in the various regions of Southeast Asia and even in Japan. In efforts to enforce its laws the Ming court closed all maritime trade supervisorates except the one at Canton early in the 16th century, and by the 1540s it had begun to reinvigorate coastal defenses against marauders throughout the southeast and the south.

These circumstances shaped the early China coast experiences of the Europeans, who first appeared in Ming China in 1514. The Portuguese had already established themselves in southern India and at Malacca, where they learned of the huge profits that could be made in the regional trade between the China coast and Southeast Asia. Becoming involved in what the Ming court considered smuggling and piracy, the Portuguese were not welcomed to China; but they would not be rebuffed and by 1557 had taken control of a settlement at the walled-off end of a coastal peninsula (modern Macau) and were trading periodically at nearby Canton. In 1575 Spaniards from Manila visited Canton in a vain effort to get official trading privileges, and soon they were developing active though illegal trade on the Kwangtung and Fukien coasts. Representatives of the Dutch East India Company, after unsuccessfully trying to capture Macau from the Portuguese in 1622, took control of coastal Taiwan in 1624 and began developing trade contacts in nearby Fukien and Chekiang provinces. In 1637 a squadron of five English ships shot its way into Canton and disposed of its cargoes there. Russia, meanwhile, had sent peaceful missions overland to Peking, and by the end of the Ming dynasty the Russians' eastward expansion across Siberia had carried them finally to the shores of the Pacific north of the Amur River.

Missionary activities

Christian missionaries from Europe were handicapped by the bad reputation their trader countrymen had acquired in China, but the Jesuit tactic of accommodating to local customs eventually got the Jesuits admitted to the mainland. Matteo Ricci was the successful pioneer, beginning his work in 1583 well trained in the Chinese language and acquainted with Confucian learning. By the time of his death in 1610, despite hostility in some quarters, Jesuit communities were established in many cities of south and central China, a church had been built in Peking under Imperial patronage, and Christianity was known and respected by many Chinese scholar-officials. Before the end of the dynasty, Jesuits had won influential converts at court (notably the grand secretary Hsü Kuang-ch'i, or Paul Hsü), had produced Chinese books on European science as well as theology, and were manufacturing Portuguese-type cannon for Ming use against the Manchu. They also held official appointments in China's Directorate of Astronomy, which had the important responsibility of determining the official calendar. Both European technology and European ideas were beginning to have some effect on China, albeit still very limited.

## ECONOMIC POLICY AND DEVELOPMENTS

**Population.** Ming China's northward orientation in foreign relations was accompanied by a flow of Chinese migrants from the crowded South back into the vast North China Plain, and by a concomitant shift in emphasis from an urban and commercial way of life back to a rural and agrarian pattern. Thus, demographic and economic trends that had characterized China for centuries—the southward movement of population and the urbanization and commercialization of life—were arrested or even reversed.

Migration to the North

The North China Plain had been neglected since early Sung times, and its rehabilitation became a high-priority project of the early Ming emperors. The Ming founder's ancestral home was in the North, and his son, the Yung-lo emperor, won the throne from a personal power base in the newly recovered North at Peking. Securing the northern frontier was the major political goal of both these emperors, and both had reasons for being somewhat suspicious of Southerners and hostile toward them. In consequence, both emperors regularly moved well-to-do city dwellers of the Yangtze Delta region to northern towns for their cultural adornment, resettled peasants from the overpopulated southeast into the vacant lands of the North for their agrarian redevelopment, and instituted water-control projects to restore the productivity of the Huang and Huai river basins. (Notable among these were rehabilitation and extension of the Grand Canal, which reopened in 1415.) Colonists were normally provided with seeds, tools, and animals and were exempted from taxes for three years. The numerous army garrisons that were stationed in the North for defense of the frontier and of the

post-1420 capital at Peking were also given vacant lands to develop and were encouraged to become self-supporting. Such government measures were supplemented, following political reunification, by popular migration into the relatively frontier-like and open North. Rehabilitation of the North was no doubt also facilitated by the new availability of sorghum for dry farming. All these elements produced a substantial revival of the North. In Yüan times censuses credited the northern provinces with only one-tenth of the total Chinese population, but by the late 16th century they claimed 40 percent of the registered total; and until the late Ming years they were productive enough to sustain themselves as well as most of the large frontier defense forces. Suspension of government incentives late in the 15th century caused the northwest to enter into agrarian decline, and Shensi eventually became impoverished and bandit-infested. Support of the frontier defenses became an increasing burden on the central government.

During the migrations back to the North, the registered populations of the largest urban centres of the southeast declined. For example, between 1393 and 1578 Nanking declined from 1,193,000 to 790,000; Chekiang Province from 10,487,000 to 5,153,000; and Kiangsi Province from 8,982,000 to 5,859,000. Despite this leveling trend in the regional distribution of population, the South, and especially the southeast, remained the most populous, the wealthiest, and the most cultured area of China in Ming times. Such great southeastern cities as Nanking, Su-chou, and Hang-chou remained the major centres of trade and manufacturing, of entertainment, and of scholarship and the arts. Peking was their only rival in the North—solely because of its being the centre of political power.

Population changes

Although official census figures suggest that China's overall population remained remarkably stable in Ming times at a total of about 60,000,000, modern scholars have estimated that there was, in fact, substantial growth, probably to a total well over 100,000,000 and perhaps almost as high as 150,000,000 in the early 17th century. Domestic peace and political stability in the 15th century clearly set the stage for great general prosperity in the 16th century. This can be accounted for in part as the cumulative result of the continuing spread of early ripening rice and of cotton production—new elements that had been introduced into the Chinese economy in Sung and Yüan times. The introduction in the 16th century of food crops originating in America—peanuts (groundnuts), corn (maize), and sweet potatoes—created an even stronger agrarian basis for rapidly escalating population growth in the Ch'ing period.

**Agriculture.** Neofeudal land-tenure developments of late Sung and Yüan times were arrested with the establishment of the Ming dynasty. Great landed estates were confiscated by the government, fragmented, and rented out; and private slavery was forbidden. In the 15th century, consequently, independent peasant landholders dominated Chinese agriculture. But the Ming rulers were not able to provide permanent solutions for China's perennial land-tenure problems. As early as the 1420s the farming population was in new difficulties despite repeated tax remissions and other efforts to ameliorate its condition. Large-scale landlordism gradually reappeared, as powerful families encroached upon the lands of poor neighbours. Sung-style latifundia do not seem to have reemerged, but by the late years of the dynasty, sharecropping tenancy was the common condition of millions of peasants, especially in central and southeastern China—and a new gulf had opened between the depressed poor and the exploitative rich. The later Ming government issued countless pronouncements lamenting the plight of the common man but never undertook any significant reform of land-tenure conditions.

**Taxation.** The Ming laissez-faire policy in agrarian matters had its counterpart in fiscal administration. The Ming

Collection of land taxes

state took the collection of land taxes—its main revenues by far—out of the hands of civil service officials and entrusted this responsibility directly to well-to-do family heads in the countryside. Each designated tax captain was, on the average, responsible for tax collections in an area for which the land-tax quota was 10,000 piculs (one picul = 107 litres) of grain. In collaboration with the *li-chia* community chiefs of his fiscal jurisdiction, he saw to it that tax grains were collected and then delivered, in accordance with complicated instructions: some to local storage vaults under control of the district magistrate, and some to military units, which, by means of the Grand Canal, annually transported more than 3,000,000 piculs northward to Peking. In the early Ming years venal tax captains seem to have been able to amass fortunes by exploiting the peasantry. Later, however, tax captains normally faced certain ruin because tax-evading manipulations by large landlords thrust tax burdens increasingly on those least able to pay and forced tax captains to make up deficiencies in their quotas out of their personal reserves.

The land-tax rate was highly variable, depending not on the productivity of any plot but on the condition of its tenure, which might be as freehold or as one of several categories of land rented from the government. The land tax was calculated together with labour levies, or corvée, which, though nominally assessed against persons, were assessed against land in normal practice. Corvée obligations also varied widely and were usually payable in paper money or in silver rather than in actual service. Assessments against a plot of land might include several other considerations as well, so that a farmer's tax bill was a complicated reckoning of many different tax items. Efforts to simplify land-tax procedures in the 16th century, principally initiated by conscientious local officials, culminated in the universal promulgation of a consolidated-assessment scheme called "a single whip" (*i-t'iao-pien*) in 1581. Its main feature was the reduction of land tax and corvée obligations to a single category of payment in bulk silver or its grain equivalent. This reform was little more than a bookkeeping change at best, and it was not universally applied. Land-tax inequities were unaffected, and assessments rose sharply and repeatedly from 1618 to meet spiraling costs of defense.

Other revenues

Many revenues other than land taxes contributed to support of the government. Some, such as mine taxes and levies on marketplace shops and vending stalls, were based on proprietorship; others, such as salt taxes, wine taxes, and taxes on mercantile goods in transit, were based on consumption. Of all state revenues, more than half always seem to have remained in local and provincial granaries and treasuries; and of those forwarded to the capital, about half seem normally to have disappeared into the emperor's personal vaults. Revenues at the disposal of the central government were always relatively small. Prosperity and fiscal caution had resulted in the accumulation of huge surpluses by the 1580s, both in the capital and in many provinces; but thereafter the Sino-Japanese war in Chosŏn, unprecedented extravagances on the part of the long-lived Wan-li emperor, and defense against domestic rebels and the Manchu bankrupted both the central government and the Imperial household.

**Coinage.** Copper coins were used throughout the Ming dynasty. Paper money was used for various kinds of payments and grants by the government, but it was always nonconvertible and, consequently, lost value disastrously. It would, in fact, have been utterly valueless except that it was prescribed for the payment of certain types of taxes. The exchange of precious metals was forbidden in early Ming times, but gradually bulk silver became common currency, and after the mid-16th century government accounts were reckoned primarily in taels (ounces) of silver. By the end of the dynasty silver coins produced in Mexico, introduced by Spanish sailors based in the Philippines, were becoming common on the south coast.

Because during the last century of the Ming dynasty there emerged a genuine money economy, and because concurrently there developed some relatively large-scale mercantile and industrial enterprises under private as well as state ownership (most notably in the great textile centres of the southeast), some modern scholars have considered the Ming age one of "incipient capitalism" from which European-style mercantilism and industrialization might have evolved had it not been for the Manchu conquest and expanding European imperialism. It would seem clear, however, that private capitalism in Ming times flourished only insofar as it was condoned by the state, and it was

never free from the threat of state suppression and confiscation. State control of the economy—for that matter, of society in all its aspects—remained the dominant characteristic of Chinese life in Ming times as earlier.

## CULTURE

Role of the state in intellectual life

The predominance of state power also marked the intellectual and aesthetic life of Ming China. By requiring use of their interpretations of the Classics in education and in the civil service examinations, the state prescribed the Neo-Confucianism of the great Sung thinkers Ch'eng I and Chu Hsi as the orthodoxy of Ming times; by patronizing or commandeering craftsmen and artists on a vast scale, it set aesthetic standards for all the minor arts, for architecture, and even for painting; and by sponsoring great scholarly undertakings and honouring practitioners of traditional literary forms, the state established norms in these realms as well. It has consequently been easy for historians of Chinese culture to categorize the Ming era as an age of bureaucratic monotony and mediocrity. But the stable, affluent Ming society, in fact, proved irrepressibly creative and iconoclastic. Drudges by the hundreds and thousands may have been content with producing second-rate imitations or interpretations of T'ang and Sung masterpieces in all genres, but independent thinkers, artists, and writers were striking out in many new directions. The final Ming century, especially, was a time of intellectual and artistic ferment akin to the most seminal ages of the past.

**Philosophy and religion.** Taoism and Buddhism by Ming times had declined into ill-organized popular religions, and what organization they had was regulated by the state. State espousal of Chu Hsi thought, and state repression of noted early Ming litterateurs, such as the poet Kao Ch'i and the thinker Fang Hsiao-ju, made for widespread philosophical conformity during the 15th century. This was perhaps best characterized by the scholar Hsüeh Hsüan's insistence that the Way had been made so clear by Chu Hsi that nothing remained but to put it into practice. Philosophical problems about man's identity and destiny, however, especially in an increasingly autocratic system, rankled in many minds; and new blends of Confucian, Taoist, and Buddhist elements appeared in a sequence of efforts to find ways of personal self-realization in contemplative, quietistic, and even mystical veins. These culminated in the antirationalist individualism of the famed scholar-statesman Wang Yang-ming, who denied the external "principles" of Chu Hsi and advocated striving for wisdom through cultivation of the innate knowledge of one's own mind and attainment of "the unity of knowledge and action." Wang's followers carried his doctrines to extremes of self-indulgence, preached to the masses in gatherings resembling religious revivals, and collaborated with so-called mad Ch'an Buddhists to spread the notion that Confucianism, Taoism, and Buddhism are equally valid paths to the supreme goal of individualistic self-fulfillment. Through the 16th century intense philosophical discussions were fostered, especially in rapidly multiplying private academies (*shu-yüan*). Rampant iconoclasm climaxed with Li Chih, a zealous debunker of traditional Confucian morality, who abandoned a bureaucratic career for Buddhist monkhood of a highly unorthodox type. Excesses of this sort provoked occasional suppressions of private academies, occasional persecutions of heretics, and sophisticated counterarguments from traditionalistic, moralistic groups of scholars, such as those associated with the Tung-lin Academy near Su-chou, who blamed the late Ming decline of political efficiency and morality on widespread subversion of Chu Hsi orthodoxy. The zealous searching for personal identity was only intensified, however, when the dynasty finally collapsed.

Ming porcelain

**Fine arts.** In the realm of the arts, the Ming period has long been esteemed for the variety and high quality of its state-sponsored craft goods—cloisonné and, particularly, porcelain wares. The sober, delicate monochrome porcelains of the Sung dynasty were now superseded by rich, decorative polychrome wares. The best known of these are of blue-on-white decor, which gradually changed from floral and abstract designs to a pictorial emphasis. From this eventually emerged the "willow-pattern" wares that became export goods in great demand in Europe. By late Ming times, perhaps because of the unavailability of the imported Iranian cobalt that was used for the finest blue-on-white products, more flamboyant polychrome wares of three and even five colours predominated. Painting—chiefly portraiture—followed traditional patterns under Imperial patronage, but independent gentlemen painters became the most esteemed artists of the age, especially four masters of the Su-chou area: Shen Chou, Ch'iu Ying, T'ang Yin, and Wen Cheng-ming. Their work, always of great technical excellence, became less and less academic in style; and out of this tradition, by the late years of the dynasty, emerged a conception of the true painter as a professionally competent but deliberately amateurish artist bent on individualistic self-expression. Notably in landscapes, a highly cultivated and somewhat romantic or mystical simplicity became the approved style, perhaps best exemplified in the work of Tung Ch'i-ch'ang.

**Literature and scholarship.** As was the case with much of the painting, Ming poetry and belles lettres were deliberately composed "after the fashion of" earlier masters, and groups of writers and critics earnestly argued about the merits of different T'ang and Sung exemplars. No Ming practitioner of traditional poetry has won special esteem, though Ming literati churned out poetry in prodigious quantities. The historians Sung Lien and Wang Shih-chen and the philosopher-statesman Wang Yang-ming were among the dynasty's most noted prose stylists, producing expository writings of exemplary lucidity and straightforwardness. Perhaps the most admired master was Kuei Yu-kuang, whose most famous writings are simple essays and anecdotes about everyday life—often rather loose and formless but with a quietly pleasing charm, evoking character and mood with artless-seeming delicacy. The iconoclasm of the final Ming decades was mirrored in a literary movement of total individual freedom, championed notably by Yüan Tsung-tao; but writings produced during this period were later denigrated as being insincere, coarse, frivolous, and so strange and eccentric as to make impossible demands on the readers.

The importance of popular fiction

The late Ming iconoclasm did successfully call attention to popular fiction in colloquial style. In retrospect, this must be reckoned the most significant literary work of the late Yüan and Ming periods, despite its being disdained by the educated elite of the time. The late Yüan–early Ming novels *San-kuo chih yen-i* (*Romance of the Three Kingdoms*) and *Shui-hu chuan* (*The Water Margin,* also translated as *All Men Are Brothers*) became the universally acclaimed masterpieces of the historical and picaresque genres, respectively. Sequels to each were produced throughout the Ming period. Wu Ch'eng-en, a 16th-century local official, produced *Hsi-yu chi* (*Journey to the West,* also partially translated as *Monkey*), which became China's most treasured novel of the supernatural; and late in the 16th century an unidentifiable writer produced *Chin P'ing Mei* (also translated as *Golden Lotus*), a realistically Rabelaisian account of life and love among the bourgeoisie, which established yet another genre for the novel. By the end of the Ming period iconoclasts such as Li Chih and Chin Sheng-t'an, both of whom published editions of *Shui-hu chuan,* made the then astonishing assertion that this and other works of popular literature should rank alongside the greatest poetry and literary prose as treasures of China's cultural heritage. Colloquial short stories also proliferated in Ming times, and collecting anthologies of them became a fad of the last Ming century. The master writer and editor in this realm was Feng Meng-lung, whose creations and influence dominate the best-known anthology, *Chin-ku ch'i-kuan* ("Wonders Old and New"), published in Su-chou in 1624.

Operatic drama, which had emerged as a major new art form in Yüan times, was popular throughout the Ming dynasty, and Yüan masterpieces in the tightly disciplined four-act *tsa-chü* style were regularly performed. Ming contributors to the dramatic literature were most creative in a more rambling, multiple-act form known as "southern drama" or as *ch'uan-ch'i.* Members of the Imperial clan and respected scholars and officials such as Wang Shih-chen and particularly T'ang Hsien-tsu wrote for the stage.

A new southern opera aria form called *k'un-ch'ü,* originating in Su-chou, became particularly popular and provided the repertoire of sing-song girls throughout the country. Sentimental romanticism was a notable characteristic of Ming dramas.

Ming scholarship

Perhaps the most representative of all Ming literary activities, however, were voluminous works of sober scholarship in many realms. Ming literati were avid bibliophiles, both collectors and publishers. They founded many great private libraries, such as the famed T'ien-i-ko collection of the Fan family at Ning-po. They also began producing huge anthologies (*ts'ung-shu*) of rare or otherwise interesting books and thus preserved many works from extinction. The example was set in this regard by an Imperially sponsored classified anthology of all the esteemed writings of the whole Chinese heritage completed in 1407 under the title *Yung-lo ta-tien* ("Great Canon of the Yung-lo Era"). Its more than 11,000 volumes being too numerous for even the Imperial government to consider printing, it was preserved only in manuscript copies. Private scholars also produced great illustrated encyclopaedias, including *Pen-ts'ao kang-mu* (late 16th century; "Index of Native Herbs"), a monumental materia medica listing 1,892 herbal concoctions and their applications; *San-ts'ai t'u-hui* (1607–09; "Assembled Pictures of the Three Realms"), a work on subjects such as architecture, tools, costumes, ceremonies, animals, and amusements; *Wu-pei chih* (1621; "Treatise on Military Preparedness"), on weapons, fortifications, defense organization, and war tactics; and *T'ien-kung k'ai-wu* (1637; "Creations of Heaven and Human Labour"), on industrial technology. Ming scholars also produced numerous valuable geographical treatises and historical studies. Among the creative milestones of Ming scholarship, which pointed the way for the development of modern critical scholarship in early Ch'ing times, were the following: a work by Mei Tsu questioning the authenticity of sections of the ancient *Shu Ching* ("Classic of History"); a phonological analysis by Ch'en Ti of the ancient *Shih Ching* ("Classic of Poetry"); and a dictionary by Mei Ying-tso that for the first time classified Chinese characters under 214 radicals and subclassified them by number of brushstrokes—the arrangement of most standard modern dictionaries.

One of the great all-around literati of Ming times, representative in many ways of the dynamic and wide-ranging activities of the Ming scholar-official at his best, was Yang Shen. Yang won first place in the metropolitan examination of 1511, remonstrated vigorously against the caprices of the Cheng-te and Chia-ching emperors, and was finally beaten, imprisoned, removed from his post in the Hanlin Academy, and sent into exile as a common soldier in Yunnan. But he produced poetry and belles lettres in huge quantities, as well as a study of bronze and stone inscriptions through history, a dictionary of obsolete characters, suggestions about the phonology of ancient Chinese, and a classification of fish in Chinese waters. (C.O.Hu.)

## The Ch'ing dynasty

### EARLY CH'ING

**The rise of the Manchu.** The Manchu, who ruled China from 1644 to 1911, were descendants of the Juchen tribes who had ruled North China as the Chin dynasty in the 12th century. From the 15th century they had paid tribute to the Ming and been organized in the commandery system, so they had long had extensive and regular contact with the Chinese state and, more importantly, with the Chinese military officers stationed in the Ming frontier garrisons. By the 16th century these officers had become a hereditary regional military group in southern Manchuria, the Manchu homeland. Transformed by their long residence on the frontier, the Chinese soldiers mingled with the "barbarians," adopting Manchu names and tribal customs. Still other Chinese were in the area as enslaved "bond servants" who worked the land or helped manage the trade in ginseng root, precious stones, and furs with China and Korea. Later, after the conquest of China, many of these bond servants became powerful officials who were sent on confidential missions by the emperor and staffed the powerful Imperial Household Department.

Rule of Nurhachi and Abahai

Under Nurhachi and his son Abahai, the Aisin Gioro clan of the Chien-chou tribe won hegemony among the rival Juchen tribes of the northeast, then extended its control through warfare and alliances into Inner Mongolia and Korea. Nurhachi created large permanent civil-military units called "banners" to replace the small hunting groups used in his early campaigns. A banner was composed of smaller companies; it included some 7,500 warriors and their households, including slaves, under the command of a chieftain. Each banner was identified by a coloured flag that was yellow, white, blue, or red, either plain or with a border design. Originally there were four, then eight, Manchu banners; new banners were created as the Manchu conquered new regions, and eventually there were Manchu, Mongol, and Chinese banners, eight for each ethnic group. By 1648 less than 16 percent of the bannermen were actually of Manchu blood. The Manchu conquest was thus achieved with a multiethnic army, led by Manchu nobles and Han Chinese generals. Han Chinese soldiers were organized into the Army of the Green Standard, which became a sort of Imperial constabulary force posted throughout China and on the frontiers.

Modern scholarship on the rise of the Manchu emphasizes the contributions of Chinese collaborators to the Manchu cause. The Manchu offered rewards and high positions to these Chinese, who not only brought military skills and technical knowledge with them but also encouraged the adoption of Chinese institutional models. From Chinese and Korean artisans the Manchu learned iron-smelting technology and acquired the advanced European artillery of the Ming. They created a replica of the Ming central government apparatus in their new capital, Mukden (modern Shen-yang), established in 1625. Whereas Nurhachi had initially based his claim to legitimacy on the tribal model, proclaiming himself khan in 1607, he later adopted the Chinese political language of the T'ien Ming ("Mandate of Heaven") and in 1616 created the Hou (Later) Chin dynasty. Abahai continued to manipulate the political symbols of both worlds by acquiring the great seal of the Mongol khan in 1635, and thus the succession to the Yüan dynasty, and by taking on a Chinese dynastic name, Ch'ing, for his own dynasty the following year.

Factors in the Ming's downfall

The downfall of the Ming house was the product of factors that extended far beyond China's borders. In the 1630s and '40s China's most commercialized regions, the Yangtze Delta and the southeast coast, suffered an acute economic depression brought on by a sharp break in the flow of silver entering ports through foreign trade from Acapulco, Malacca, and Japan. The depression was exacerbated by harvest shortfalls resulting from unusually bad weather during 1626–40. The enervated government administration failed to respond adequately to the crisis, and bandits in the northwest expanded their forces and began invading north and southwest China. One of these bandit leaders, Li Tzu-ch'eng, marched into Peking in 1644 unopposed, and the Emperor, forsaken by his officials and generals, committed suicide. A Ming general, Wu San-kuei, sought Manchu assistance against Li Tzu-ch'eng. Dorgon, the regent and uncle of Abahai's infant son (who became the first Ch'ing emperor), defeated Li and took Peking, where he declared the Manchu dynasty.

It took the Manchu several decades to complete the military conquest of China. In 1673 the conquerors confronted a major rebellion led by three generals (among them Wu San-kuei), former Ming adherents who had been given control over large parts of south and southwest China. This revolt, stimulated by Manchu attempts to cut back on the autonomous power of these generals, was finally suppressed in 1681. In 1683 the Ch'ing finally eliminated the last stronghold of Ming loyalism on Taiwan.

**The Ch'ing Empire.** After 1683 the Ch'ing rulers turned their attention to consolidating control over their frontiers. Taiwan became part of the empire, and military expeditions against perceived threats in north and west Asia created the largest empire China has ever known. From the late 17th to the early 18th century Ch'ing armies destroyed the Oyrat Empire based in Dzungaria and incorporated into the empire the region around the Koko Nor (Blue

Lake) in Central Asia. In order to check Mongol power, a Chinese garrison and a resident official were posted in Lhasa, the centre of the Dge-lugs-pa (Yellow Hat sect) of Buddhism that was influential among Mongols as well as Tibetans. By the mid-18th century the land on both sides of the Tien Shan range as far west as Lake Balkhash had been annexed and renamed Sinkiang ("New Dominion").

Internal migration

Military expansion was matched by the internal migration of Chinese settlers into parts of China that were dominated by aboriginal or non-Han ethnic groups. The evacuation of the south and southeast coast during the 1660s spurred the westward migration of an ethnic minority, the Hakka, who moved from the hills of southwest Fukien, northern Kwangtung, and southern Kiangsi. Although the Ch'ing dynasty tried to forbid migration into its homeland, Manchuria, in the 18th and 19th centuries Chinese settlers flowed into the fertile Liao River basin. Government policies encouraged Han movement into the southwest during the early 18th century, while Chinese traders and assimilated Chinese Muslims moved into Sinkiang and the other newly acquired territories. This period was punctuated by ethnic conflict stimulated by the Han Chinese takeover of former aboriginal territories and by fighting between different groups of Han Chinese.

*Political institutions.* The Ch'ing had come to power because of their success at winning Chinese over to their side; in the late 17th century they adroitly pursued similar policies to win the adherence of the Chinese literati. Ch'ing emperors learned Chinese, addressed their subjects using Confucian rhetoric, reinstated the civil service examination system and the Confucian curriculum, and patronized scholarly projects, as had their predecessors. They also continued the Ming custom of adopting reign names, so that Hsüan-yeh, for example, is known to history as the K'ang-hsi emperor. The Ch'ing rulers initially used only Manchu and bannermen to fill the most important positions in the provincial and central governments (half of the powerful governors-general throughout the dynasty were Manchu), but Chinese were able to enter government in greater numbers in the 18th century, and a Manchu-Han dyarchy was in place for the rest of the dynasty.

The early Ch'ing emperors were vigorous and forceful rulers. The first emperor, Fu-lin (reign name, Shun-chih), was put on the throne when a child of six *sui* (about five years in Western calculations). His reign (1644–61) was dominated by his uncle and regent, Dorgon, until Dorgon died in 1650. Because the Shun-chih emperor had died of smallpox, his successor, the K'ang-hsi emperor, was chosen in part because he had already survived a smallpox attack. The K'ang-hsi emperor (reigned 1661–1722) was one of the most dynamic rulers China has known. During his reign the last phase of the military conquest was completed, and campaigns were pressed against the Mongols to strengthen Ch'ing security on its Inner Asian borders. China's literati were brought into scholarly projects, notably the compilation of the Ming history, under Imperial patronage.

The K'ang-hsi emperor's designated heir, his son Yin-jeng, was a bitter disappointment, and the succession struggle that followed the latter's demotion was perhaps the bloodiest in Ch'ing history. Many Chinese historians still question whether the K'ang-hsi emperor's eventual successor, his son Yin-chen (reign title, Yung-cheng), was truly the emperor's deathbed choice. During the Yung-cheng reign (1722–35) the government promoted Chinese settlement of the southwest and tried to integrate non-Han aboriginal groups into Chinese culture; it reformed the fiscal administration and rectified bureaucratic corruption.

Ch'ien-lung reign

The Ch'ien-lung reign (1735–96) marked the culmination of the early Ch'ing. The Emperor had inherited an improved bureaucracy and full treasury from his father and expended enormous sums on the military expeditions known as the Ten Great Victories. He was both noted for his patronage of the arts and notorious for the censorship of anti-Manchu literary works that was linked with the compilation of the *Ssu-k'u ch'üan-shu* ("Complete Library of the Four Treasuries"; Eng. trans. under various titles). The closing years of his reign were marred by intensified court factionalism centred on the meteoric rise to political power of an Imperial favourite, a young officer named Ho-shen. Yung-yen, who reigned as the Chia-ch'ing emperor (1796–1820), lived most of his life in his father's shadow. He was plagued by treasury deficits, piracy off the southeast coast, and uprisings among aboriginal groups in the southwest and elsewhere. These problems, together with new pressures resulting from an expansion in opium imports, were passed on to his successor, the Tao-kuang emperor (reigned 1820–50).

The early Ch'ing emperors succeeded in breaking from the Manchu tradition of collegial rule. The consolidation of Imperial power was finally completed in the 1730s, when the Yung-cheng emperor destroyed the power base of rival princes. By the early 18th century the Manchu had adopted the Chinese practice of father–son succession but without the custom of favouring the eldest son. Because the identity of the Imperial heir was kept secret until the emperor was on his deathbed, Ch'ing succession struggles were particularly bitter and sometimes bloody.

The Manchu also altered political institutions in the central government. They created an Imperial Household Department to forestall the usurpation of power by eunuchs that had plagued the Ming ruling house, and they staffed this agency with bond servants. The Imperial Household Department became a power outside the control of the regular bureaucracy. It managed the large estates that had been allocated to bannermen and supervised various government monopolies, the Imperial textile and porcelain factories in central China, and the customs bureaus scattered throughout the empire. The size and strength of the Imperial Household Department reflected the accretion of power to the throne that was part of the Ch'ing political process. Similarly, revisions of the system of bureaucratic communication and the creation in 1729 of a new top decision-making body, the Grand Council, permitted the emperor to control more efficiently the ocean of government memorandums and requests.

Imperial Household Department

*Foreign relations.* The Manchu inherited the tributary system of foreign relations from previous dynasties. This system assumed that China was culturally and materially superior to all other nations, and it required those who wished to trade and deal with China to come as vassals to the emperor, who was the ruler of "all under Heaven." The tributary system was used by the Ch'ing Board of Rites to deal with the countries along China's eastern and southern borders and with the European nations that sought trade at the ports of south and southeast China.

The tributary system operated in its fullest form in the Ch'ing treatment of Korea. The Korean court used the Chinese calendar, sent regular embassies to Peking to present tribute, and consulted the Chinese on the conduct of foreign relations. The Ch'ing emperor confirmed the authority of the Korean rulers, approved the Korean choice of consorts and heirs, and bestowed noble ranks on Korean kings. The Korean envoy performed the kowtow (complete prostration and knocking of the head on the ground) before the Ch'ing emperor and addressed him using the terms appropriate to someone of inferior status.

Central Asia was another matter. Tribes on the Inner Asian frontiers had repeatedly invaded China, and the Manchu, who had been part of the world of the steppe, were keenly aware of the need for military supremacy on China's northern borders. Inner Asian affairs were handled by a new agency, the Court of Colonial Affairs, that was created before 1644. Ch'ing policies toward Central Asia frequently deviated from the tributary ideal. Chinese relations with Russia are a case in point. The early Ch'ing rulers attempted to check the Russian advance in north Asia and used the Russians as a buffer against the Mongols. The Sino-Russian Treaty of Nerchinsk (1689), which tried to fix a common border, was an agreement between equals. The Treaty of Kyakhta (1727) extended agreement on the borders to the west and opened markets for trade. When Chinese ambassadors went to Moscow (1731) and St. Petersburg (1732) to request that Russia remain neutral during the Chinese campaigns against the Oyrat in Central Asia, they performed the kowtow before the Empress.

Foreign trade was not always restricted to the formal exchanges prescribed by the tributary system. Extensive trad-

ing was carried out in markets along China's borders with Korea, at the Russo-Mongolian border town of Kyakhta, and at selected ports along the coast, whence ships traded with Southeast Asia. Perhaps the most striking example of trade taking precedence over tribute was the Ch'ing trade with Japan. The Tokugawa shogunate viewed the Ch'ing as barbarians whose conquest sullied China's claim to moral superiority in the world order. They refused to take part in the tributary system and themselves issued trade permits (counterparts of the Chinese tributary tallies) to Chinese merchants coming to Nagasaki after 1715. The Ch'ing need for Japanese copper, a money metal in China, required that trade with Japan be continued, and it was.

Relations with Japan

*Economic development.* In the 1640s and '50s the Manchu abolished all late Ming surtaxes and granted tax exemptions to areas ravaged by war. Tax remissions were limited, however, by the urgent need for revenues to carry on the conquest of China. It was not until the 1680s, after the consolidation of military victory, that the Ch'ing began to permit tax remissions on a large scale. The permanent freezing of the *ting* (corvée quotas) in 1712 and the subsequent merger of the *ting* and land tax into a single tax that was collected in silver were part of a long-term simplification of the tax system. The commutation of levies from payment in kind to payment in money and the shift from registering males to registering land paralleled the increasing commercialization of the economy.

A healthy tax base required that land be brought under cultivation. Because more than one-quarter of the total cultivated land had slipped off the tax rolls in the early 17th century, the restoration of agriculture was an important goal. The new dynasty began to resettle refugees on abandoned land with offers of tax exemptions for several years and grants of oxen, tools, seeds, or even cash in some areas. In the late 17th century the resettlement of the Ch'eng-tu Basin in western China and of Hunan, Hupeh, and the far southwest proceeded on this basis.

Land reclamation went hand in hand with the construction and reconstruction of water-control projects. This was an activity so characteristic of a new dynasty that one can speak of "hydraulic cycles" moving in tandem with political consolidations in China. These water-control projects varied in scale with terrain and ecology. In central and South China irrigation systems were the foundation for rice cultivation and were largely the product of private investment and management. In North China control of the heavily silted Huang Ho (Yellow River), which frequently inundated the eastern portion of the North China Plain, required large-scale state management and coordination with the related water level of the Grand Canal, the major north–south waterway supplying Peking.

The preferred crops—rice in central and South China, wheat in North China—retained their primacy in Ch'ing agriculture. In the course of the dynasty, the cultivation of wheat and other northern staple grains continued to creep southward; rice was transplanted to the best lands on the frontiers, and there was a gradual intensification of the cropping cycle. Both on the frontiers and within China proper new lands were opened for settlement using the New World crops that had been introduced into China in the late 16th century. Corn (maize) and the Irish potato permitted Chinese to cultivate the marginal hilly lands. The sweet potato provided insurance against famine, while peanuts (groundnuts) were a new source of oil in the peasant diet. Tobacco, another 16th-century import, competed with rice and sugarcane for the best lands in South China and became an important cash crop.

Once the economy had been restored, the Ch'ing state attempted to keep it running smoothly. For the most part, the state did not actively intervene in what was becoming an extremely complex market economy. The major exception was its successful effort to offset regional food shortages in years of crop failure. Every province was supposed to purchase or retain reserves in the "ever-normal" granaries located in each county, so named because they were intended to stabilize the supply, and hence the price, of grain. Even relatively uncommercialized hinterlands were thus armoured against famine. The ability of the government to respond effectively to food scarcity was dependent on its information gathering. During the 18th century data on local grain prices became a regular feature of county, prefectural, and provincial reports.

The Ch'ing government played a relatively minor role in the commercial economy. There were state monopolies in salt, precious metals, pearls, and ginseng, but the long-run trend was to reduce the number of monopolies. The state barely began to tap the growing revenue potential of trade, just as it failed to tap the expanding agricultural base. Its rare interventions in trade were motivated by a desire to dampen economic fluctuations in employment. Its major goal was stability, not growth.

And yet the early Ch'ing was a period of economic growth and development. With the imposition of the Ch'ing peace, the economy resumed a commercial expansion that had begun in the 16th century. This expansion in turn stimulated specialization in crops sent to market, which included raw materials to be used in the textile industry as well as consumption goods such as tea, sugar, and tobacco. Profit enticed merchants, landlords, and peasants to buy or rent land to produce cash crops. A new kind of managerial landlord, who used hired labour to grow market crops, emerged in the 18th century.

Economic growth

The tenant's position improved vis-à-vis the landlord's, a wage-labour force arose in agriculture, and land was increasingly used as a marketable commodity. Systems that guaranteed tenants permanent rights of cultivation spread in the 18th century through the wet-rice cultivation zone and in some dryland cultivation systems. Multiple layers of rights to the land generally benefited the tenant and improved incentives to maintain the fertility of the soil and to raise output. There was a general shift from servile to contractual labour in agriculture that was part of a long trend toward the elimination of fixed status and the increased mobility of labour and land.

Equally important processes of commercialization gained momentum with the recovery of the domestic economy. The 16th-century boom created new layers of rural markets that linked villages more firmly to a market network. Although the majority of economic transactions continued to take place within local and intermediate markets, there was a significant expansion of interregional and national trade in grain, tea, cotton, and silk. In the 18th century Shanghai became a thriving entrepôt for the coastal trade that extended from Manchuria to South China.

The most dramatic economic innovations of the 18th century resulted from the needs of long-distance traders for credit and new mechanisms that would ease the transfer of funds. Native banks, as they were called by foreigners in the 19th century, accepted deposits, made loans, issued private notes, and transferred funds from one region to another. Promissory notes issued by native banks on behalf of merchants facilitated the purchase of large quantities of goods, and money drafts and transfer accounts also helped ease the flow of funds. By the early 19th century, paper notes may have constituted one-third or more of the total volume of money in circulation. The demands of large-scale, long-distance trade had, without government participation, inspired merchants to transform a metallic monetary system into one in which paper notes supplemented copper coins and silver.

Emergence of a banking system

Customary law evolved outside the formal legal system to expedite economic transactions and enable strangers to do business with one another. Business partnerships in mining, commerce, and commercial agriculture could be formalized and protected through written contract. Reliance on written contracts for the purchase and mortgaging of land, purchase of commodities and people, and hiring of wage labourers became commonplace.

The early Ch'ing economy was intimately tied to foreign trade, which consisted of a junk trade with ports in Southeast Asia, Japan, and the Philippines and the expanding trade conducted by Europeans. After 1684, when the ban on maritime trade was lifted, Western traders flocked to Canton, and foreign commerce was finally confined to this port in 1759. The "Canton system" of trade that prevailed from that year until 1842 specified that Europeans had to trade through the cohong, a guild of Chinese firms that had monopoly rights to the trade in tea and silk.

From 1719 to 1833 the tonnage of foreign ships trading at Canton increased more than 13-fold. The major export was tea; by 1833 tea exports were more than 28 times the export levels of 1719. Silk and porcelain were also exported in increasing quantities through the early 18th century. Although exports remained a small fraction of total output, the effect of foreign trade on the Chinese economy was direct and perceptible. Its repercussions were not limited to the merchants and producers involved in specific export commodities but also had a general impact on domestic markets through the monetary system.

The Chinese economy had long been based on a metallic currency system in which copper cash was used for daily purchases and silver for large business transactions and taxes. The exchange ratio between silver and copper cash was responsive to fluctuations in the supply of the metals, and changes in the exchange ratio affected all citizens. The economic expansion of the 18th century brought rising demand for silver and copper. Although domestic production of copper increased, silver was primarily obtained from abroad. After 1684 the net balance of trade was consistently in China's favour, and silver flowed into the Chinese economy. Perhaps 10,000,000 Spanish silver dollars a year came into China during the early Ch'ing, and in the 18th century Spanish silver dollars became a common unit of account in the southeast and south.

**Ch'ing society.** Chinese society continued to be highly stratified during the early Ch'ing. Hereditary status groups in the society ranged from the descendants of the Imperial line down to the "mean people" at the bottom of the social ladder. Many professions were hereditary: bannermen, brewers, dyers, doctors, navigators, and Taoist priests usually passed on their occupations to at least one son in each generation. The "mean people" included remnants of aboriginal groups who had survived Chinese expansion and settlement and certain occupational groups, including prostitutes, musicians, actors, and local government underlings (jailers and gatekeepers, for example). Ch'ing laws forbade intermarriage between respectable commoners ("good people") and the mean people, who were also barred from sitting for the civil service examinations. Despite attempts in the 1720s to return some of these mean people to ordinary commoner status, the social stigma persisted throughout the dynasty.

Servitude was commonplace in Ch'ing society. The Manchu had enslaved prisoners of war, and in China persons could be sold by their families. Many well-to-do households owned some domestic servants. Grouped with the mean people in Ch'ing law, some servants nonetheless achieved considerable power and authority. Bond servants of the Imperial house ran the powerful Imperial Household Department and themselves owned slaves. Servile tenants of the wealthy Hui-chou merchants were sometimes raised as companions to the master's son and trusted to help run the long-distance trade on which Hui-chou fortunes were based. Servitude in some cases was thus an important avenue for social advancement.

Social mobility

Social mobility increased during the early Ch'ing, supported by a pervasive belief that it was possible for a peasant boy to become the first scholar in the land. An ethic that stressed education and hard work motivated many households to invest their surplus in the arduous preparation of sons for the civil service examinations. Although the most prestigious career in Ch'ing society remained that of the scholar-official, the sharpened competition for degrees in the prosperous 18th century brought a significant expansion in socially acceptable forms of achievement. At one pole, alienated literati deliberately eschewed the morally ambiguous role of official to devote their energies to scholarship, painting, poetry, and the other arts. Others turned to management of their localities and assumed leadership in public welfare, mediation of disputes, and local defense. Families with a long tradition of success in examinations and official service were increasingly preoccupied with strategies for ensuring the perpetuation of their elite status and countering the inexorable division of family estates stemming from the Chinese practice of partible inheritance. Downward mobility was a more general phenomenon than upward mobility in Ch'ing society: those at the bottom of the social scale did not marry and reproduce themselves, while the wealthy practiced polygyny and tended to have many children.

In China's long-settled and densely populated regions, degree holders who confronted the prospect of downward mobility for their sons were profoundly disturbed about the circumstances that permitted rich merchants to mimic their way of life. The money economy and its impersonal values penetrated more deeply into Chinese society than ever before, challenging former indicators of status for preeminence. Alarmed, the Chinese elite joined the Ch'ing state in trying to propagate traditional values and behaviour. Morality books, published in increasing quantities from the late 16th century onward, tied virtuous behaviour to concrete rewards in the form of educational success, high office, and sons. The Ch'ing bestowed titles, gifts, and Imperial calligraphy on virtuous widows and encouraged the construction of memorial arches and shrines in their honour to reinforce this female role. Rural lectures (*hsiang-yüeh*) were public ceremonies staged for citizens that combined religious elements with the recitation of the Sacred Edict promulgated by the emperor.

*Social organization.* The basic unit of production and consumption in Chinese society remained the *chia* ("family"), consisting of kin related by blood, marriage, or adoption, with a common budget and common property. The Chinese family system was patrilineal; daughters married out, while sons brought in wives and shared the residence of their fathers. The head of the family, the patriarch, had the power to direct the activities of each member in an effort to optimize the family's welfare. The family was a metaphor for the state, and family relations were the foundation of the hierarchical social roles that were essential in the Confucian vision of a morally correct society.

In southeast and South China the early Ch'ing saw an expansion of extended kinship organizations based on descent from a common ancestor. In these areas lineages became a powerful tool for collective action and local dominance, using revenues from corporate property to support education, charity, and ancestral rites. Other types of lineages, possessing little corporate property, existed in other parts of China. These lineages seem to have been composed of only the most elite lines within a descent group, who focused their efforts on national rather than local prominence and emphasized their marriage networks rather than ties to poorer kinsmen.

Kinship was of limited use to the increasingly numerous sojourners who were working away from home in the early Ch'ing. Other kinds of organizations emerged to meet the needs of a more mobile population. The share partnership permitted unrelated persons to pool their resources to start a business, and it was used to finance a wide variety of enterprises, including mining ventures, coastal and overseas shipping, commercial agriculture, money shops, and theatres. The trading empires created by the Hui-chou and Shansi merchants were examples of how such partnerships, cemented by kinship and native-place ties, could be used for large-scale business operations.

The *hui-kuan* and *kung-so* organizations

"Native place" was the principle used to organize the *hui-kuan* (native-place associations) that spread throughout Ch'ing market centres. Some *hui-kuan* were primarily intended for officials and examination candidates; these were located in the capitals of provinces and in Peking. Others, located in the southwest, were for immigrants, but the vast majority were created and used by merchants. The *hui-kuan* provided lodging and a place to meet fellow natives, receive financial aid, and store goods. In the course of the 18th century another kind of organization that encompassed all the men in a trade, the *kung-so* (guild), emerged in China's cities. *Hui-kuan* frequently became subunits of *kung-so,* and both groups participated in the informal governance of cities.

New kinds of social organization also emerged on China's frontiers. Native-place ties were frequently expressed in worship of a deity, so that a temple or territorial cult would become a vehicle for collective action. White Lotus sectarianism appealed to other Chinese, most notably to women and to the poor, who found solace in worship of the Eternal Mother who was to gather all her children at

the millennium into one family. The Ch'ing state banned the religion, and it was generally an underground movement. Although the White Lotus faith was practiced by boatmen on the Grand Canal with no attempts to foment uprising, its millenarian message spurred spectacular rebellions, the most notable being the White Lotus Rebellion at the close of the 18th century.

A new form of social organization, based on sworn brotherhood, emerged among male sojourners in southeast China in the late 18th century. The Triad fraternities built on kinship, native-place, and contractor–worker ties but added special rituals that bound fellow workers together as "brothers" in discipleship to a monk founder. Secret lore, initiation rituals, and an elaborate origin myth evolved, but the fraternities tended to be highly decentralized autonomous units. Appearing first on Taiwan, the Triads expanded with transport workers into South China and became a powerful organization that dominated the Chinese underworld.

*State and society.* The state barred literati from using the academies and literary societies for explicitly political activities. Scholars in Peking and in the rich cities of the Yangtze Delta turned from politics to the study of texts that marked the empirical school of scholarship (*k'ao-cheng hsüeh*). Influenced by their knowledge of European mathematics and mathematical astronomy, these scholars laid down new rules for verifying the authenticity of the Classical texts and, by revealing flaws in previously accepted canons, challenged the Neo-Confucian orthodoxy. Turning away from the Confucian quest for sagehood, the empirical scholars were increasingly secular and professional in their pursuit of textual studies. Scholarly associations, poetry societies, and academies were the organizational loci for the empirical schools. Great libraries were created, rare texts were reprinted, and compilation projects proliferated, culminating in the great government-sponsored *Ssu-k'u ch'üan-shu* (1772–82), which undertook to collect for reprinting the best editions of the most important books produced in China, using as selection criteria the methods of the empirical school.

Spread of urban culture

Ch'ing society saw the expansion and extension of a national urban culture into various parts of the empire. Urban culture circulated through the market network into the hinterland, as sojourners disseminated culture from localities into the cities and back again. The dissemination of this culture was also supported by increased functional literacy and the expansion of large-scale printing for commercial and scholarly audiences. There was a wide variety of written materials available in market towns and cities—collections of winning examination essays, route books for commercial travelers, religious pamphlets and scriptures, novels, short story collections, jokebooks, and almanacs. Storytelling, puppet plays, and regional drama in rural and urban places provided yet another mode of cultural dissemination. In China's cities sojourning merchants sponsored visits of drama troupes from their own localities; in this way, regional drama also spread outside its own territory. Drama was the bridge connecting the oral and written realms, the "living classroom" for peasants who learned about cultural heroes and history through watching plays. The expansion of a national urban culture supported the state's efforts to systematize and standardize Chinese society.

China's non-Han minorities found themselves surrounded by an aggressive, expansionist Han Chinese culture during the early Ch'ing. Early attempts by the emperors to protect minorities from the Han onslaught were largely unavailing, and some rulers such as the Yung-cheng emperor actually tried to hasten the assimilation of aboriginal groups into the Chinese order. The Ch'ing categorized these ethnic minorities into two groups: those who were "raw," or still possessed of their own culture, and those who were "cooked," or assimilated. Despite violent resistance, the ethnic minorities were gradually assimilated or pushed further south and west during the early Ch'ing.

*Trends in the early Ch'ing.* The tripling of China's population from the beginning of the Ch'ing dynasty to the mid-19th century rested on the economic expansion that followed the consolidation of Manchu rule. This population growth has been frequently cited as the major cause of the decline of China in the 19th century. Certainly the year 1800 saw the Ch'ing state's surpluses, sufficient through the 18th century to pay for numerous military expeditions, exhausted in the long campaign to quell the White Lotus Rebellion. Whereas fiscal reforms had strengthened the state in the 18th century, fiscal weakness plagued Ch'ing governments in the 19th. The vaunted power of the Ch'ing armies also waned after 1800, in part because of new modes of warfare. Increased commercialization had tied more and more Chinese into large market fluctuations. In the 18th century the world market economy into which China was increasingly integrated worked in its favour and stimulated a long period of internal prosperity. But the favourable trend was reversed in the 1820s and '30s, when rising importation of opium altered the net balance of trade against China and ushered in a period of economic depression. (E.S.R.)

LATE CH'ING

**Western challenge, 1839–60.** The opium question, the direct cause of the first Sino-British clash in the 19th century, began with a late 18th-century British attempt to counterbalance their unfavourable China trade with traffic in Indian opium. After monopolizing the opium trade in 1779, the East India Company's government began to sell the drug at auction to private British traders in India, who shipped it to buyers in China. The silver acquired from the sale of opium in China was sold at Canton for the company's bills of exchange, payable in London, and was used by the company to purchase its large annual tea cargo for sale in Europe. This "triangular trade" became a major vehicle for realizing the potential gains from the British conquest of India, providing as it did a means to repatriate the company's Indian revenue in opium in the form of Chinese teas. In 1819 the company began to handle larger amounts of opium. Substantial social and economic disruption resulted from the spread of the opium habit and of corruption among petty officials and from a fall in the value of copper in China's bimetallic monetary system as silver was drained from the economy. The Peking court repeatedly banned the importing of opium but without success, because the prohibition itself promoted corruption among the officials and soldiers concerned. There was no possibility of the opium question being solved as a domestic affair.

Trade in tea and opium

After the turn of the 19th century the main vehicle of opium smuggling was the country traders who were allowed only to manage the inter-Asian trade under the company's license. Without protection from the company, they cultivated the opium market in China on their own. They defied the opium ban in China and gradually became defiant toward Chinese law and order in general, having nothing in mind but making money. After Parliament revoked the East India Company's monopoly in 1834, William John Napier was appointed chief superintendent of British trade in China and arrived at Canton. He tried to negotiate with the Canton authorities on equal footing, but the latter took his behaviour as contrary to the established Sino-foreign intercourse. His mission failed.

In Peking an 1836 proposal to relax the opium restraint acquired much support. But the Tao-kuang emperor appointed a radical patriot, Lin Tse-hsü, as Imperial commissioner for an anti-opium campaign. Chinese anti-opium efforts, in fact, began to make considerable headway in controlling the Chinese side of the smuggling trade in late 1838 and early 1839. The critical foreign side of the opium trade was, however, beyond Commissioner Lin's direct reach. Arriving at Canton in March 1839, Lin confiscated and destroyed more than 20,000 chests of opium. Skirmishes began after September between the Chinese and the British.

*The first Opium War and its aftermath.* In February 1840 the British government decided on an expedition, and Rear Adm. George Elliot was appointed first commissioner and plenipotentiary to China. In June, 16 British warships arrived in Hong Kong and sailed northward to the mouth of the Pei River to press China with its demands. No agreement was reached. In May 1841 the

British attacked the walled city of Canton and received a ransom of $6,000,000, which provoked a counterattack on the part of the Cantonese. This was the beginning of a continuing conflict between the British and the Cantonese.

The Ch'ing had no effective tactics against the powerful British Navy. They retaliated merely by setting burning rafts on the enemy's fleet; and they encouraged people to take the heads of the enemies, for which they offered a prize. The Imperial banner troops, although they sometimes fought fiercely, were ill-equipped and lacked training for warfare against the more modern British forces. The Green Standard battalions were similarly in decay and without much motivation or good leadership. To make up the weakness, local militias were urgently recruited, but they were useless. The British proclaimed that their aim was to fight the government officials and soldiers who abused the people, not to make war against the Chinese population. And indeed there was a deep rift between the government and people, of which the British could easily take advantage, a weakness in Ch'ing society that became apparent in the crisis of the war.

Elliot's successor, Henry Pottinger, arrived at Macau in August and campaigned northward, seizing Amoy, Tinghai, and Ning-po. Reinforced from India, he resumed action in May 1842 and took Wu-sung, Shanghai, and Chen-chiang. Nanking yielded in August, and peace was restored with the Treaty of Nanking. The main provisions

he Treaty f Nanking

of the treaty were the cession of Hong Kong, the opening of five ports to British trade, the abolition of the cohong system of trade, equality of official recognition, and an indemnity of $21,000,000. This was the result of the first clash between China, which had regarded foreign trade as a favour given by the Heavenly empire to the poor barbarians, and the British, to whom trade and commerce had become "the true herald of civilization."

The Treaty of Nanking was followed by two supplementary arrangements with the British in 1843. In July 1844 China signed the Treaty of Wanghia with the United States and in October the Treaty of Whampoa with France. These arrangements made up a complex of foreign privileges by virtue of the most-favoured-nation clauses (guaranteeing trading equality) conceded to every signatory. All in all, they provided a basis for such later evils as the loss of tariff autonomy, extraterritoriality (exemption from the application or jurisdiction of local law or tribunals), and the free movement of missionaries.

With the signing of the treaties—which began the so-called treaty-port system—the Imperial commissioner Ch'i-ying, newly stationed at Canton, was put in charge of foreign affairs. Following a policy of appeasement, his dealing with foreigners started fairly smoothly. But contrary to the British expectation, the amount of trade dropped after 1846; and, to their dissatisfaction, the question of opium remained unsettled in the postwar arrangements. The core of the Sino-Western tension, however, rested in an antiforeign movement in Kwangtung.

*Antiforeign movement and the Arrow, or second Opium, War.* At the signing of the Treaty of Nanking there was a difference between China and Britain as to whether foreigners were allowed to enter the walled city of Canton. Though Canton was declared open in July 1843, the British faced Cantonese opposition. After 1847 trouble rapidly grew, and as a result of an incident at nearby Fo-shan, a promise was given the British that they would be allowed to enter the city in 1849. Yet troubles continued. As a result of his inability to control the situation, Ch'i-ying was recalled in 1848 and replaced with the less compliant Hsü Kuang-chin. As the promised date neared, the Cantonese demonstrated against British entry. Finally, the British yielded, and the antiforeigners won a victory despite the fact that the Peking court conceded a "temporary entrance" into the city.

After the Cantonese resistance in 1841, the gentry in Kwangtung began to build a more organized antiforeign movement, promoting the militarization of village society. The city of Canton was also a centre of diffusion of xenophobia because the scholars at the city's great academies were proclaiming the Confucian theory that uncultured barbarians should be excluded. The inspired antiforeign mood also contained a strong antigovernment sentiment and perhaps a tendency toward provincialism; the Cantonese rose up against the barbarians to protect their own homeland, without recourse to the government authorities.

In the strained atmosphere in Canton, where the xenophobic governor-general, Yeh Ming-ch'en, was inciting the Cantonese to annihilate the British, the *Arrow* incident occurred in October 1856. Canton police seized the *Arrow,* a Chinese-owned but British-registered ship flying a British flag, and charged its Chinese crew with piracy and smuggling. The British consul Harry Parkes sent a fleet to fight its way up to Canton. French forces joined the venture on the plea that a missionary had been officially executed in Kwangsi. The British government sent an expedition under Lord Elgin as a plenipotentiary. The Russians and the Americans abstained but sent their representatives for diplomatic maneuvering. At the end of 1857 an Anglo-French force occupied Canton; in March 1858 they took the Ta-ku fort and marched to Tientsin.

Foreign gains

The Ch'ing representatives had no choice but to comply with the demands of the British and French; the Russian and U.S. diplomats also gained the privileges their militant colleagues secured by force. During June four Tientsin treaties were concluded that provided for, among other measures, the residence of foreign diplomats in Peking and the freedom of Christian missionaries to evangelize their faith.

In 1859 when the signatories arrived off the Ta-ku fort on their way to sign the treaties in Peking, they were repulsed, with heavy damage inflicted by the gunfire from the fort. In 1860 an allied force invaded Peking, driving the Hsien-feng emperor (reigned 1850–61) out of the capital to the summer palace at Cheng-te. A younger brother of the emperor, Prince Kung Ch'in-wang, was appointed Imperial commissioner in charge of negotiation. But the famous summer palace was destroyed by the British in October. Following the advice of the Russian negotiator, Prince Kung exchanged ratification of the 1858 treaties; in addition, he signed new conventions with the British and the French. The U.S. and Russian negotiators had already exchanged the ratification in 1859, but the latter's diplomatic performance in 1860 was remarkable.

Russian interests in the East had been activated in competition with the British effort to open China. A Russian spearhead, directed to Kuldja (modern I-ning) by way of the Irtysh River, resulted in the Sino-Russian Treaty of Kuldja in 1851, which opened Kuldja and Chuguchak (modern T'a-ch'eng) to Russian trade. Another drive was directed to the Amur watershed under the initiative of Nikolay Muravyov, who had been appointed governor-general of eastern Siberia in 1847. By 1857 Muravyov had sponsored four expeditions down the Amur; in the third one, in 1856, the left bank and lower reaches of the river had actually been occupied by the Russians. In May 1858 Muravyov pressed the Ch'ing general I-shan to sign a treaty at Aigun (modern Ai-hui), by which the territory on the northern bank of the Amur was ceded to Russia and the land between the Ussuri River and the sea was placed in joint possession by the two nations, pending further disposition. But Peking refused to ratify the treaty. When the Anglo-French allies attacked North China in 1860, the Russian negotiator Nikolay Ignatyev acted as China's friend and mediator in securing the evacuation of the invaders from Peking. Soon after the allies had left Peking, Ignatyev secured, as a reward for his mediatory effort, the Sino-Russian Treaty of Peking, which confirmed the Treaty of Aigun and ceded to Russia the territory between the Ussuri and the sea.

Effects of the treaties of 1858–60

The 1858–60 treaties extended the foreign privileges granted after the first Opium War and confirmed or legalized the developments in the treaty-port system. The worst effects for the Ch'ing authorities were not the utilitarian rights, such as trade, commerce, and tariff, but the privileges that affected the moral and cultural values of China. The right to propagate Christianity threatened Confucian values, the backbone of the Imperial system. The permanent residence of foreign representatives in Peking signified an end to the long-established tributary relationship between China and other nations. The partial collapse of

the tribute system meant a loss of the emperor's virtue, a serious blow to dynastic rule in China.

During the turbulent years 1858–60, the Ch'ing bureaucracy was divided between the war and peace parties; and it was the peace party's leaders—Prince Kung, Kuei-liang, and Wen-hsiang—who took charge of negotiating with the foreigners. But they did so because the imminent crisis forced them to and not as a matter of principle.

In 1861 in response to the settlement of the foreign representatives in the capital, the Tsungli Yamen (office for General Management) was opened to deal with foreign affairs, with a main staff filled by the peace-party leaders. The Ch'ing officials themselves, however, deemed this as still keeping a faint silhouette of the tribute system.

The delay and difficulty in the Ch'ing adjustment to the West may possibly be ascribed to both external and internal factors. The Chinese must have seen the Westerners who had appeared in China as purveyors of poisonous drugs, as barbarians in the full sense of the word, from whom they could learn nothing. But the Chinese staunchly held to their tradition, which also had two aspects—ideological and institutional. The core of the ideological aspect was the Confucian distinction between China and foreign nations. The institutional aspect had recently been much studied, however, and precedents in Chinese history had been found, for example, of treaty ports with foreign settlements, consular jurisdiction, and employment of Westerners as Imperial personnel; thus, the Chinese regarded the Western impact as an extension of their tradition rather than as a totally new situation that necessitated a new adjustment. And at least until 1860 the Ch'ing leaders remained withdrawn in the shell of tradition, making no effort to cope with the new environment by breaking the yoke of the past.

**Popular uprising.** The third quarter of the 19th century saw a series of uprisings, again as a result of social discontent.

*The Taiping Rebellion.* In the first half of the 19th century the province of Kwangtung, the homeland of the Taiping people, was beset with accelerating social unrest. After the first Opium War, government prestige declined, and officials lost their capacity to reconcile communal feudings. The greatest among such conflicts was that between the native settlers and the so-called guest settlers—the Hakka, who had migrated to Kwangsi and western Kwangtung, mainly from eastern Kwangtung. The Pai Shang-ti Hui (God Worshipers' Society) was founded by Hung Hsiu-ch'üan, a fanatic who believed himself a son of God, and his protégé, Feng Yün-shan, an able organizer. Their followers were collected from among miners, charcoal workers, and poor peasants in central Kwangsi, most of whom belonged to the Hakka. In January 1851 a new state named T'ai-p'ing T'ien-kuo (Heavenly Kingdom of Great Peace) was declared in the district of Kuei-p'ing in Kwangsi with Hung Hsiu-ch'üan proclaimed Heavenly king. That September the Taiping shifted their base to the city of Yung-an, where they were besieged by the Imperial army until April 1852. Then they broke the siege and rushed into Hunan. Absorbing some secret-society members and outlaws, they dashed to Wu-han, the capital of Hupeh, and proceeded along the Yangtze to Nanking, which they captured in March 1853 and made their capital.

Taiping religion

The core of the Taiping religion was a monotheism tinged with fundamental Protestant Christianity. But it was mixed with a hatred of the Manchu and an intolerance of the Chinese cultural tradition. In the early years of the rebellion this politico-religious faith sustained the fighting spirit of the Taiping. In the ideal Taiping vision the population was to give all of its belongings to a "general treasury," which would be shared by all alike. While this extreme egalitarianism was rarely implemented outside the original Hakka core from Kwangsi, it probably at times attracted the distressed and lured them to the Taiping cause. The origin of many Taiping religious ideas, morals, and institutions can be traced to China's Confucian tradition; but the Taiping's all-out antiregime struggle, motivated by strong religious beliefs and a common sharing, also had precedents in earlier religious rebellions.

After the Taiping settled in Nanking, village officials were appointed, and redistribution of farmland was planned in accordance with an idea of primitive communism. But in fact the land reform was impracticable. The village officials' posts were filled mainly by the former landlords or the clerks of the local governments, and the old order in the countryside was not replaced by a new one that the oppressed people could dominate.

In May 1853 the Taiping sent an expedition to northern China, which reached the neighbourhood of Tientsin but finally collapsed during the spring of 1855. After this, the Yangtze Valley provinces were the main theatre of struggle. Of the government armies in those years, the Green Standards were too ill disciplined, and not much could be expected of the bannermen. The Ch'ing government had no choice but to rely on the local militia forces, such as the "Hunan Braves" (later called the Hunan Army), organized by Tseng Kuo-fan in 1852, and the "Huai Braves" (later called the Huai Army), organized by Li Hung-chang in 1862. These armies were composed of the village farmers, inspired with a strong sense of mission for protecting the Confucian orthodoxy, and were used for wider operations than merely protecting their own villages. The necessary funds for maintaining them were provided initially by local gentry.

The fall of Nanking

The Taiping were gradually beaten down: with the capture of An-ch'ing, the capital of Anhwei, in October 1861 by the Hunan Army, the revolutionary cause was doomed. But the fall of Nanking was accelerated by the cooperation of Chinese mercenaries equipped with Western arms, commanded by an American, Frederick T. Ward; a Briton, Charles G. Gordon; and others. Nanking's fall in July 1864 marked the end of one of the greatest civil wars in world history. The main cause of the Taiping failure was internal strife among the top leaders in Nanking. Not only did they give themselves over to luxury, but also their energy was exhausted and their leadership lost by a fratricidal conflict that occurred in 1856. In addition, religious fanaticism, though it inspired the fighters, became a stumbling block that interfered with the rational and elastic attitude necessary to handle delicate military and administrative affairs. The intolerance toward traditional culture alienated the gentry and the people alike. Presumably, the failure of the land-redistribution policy also estranged the landless paupers from the Taiping cause.

*The Nien Rebellion.* Often in the first half of the 19th century plundering gangs, called *nien,* ravaged northern Anhwei, southern Shantung, and southern Honan. In mid-century, however, their activities were suddenly intensified, partly by the addition to their numbers of a great many starving people who had lost their livelihood from repeated floods of the Huang Ho in the early 1850s, and partly because they had become emboldened by the Taiping advance north of the Yangtze. From 1856 to 1859 the Nien leaders consolidated their bases north of the Huai River by winning over the masters of the earth-wall communities, consolidated villages that had been fortified for self-defense against the Taiping. The Nien strategy was to use their powerful cavalry to plunder the outlying areas and carry the loot to their home bases.

Many influential clans, with all their members, joined the Nien cause; and among the Nien leaders, the clan chiefs played an important role. Gentry of lower strata also joined the Nien. The greater part of the Nien force consisted of poor peasants, although deserters from the government-recruited militias and salt smugglers were important as military experts. The real cause of their strength was supposed to have been the people's support and sympathy for their leaders; but there was difficulty in creating a power centre, because the Nien's basic social unit was the earth-wall community, where a powerful master exercised autonomy. In 1856 Chang Lo-hsing received the title of "lord of the alliance" of the Nien, but he was far too weak to form a centre. Imperial pacification was launched by Gen. Seng-ko-lin-ch'in, who led a powerful cavalry into the affected area in 1862; but his pursuit was ineffective, and the general himself was killed in Shantung in May 1865. Thus, the last Imperial crack unit disappeared. Tseng Kuo-fan succeeded Seng as general and enforced a

policy of detaching the earth-wall masters from their men and of employing the latter as his troops. Finally, Li Hung-chang succeeded Tseng in 1866 and set up encirclement lines along the Huang Ho and the Grand Canal, by means of which he destroyed the revolts in 1868.

*Muslim rebellions.* Muslim rebellions in Yunnan and in Shensi and Kansu originated from clashes between the Chinese and Muslims in those provinces. Religious antipathy must be taken into account, but more important were the social and political factors. In the frontier provinces the late dynastic confusions were felt as keenly as elsewhere, which aggravated the problems between the Chinese and the Muslims. Yunnan had been haunted by Muslim-Chinese rivalries since 1821, but in Shensi small disturbances had been seen as early as the Ch'ien-lung reign. Government officials supported the Chinese, and the Muslims were obliged to rise up against both the Chinese and the authorities.

Clashes between Chinese and Muslims

A rivalry between the Chinese and Muslim miners in central Yunnan triggered a severe clash in 1855, which developed into a slaughter of a great many Muslims in and around the provincial capital, K'un-ming, the following April. This caused a general uprising of Yunnan Muslims, which lasted until 1873. Lack of a unified policy weakened the Muslims, and the rebellion was brought to an end partly through the pacifiers' policy of playing the rebel leaders off against one another.

Another Muslim uprising, in Shensi in 1862, promptly spread to Kansu and East Turkistan and lasted for 15 years. The general cause of the trouble was the same as in Yunnan, but the Taiping advance to Shensi stimulated the Muslims into rebellion. The first stage of the uprising developed in the Wei Valley in Shensi; in the next stage the rebels, defeated by the Imperial army, fled to Kansu, which became the main theatre of fighting. Encouraged by the Nien invading Shensi at the end of 1866, the core of the rebel troops returned to Shensi, and sporadic clashes continued in the two provinces. In the last phase, Tso Tsung-t'ang, a former protégé of Tseng Kuo-fan, appeared in Shensi with part of the Huai Army and succeeded in pacifying the area in 1873.

In Shensi and Kansu there were many independent Muslim leaders, but they had neither a common headquarters nor unified policy, and there were no all-out revolutionaries. Pacification was delayed because the Imperial camp was preoccupied with the Taiping and the Nien and could not afford the expenditure needed for an expedition to the remote border provinces.

*Effects of the rebellions.* To meet the large popular uprisings, the Ch'ing authorities had to rely on local armies, which were financed by the provincial and local gentry class. To meet this need, a special tax on goods in transit (called the likin) was started in 1853, the proceeds of which remained largely outside the control of the central government. The provincial governors-general and governors came to enlarge their military and financial autonomy, bringing about a trend of decentralization. Moreover, the locus of power shifted from the Manchu to those Chinese who had played the main part in putting down the rebellions. The Hunan Army was gradually disbanded after the fall of Nanking; but the Huai Army, after its success against the Muslims, served as a strong basis for the political maneuvers of its leader, Li Hung-chang, until its defeat and collapse in the Sino-Japanese War in 1894–95.

Shift of power away from the Manchu

The rebellions brought immeasurable damage and devastation to China. Both the Taiping and the pacifiers were guilty of brutality and destruction. A contemporary estimate of 20,000,000 to 30,000,000 victims is certainly far less than the real number. In the course of the Taiping Rebellion the lower Yangtze provinces lost much of their surplus population, but thereafter the region was resettled by immigrants from less damaged areas. Its ruined industry and agriculture had not fully recovered even by the beginning of the 20th century. The area of the Muslim rebellions, too, suffered catastrophic devastation and depopulation.

During the first half of the 19th century a number of natural disasters left large hordes of starving victims who had no choice but to join the Taiping and other rebel groups. The worst calamity, however, was a drought that attacked the northern provinces of Shansi, Shensi, and Honan in 1877–78 and caused hardship for 9,000,000 to 13,000,000 people. These disasters were a serious setback to China, which had just begun to promote industrialization to meet the Western challenge.

**The Self-Strengthening Movement.** Upon the Hsien-feng emperor's death at Cheng-te in 1861, his antiforeign entourage entered Peking and seized power; but Tz'u-hsi, the mother of the newly enthroned boy emperor Tsai-ch'un (reigned as the T'ung-chih emperor, 1861–75), and Prince Kung succeeded in crushing their opponents by a coup d'état in October. There emerged a new system in which the leadership in Peking was shared by Tz'u-hsi and another empress dowager, Tz'u-an, in the palace and by Prince Kung and Wen-hsiang, with the Tsungli Yamen as their base of operation. The core of their foreign policy was expressed by Prince Kung as "overt peace with the Western nations in order to gain time for recovering the exhausted power of the state."

*Foreign relations in the 1860s.* The Tsungli Yamen had two offices attached to it: the Inspectorate General of Customs and the Language School, called T'ung-wen-kuan. The former was the centre for the Maritime Custom Service, administered by Western personnel appointed by the Ch'ing. The latter was opened to train the children of bannermen in foreign languages, and later some Western sciences were added to its curriculum; but the quality of candidates for the school was not high. Similar schools were opened in Shanghai and Canton.

A superintendent of trade for the three northern ports (later known as high commissioner for *pei-yang,* or "northern ocean") was established in 1861 at Tientsin, parallel to a similar, existing post at Shanghai (later known as high commissioner for *nan-yang,* or "southern ocean"). The creation of the new post was presumably aimed at weakening the foreign representatives in Peking by concentrating foreign affairs in the hands of the Tientsin officials.

In 1865–66 the British strongly urged the Ch'ing authorities to make domestic reforms and to become westernized. Prince Kung asked the high provincial officials to submit their opinions about the proposed reforms. The consensus advocated diplomatic missions abroad and the opening of mines but firmly argued against telegraph and railway construction. Against this background, a roving mission was sent to the United States in 1868, which then proceeded to London and Berlin. This first mission abroad was a success for China, but its very success had an adverse effect on China's modernization by encouraging the conservatives, who learned to regard the Westerners as easy to manipulate.

First mission abroad

The Anglo-Chinese Treaty of Tientsin provided for its revision in the year 1868, at which time the Ch'ing were able to negotiate with due preparations and in a mood of peace for the first time after the Opium Wars. The result was the Alcock Convention of 1869, which limited the unilateral most-favoured-nation clause, a sign of gradual improvement in China's foreign relations; but under pressure from British merchants in China, the London government refused to ratify it. The resentment engendered by the refusal, together with an anti-Christian riot at Tientsin in 1870, brought an end to the climate of Sino-foreign cooperation that had prevailed in the 1860s.

The post-Opium Wars arrangements forced China to remove the ban on Christianity, but the Peking court tried to keep the fact secret and encouraged provincial officials to prohibit the religion. The pseudo-Christian Taiping movement furthered the anti-Christian move on the part of royalists. Under such circumstances, anti-Christian riots had spread throughout the country, culminating in the Tientsin Massacre in 1870, in which a French consul and two officials, 10 nuns, and two priests died, and in which three Russian traders were killed by mistake. At the negotiating table the French sternly demanded the lives of three responsible Chinese officials as a preventive against further such occurrences, but the Ch'ing negotiators, Tseng Kuo-fan and Li Hung-chang, were successful at least in refusing the demanded execution. After the incident, however, Tseng was denounced for his infirm

stand and Prince Kung's political influence began to wane in the growing antiforeign climate.

As to the nature of the anti-Christian movement, there are various interpretations: some emphasize the antiforeign Confucian orthodoxy, while others stress the patriotic and nationalistic reaction against the missionaries' attempt to westernize the Chinese. Still others point to the Christian support of the oppressed in their struggle against the official and gentry class. What is clear, however, is that Christianity sowed dissension and friction in the already disintegrating late Ch'ing society and undermined the prestige of the Ch'ing dynasty and the Confucian orthodoxy.

*Industrialization for "self-strengthening."* Stimulated by the military training and techniques exhibited during the Westerners' cooperation against the Taiping and supported by Prince Kung in Peking, the Self-Strengthening Movement was launched by the anti-Taiping generals Tseng Kuo-fan, Li Hung-chang, and Tso Tsung-t'ang, who sought to consolidate the Ch'ing power by introducing Western technology. The ideological champion of the movement was Feng Kuei-fen, who urged "the use of the barbarians' superior techniques to control the barbarians" and proposed to give the gentry stronger leadership than before in local administration.

Introduction of Western technology

In the first period of modern industrial development (1861–72), effort was focused on the manufacture of firearms and machines, the most important enterprises being the Kiangnan Arsenal in Shanghai, the Tientsin Machine Factory, and the Fu-chou Navy Yard. There were many other smaller ones. But the output was disappointing—the shipyard at Fu-chou, for example, built 15 vessels during the five years after 1869 as scheduled, but thereafter it declined and was destroyed in 1884 during the Sino-French War—and the weapons industry was significant not so much for its direct military purpose as for the introduction of Western knowledge and techniques through the many educational facilities that were attached to each installation.

In the second period (1872–94), weight shifted from the weapons industry to a wider field of manufacture, and the operation shifted from direct government management to a government-supervised and merchant-managed method. Leading among the several enterprises of the second period were the China Merchants' Steam Navigation Company and the K'ai-p'ing Coal Mines. These enterprises were sponsored by high provincial officials—the central figure was Li Hung-chang—but their management was left to joint operation by shareholders' representatives and the lower officials appointed by the sponsors.

Bureaucratic malpractice and corruption

Management, however, was beset with bureaucratic malpractices. The seat of decision making and responsibility was obscure, business was spoiled by nepotism and corruption, and the sponsors tended to use the enterprises as a basis for their regional power. The central government not only was unable to supply capital but also looked for every opportunity to exploit these enterprises as it had exploited the monopolistic salt business after which those companies were modeled. Under such circumstances, the enterprises inevitably slid into depression after some initial years of apparent success.

Compounding the problems were the compradors (Chinese agents employed by foreign firms in China) who, acting as a link between Chinese commerce and the foreign firms in the treaty ports, accumulated vast wealth from the new enterprises. Though active in supplying capital and managerial personnel to the enterprises, the compradors themselves lacked technical training and knowledge and often indulged in speculation and embezzlement. Each comprador belonged to an exclusive community by strong family or regional ties that focused his concerns on his community rather than on national interests.

These shortcomings were deeply rooted in the late Ch'ing social conditions and more than offset efforts to construct and maintain the new enterprises. Thus, Chinese society as a whole did not change structurally before 1911.

**Changes in outlying areas.** With the decline of the Ch'ing power and prestige, beginning in the early 19th century, China's peripheral areas began to free themselves from the Ch'ing influence.

*East Turkistan.* To the west of Kashgaria in East Turkistan, a khanate of Khokand emerged in Ferghana after 1760 as a powerful caravan trade centre. When Muslim rebellion spread rapidly from Shensi and Kansu to East Turkistan, a Khokandian adventurer, Yakub Beg, seized the opportunity to invade Kashgaria and established power there in 1865; he soon showed signs of advancing to the Ili region in support of the British in India. In Ili rebel Muslims had set up an independent power at Kuldja in 1864, which terrorized the Russian borders in defiance of the Sino-Russian Treaty of 1851. The Russians, therefore, occupied Kuldja in 1871 and remained there for 10 years.

Having subdued the Kansu Muslim rebellion in 1873, Tso Tsung-t'ang captured Urumchi (Wu-lu-mu-ch'i) in August 1876 and restored the whole region northward to the Tien Shan range, except for the Kuldja area, and painstakingly recovered Kashgaria at the end of 1877.

Li Hung-chang hoped to regain Ili through negotiation. A treaty for the restitution of Ili, signed at Livadia in October 1879, was extremely disadvantageous to China. Upon returning home amid a storm of condemnation, the Chinese negotiator Ch'ung-hou was sentenced to death; the Russians considered this to be inhuman and they stiffened their attitude. But the minister to Britain and France, Tseng Chi-tse, son of Tseng Kuo-fan, succeeded in concluding a treaty at St. Petersburg in February 1881 that was more favourable yet still conceded the Russians many privileges in East Turkistan.

Though at a cost of nearly 58,000,000 taels in expedition and indemnity, the northwest was finally restored to China, and in 1884 a new province, Sinkiang, was established over the area, which had never before been integrated into China.

Establishment of Sinkiang

*Tibet and Nepal.* Ch'ing control of Tibet reached its height in 1792. But thereafter China became unable to protect Tibet from foreign invasion. When an army from northern India invaded western Tibet in 1841, China could not afford to reinforce the Tibetans, who expelled the enemy on their own. China was a mere bystander during a coup d'état in Lhasa in 1844 and could not protect Tibet when it was invaded by Gurkhas in 1855. Tibet thus tended to free itself from Ch'ing control.

The border dispute between Nepal and British India, which sharpened after 1801, had caused the Anglo-Nepalese War of 1814–16 and brought the Gurkhas under British influence. During the war the Gurkhas sent several missions to China in vain expectation of assistance. During political unrest in Nepal after 1832, an anti-British clique seized power and sought assistance from China to form an anti-British common front with the Ch'ing, then fighting the Opium War. But this, too, was rejected. Jang Bahadur, who had become premier of Nepal in 1846, decided on a pro-British policy; his invasion of Tibet in 1855—which took advantage of the Taiping uprising in China—gained Nepal many privileges there. Though Nepal sent quinquennial missions to China until 1906, the Gurkhas had not recognized Chinese suzerainty.

*Burma.* In 1867 the British gained the right to station a commercial agent at Bhamo in Burma, from which they could explore the Irrawaddy River up to the Yunnan border. A British interpreter accompanying a British exploratory mission to Yunnan was killed by local tribesmen on the Yunnan–Burma border in February 1875. The British minister in China, Thomas Wade, seized the opportunity to force China to comply with the Chefoo Convention (1876), which further enlarged the British rights by opening more Chinese ports to foreign trade and agreeing to a mission to delineate the Yunnan–Burmese border, though the London government put off its ratification until 1885. Kuo Sung-tao, appointed chief of a mission of apology to Britain, arrived in London in 1877. He was the first Chinese resident minister abroad, and within two years China opened embassies in five major foreign capitals.

Further enlargement of British rights

When the last king of Burma, Thibaw, tried to join with France and Italy to stave off the British pressure, Britain sent an ultimatum in October 1885, seized the capital of Mandalay, and annexed the country in January 1886. During the final bargaining with the British, the Burmese

king ignored his tributary relations with the Ch'ing; yet China proposed that the Burmese royal court be preserved even nominally so that it could send a decennial mission to China. Britain refused, but in a convention signed in July 1886 it agreed that the Burmese government should send to China a decennial envoy. This outdated practice, however, was buried in 1900.

*Vietnam.* In 1802 a new dynasty was founded in Vietnam (Dai Viet) by Nguyen Phuc Anh, a member of the royal family of Nguyen at Hue, who had expelled the short-lived Tay Son regime and had unified the country. The Ch'ing, under the Chia-ch'ing emperor, recognized the new dynasty as a fait accompli, but a controversy arose as to a name for the new country. Nguyen Phuc Anh demanded the name Nam Viet, but the Ch'ing recommended Vietnam, reversing the two syllables. Finally an agreement was reached, and Nguyen Phuc Anh became king of Vietnam.

During the rule of the second and third kings of Vietnam (1820–41) the persecution of Christians was accelerated. France resorted to arms after 1843 and, by the treaty of 1862 signed at Saigon, received three eastern provinces of Cochinchina besides other privileges concerning trade and religion. In time, French attentions were focused on the Tongking (Tonkin) delta region into which the Red River flows, providing easy access to Yunnan. But the region was beset with many disorderly gangs escaped from China, including the Black Flags under the command of Liu Yung-fu, a confederate of the Taiping. After a small French force had occupied some key points in Tongking in 1873, a treaty was signed at Saigon in March 1874 that stipulated the sovereignty and independence of Vietnam. Though this clause implied that China could not intervene in Vietnamese affairs, the Tsungli Yamen failed to file a strong protest. In 1880, however, the Ch'ing claimed a right to protect Vietnam as its vassal state. Against the French occupation of Tongking in 1882–83 and France's proclamation of protectorate status for Vietnam (under the name of Annam) in the Treaty of Hue of August 1883, the Ch'ing deployed its army in the northern frontier of Tongking. Efforts for a peaceful settlement ended in failure, and both countries prepared for war.

Sino-French warfare

In August 1884 French warships attacked Fu-chou and destroyed the Chinese fleet and dockyard there. Thereafter, however, the French navy and army were stalemated, and an armistice was reached in the spring of 1885. By the subsequent definitive treaty, the French protectorate of Vietnam was recognized, terminating the historical Sino-Vietnamese tributary relationship.

In this crisis, the attitude of the Ch'ing headquarters fluctuated between advocates of militancy and appeasement. Meanwhile, Li Hung-chang and Tseng Kuo-ch'üan were reluctant to mobilize their *pei-yang* and *nan-yang* navies in accordance with orders from Peking.

*Japan and Ryukyu.* Three years after the Meiji Restoration of 1868—which inaugurated a period of modernization and political change in Japan—a commercial treaty was signed between China and Japan, and it was ratified in 1873. Understandably it was reciprocal, because both signatories had a similar unequal status vis-à-vis the Western nations. The establishment of the new Sino-Japanese relations was supported by Li Hung-chang and Tseng Kuo-fan, who advocated positive diplomacy toward Japan.

In 1872 the Meiji government conferred on the last king of Ryukyu, Shō Tai, the title of vassal king and in the following year took over the island's foreign affairs. In reprisal for the massacre of shipwrecked Ryukyuans by Taiwanese tribesmen in 1871, the Tokyo government sent a punitive expedition to Taiwan. Meanwhile, the Japanese sent an envoy to Peking to discuss the matter, and the Ch'ing agreed to indemnify Japan. In 1877, however, the Ryukyu king asked for Ch'ing intervention to revive his former tributary relations with China; Sino-Japanese negotiations were opened at Tientsin in regard to Ryukyu's position, and an agreement was reached in 1882. But the Ch'ing refused to ratify it, and the matter was dropped.

*Korea and the Sino-Japanese War.* In Korea (Chosŏn) a boy king, Kojong, was enthroned in 1864 under the regency of his father, Taewŏn-gun, a vigorous exclusionist. In 1866 they began a nationwide persecution of Christians and repulsed the French and Americans there. The Ch'ing, although uneasy, did not intervene.

After the Meiji Restoration, Japan made many efforts, all in vain, to open new and direct intercourse with Korea. The main reasons for the failure were such trifles as some offensive words in letters addressed by the Japanese government to Korea, or Japanese envoys coming to Korea in Western-style dress, which was unacceptable to the Koreans. With a slightly improved Korean attitude toward foreigners, a Japanese envoy began talks at Pusan in 1875, but the parley was protracted. Japan impatiently sent warships to Korea; these sailed northward to Kanghwa Bay, where gunfire was exchanged between the Japanese vessels and a Korean island fort. The Treaty of Kanghwa was signed in 1876, in which Korea was defined as an independent state on an equal footing with Japan. Japan sent an envoy, Mori Arinori, to China to report on recent Korean affairs. China insisted that though Korea was independent, China could come to the support of its vassal state (Korea) in a crisis, an interpretation that Mori saw as contrary to the idea of independence in international law.

Sino-Japanese competition

From this time on, the Ch'ing strove to increase their influence in Korea; they helped open Korea to the United States and supported the efforts of pro-Chinese Koreans for modernization. But in Korea a powerful conservatism and xenophobia provided the basis for the resurgence of Taewŏn-gun. In July 1882 he expelled Queen Min and her clique and burned down the Japanese legation. The Ch'ing dispatched an army to Korea, arrested Taewŏn-gun, and urged the King to sign a treaty with Japan. Thus, the Ch'ing claim for suzerainty was substantiated.

In December 1884 another coup was attempted by a group of pro-Japanese reformists, but it failed because of the Ch'ing military presence in Korea. From these two incidents Ch'ing political influence and commercial privileges emerged much stronger, though Japan's trade in Korea far surpassed that of China in the late 1880s.

In 1860 a Korean scholar, Ch'oe Che-u, founded a popular religion called Tonghak (Eastern Learning). By 1893 it had turned into a political movement that attracted a vast number of paupers under the banner of antiforeignism and anticorruption. They occupied the southwestern city of Chŏnju in late May 1894. Both China and Japan sent expeditions to Korea, but the two interventionists arrived to find the rebels at Chŏnju already dispersed. To justify its military presence, Japan proposed to China a policy of joint support of Korean reform. When China refused on the ground that this was counter to Korean independence, a clash seemed inevitable. On July 25 the Japanese Navy defeated a Chinese fleet in Kanghwa Bay, and on August 1 both sides declared war. Japan gained victories in every quarter on land and sea.

During the crisis the Ch'ing power centre was again divided. The *pei-yang* navy was less powerful than it appeared, lacking discipline, unified command, and the necessary equipments of a modern navy. In February 1895 Li Hung-chang was appointed envoy to Japan; he signed a peace treaty at Shimonoseki on April 17, whose main items were recognition of Korean independence, indemnity of 200,000,000 taels, and the cession of Taiwan, the Pescadores Islands, and the Liaotung Peninsula. Six days later, however, Russia, Germany, and France forced Japan to restore the peninsula; Japan formally relinquished it on May 5, for which China agreed to pay 30,000,000 taels. Gaining China's favour by this intervention, the three powers began to press China with demands, which gave rise to a veritable scramble for concessions.

**Reform and upheaval.** Immediately after the triple intervention, Russia succeeded in 1896 in signing a secret treaty of alliance with China against Japan, by which Russia gained the right to construct the Chinese Eastern Railway across northern Manchuria. In November 1897 the Germans seized Chiao-chou Bay in Shantung and forced China to concede them the right to build two railways in the province. In March 1898 Russia occupied Port Arthur (modern Lü-shun) and Dairen (Lü-ta) on the Liaotung Peninsula and obtained the lease of the two ports and the right to build a railway connecting them to the Chinese

The scramble for concessions

Eastern Railway. Vying with Russia and Germany, Britain leased Wei-hai in Shantung and the New Territories opposite Hong Kong and forced China to recognize the Yangtze Valley as being under British influence. Following suit, Japan put the province of Fukien under its influence, and France leased Kuang-chou Bay, southwest of Hong Kong, and singled out three southwestern provinces for its sphere of influence. Thus, China was placed on the brink of partition, arousing a keen sense of crisis in 1898 in which the Hundred Days of Reform was staged.

*The Hundred Days of Reform, 1898.* The advocates of the Self-Strengthening Movement had regarded any institutional or ideological change as needless. But after 1885 some lower officials and comprador intellectuals began to emphasize institutional reforms and the opening of a parliament and to stress economic rather than military affairs for self-strengthening purposes. For the Peking court and high officials in general the necessity of reform had to be proved on the basis of the Chinese Classics. Some scholars tried to meet their criteria. The outstanding reform leader and ideologist K'ang Yu-wei used what he considered authentic Confucianism and Buddhist canons to show that change was inevitable in history and, accordingly, that reform was necessary. Another important reformist thinker, T'an Ssu-t'ung, relied more heavily on Buddhism than K'ang did and emphasized the people's rights and independence. Liang Ch'i-ch'ao was an earnest disciple of K'ang but later turned toward people's rights and nationalism under the influence of Western philosophy.

In April 1895, when Japanese victory appeared inevitable, K'ang began to advocate institutional reform. In August K'ang, Liang, and other reformists founded a political group called the Society for the Study of National Strengthening. Though this association was soon closed down, many study societies were created in Hunan, Kwangtung, Fukien, Szechwan, and other provinces. In April 1898 the National Protection Society was established in Peking under the slogan of protecting state, nation, and national religion. Against this background, the Kuang-hsü emperor (ruled 1875–1908) was himself increasingly affected by the ideas of reform that were broadly in the air and perhaps was also directly influenced by K'ang Yu-wei's proposals. On June 11, 1898, the Emperor began to issue a stream of radical and probably hastily prepared reform decrees that lasted for about 100 days, until September 20. The reform movement produced no practical results, however. Finally, the conservatives were provoked to a sharp reaction when they learned of a reformists' plot to remove the arch-conservative empress dowager Tz'u-hsi. On September 21 the Emperor was detained and the Empress Dowager took over the administration, putting an end to the reform movement.

Failure of the reform movement

The immediate cause of the failure lay in the power struggle between the Emperor and Tz'u-hsi. But from the beginning, prospects for reform were dim because most high officials were cool toward or opposed to the movement. In addition, the reformist–conservative confrontation overlapped with the rivalry between the Chinese and the Manchu, who considered the Chinese-sponsored reform as disadvantageous to them. As for the reformists themselves, their leaders were few in number and inexperienced in politics, and their plan was too radical.

Among the local movements for reform, that in Hunan was the most active. After 1896, journals and schools were begun there for popular enlightenment; but K'ang's radical reformism aroused strong opposition, and the Hunan movement was shattered at the end of May 1898.

Though it failed, the reform movement had a few important repercussions: it produced some degree of freedom of speech and association, furthered the dissemination of Western thought, and stimulated the growth of private enterprises. It also provided much of the substance for the "conservative" Imperial reform efforts that the Manchu court undertook after the Boxer episode.

*The Boxer Rebellion.* The crisis of 1896–98 stirred a furious antiforeign uprising in Shantung, aroused by the German advances and encouraged by the provincial governor. It was staged by a band of people called the I-ho ch'üan (Righteous and Harmonious Fists) who believed that a mysterious boxing art rendered them invulnerable to harm. The group's origin is generally supposed to have been in the White Lotus sect, though it may have begun as a self-defense organization during the Taiping Rebellion. At first the Boxers (as they were called in the West) directed their wrath against Christian converts, whom they vilified for having abandoned traditional Chinese customs in favour of an alien religion. Bands of Boxers roamed the countryside killing Chinese Christians and foreign missionaries. Developing from this anti-Christian hysteria, the Boxer Rebellion grew into a naive but furious attempt to destroy all things foreign—including churches, railways, and mines—which the people blamed for their misery and for the loss of a sacred way of life.

Some Boxer recruits were disbanded Imperial soldiers and local militiamen; others were Grand Canal boatmen deprived of a livelihood by the Western-built railways. Most recruits, though, came from the peasantry, which had suffered terribly from recent natural calamities in North China. After 1895 the Huang Ho flooded almost annually, and in 1899–1900 a serious drought struck the North. Vast numbers of starvelings turned to begging and banditry and were easy converts to the Boxers' cause.

Many local authorities refused to stop the violence. Some supported the Boxers by incorporating them into local militias. The Manchu court, meanwhile, was alarmed by the uncontrollable popular uprising but took great satisfaction at seeing revenge taken for its humiliation by the foreign powers. As a result, it assumed at first a neutral policy. On the part of the Boxers, there emerged sometime in the autumn of 1899 a move to gain access to the court under the slogan of "support for the Ch'ing and extermination of foreigners." By May 1900 the Ch'ing government had changed its policy and was secretly supporting the Boxers. Tz'u-hsi inclined toward open war when she became convinced of the dependability of the Boxers' art. Finally, incensed over a false report that the foreign powers had demanded that she return administration to the Emperor, she called on all Chinese to attack foreigners. Within days, on June 20, the Boxers' eight-week siege of the foreign legations in Peking began; a day later Tz'u-hsi declared war by ordering provincial governors to take part in the hostilities.

Government support of the Boxers

An international reinforcement of some 2,000 men had left Tientsin for Peking before the siege, but on the way it was resisted by the Boxers and forced back to Tientsin. The foreign powers then sent an expedition of some 19,000 troops, which marched to Peking and seized the city on August 14. Tz'u-hsi and the Emperor fled to Sian.

The two governors-general in the southeastern provinces, Liu K'un-i and Chang Chih-tung, who together with Li Hung-chang at Canton had already disobeyed Peking's antiforeign decrees, concluded an informal pact with foreign consuls at Shanghai on June 26, to the effect that the governors-general would take charge of the safety of the foreigners under their jurisdiction. At first the pact covered the five provinces in the Yangtze River region, but later it was extended to three coastal provinces. Thus, the foreign operations were restricted to Chihli (modern Hopeh) Province, along the northern coast.

The United States, which had announced its commercial Open Door policy in 1899, made a second declaration of the policy in July 1900—this time insisting on the preservation of the territorial and administrative entity of China. With its newly acquired territory in the western Pacific, the United States was determined to preserve its own commercial interests in China by protecting Chinese territorial integrity from the other major powers. This provided a basis for the Anglo-German agreement (October 1900) for preventing further territorial partition, to which Japan and Russia consented. Thus, partition of China was avoided by mutual restraint among the powers.

The final settlement of the disturbance was signed in September 1901. The indemnity amounted to 450,000,000 taels to be paid over 39 years. Moreover, the settlement demanded the establishment of permanent guards and the dismantling of forts between Peking and the sea, a humiliation that made an independent China a mere fiction. In addition, the southern provinces were actually

Collapse of Ch'ing prestige

independent during the crisis. These occurrences meant the collapse of the Ch'ing prestige.

After the uprising Tz'u-hsi had to declare that she had been misled into war by the conservatives and that the court, neither antiforeign nor antireformist, would promote reforms, a seemingly incredible statement in view of the court's suppression of the 1898 reform movement. But the Ch'ing court's antiforeign conservative nationalism and the reforms undertaken after 1901 were, in fact, among several competing responses to the shared sense of crisis in early 20th-century China.

Early work of Sun Yat-sen

**Reformist and revolutionist movements at the end of the dynasty.** Sun Yat-sen, a commoner with no background of Confucian orthodoxy, educated in Western-style schools in Hawaii and Hong Kong, went to Tientsin in 1894 to meet Li Hung-chang and present a reform program, but he was refused an interview. This event supposedly caused his antidynastic attitude. Soon he went to Hawaii, where he founded an anti-Manchu fraternity called the Hsing-chung hui (Revive China Society). Returning to Hong Kong, he and some friends set up a similar society under the leadership of his associate Yang Ch'ü-yün. After an abortive attempt to capture Canton in 1895, Sun sailed for England and then went to Japan in 1897, where he found much support. Tokyo became the revolutionaries' principal base of operation.

After the collapse of the Hundred Days of Reform, K'ang Yu-wei and Liang Ch'i-ch'ao had also fled to Japan. An attempt to reconcile the reformists and the revolutionaries became hopeless by 1900—Sun was slighted as a secret-society ruffian, while the reformists were more influential among the Chinese in Japan and the Japanese.

The two camps competed in collecting funds from the overseas Chinese as well as in attracting secret-society members on the mainland. The reformists strove to unite with the powerful secret Ko-lao hui (Society for Brothers and Elders) in the Yangtze River region. In 1899 K'ang's followers organized the Tzu-li chün (Independence Army) at Han-k'ou in order to plan an uprising, but the scheme ended unsuccessfully. Early in 1900 the Revive China Society revolutionaries also formed a kind of alliance with the Brothers and Elders, called the Revive Han Association. This new body nominated Sun as its leader, a decision that also gave him, for the first time, the leadership of the Revive China Society. The Revive Han Association started an uprising at Hui-chou, in Kwangtung, in October 1900, which failed after two weeks' fighting with Imperial forces.

After the Boxer disaster, Tz'u-hsi reluctantly issued a series of reforms, which included abolishing the civil service examination, establishing modern schools, and sending students abroad. But these measures could never repair the damaged Imperial prestige; rather, they inspired more anti-Manchu feeling and raised the revolutionary tide. But there were other factors that intensified the revolutionary cause: the introduction of social Darwinist ideas by Yen Fu after the Sino-Japanese War countered the reformists' theory of change based on the Chinese Classics; and Western and revolutionary thoughts came to be easily and widely diffused through a growing number of journals and pamphlets published in Tokyo, Shanghai, and Hong Kong.

Spread of revolutionary ideas

Nationalists and revolutionists had their most enthusiastic and numerous supporters among the Chinese students in Japan, whose numbers increased rapidly between 1900 and 1906. The Tsungli Yamen sent 13 students to Japan for the first time in 1896; in 10 years the figure rose to some 8,000. Many of these students began to organize themselves for propaganda and immediate action for the revolutionary cause. In 1902–04 revolutionary and nationalistic organizations, including the Chinese Educational Association, the Society for Revival of China, and the Restoration Society, appeared in Shanghai. The anti-Manchu tract "Revolutionary Army" was published in 1903, and more than 1,000,000 copies were issued.

Dealing with the young intellectuals was a new challenge for Sun Yat-sen, who hitherto had concentrated on mobilizing the uncultured secret-society members. He had also to work out some theoretical planks, though he was not a first-class political philosopher. The result of his response was the Three Principles of the People—nationalism, democracy, and socialism—the prototype of which came to take shape by 1903. He expounded his philosophy in America and Europe during his travels there in 1903–05, returning to Japan in the summer of 1905. The activists in Tokyo joined him to establish a new organization called the United League (T'ung-meng hui); under Sun's leadership, the intellectuals increased their importance.

*Sun Yat-sen and the United League.* Sun's leadership in the league was far from undisputed. His understanding that the support of foreign powers was indispensable for Chinese revolution militated against the anti-imperialist trend of the young intellectuals. Only half-heartedly accepted was the principle of people's livelihood, or socialism, one of his Three Principles. Though various evaluations are given to his socialism, it seems certain that it did not reflect the hopes and needs of the commoners.

Ideologically, the league soon fell into disharmony; Chang Ping-lin, an influential theorist in the Chinese Classics, came to renounce the Three Principles of the People; others deserted to anarchism, leaving anti-Manchuism as the only common denominator in the league. Organizationally, too, the league became divided; a Progressive Society (Kung-chin hui), a parallel to the league, was born in Tokyo in 1907; a branch of this new society was soon opened at Wu-han with the ambiguous slogan of "equalization of human right." The next year Chang Ping-lin tried to revive the Restoration Society.

Effect of the Russo-Japanese War

*Constitutional movements after 1905.* Japan's victory in the Russo-Japanese War (1904–05) aroused a cry for constitutionalism in China. Unable to resist the intensifying demand, the court decided in September 1906 to adopt a constitution, and in November it reorganized the traditional six boards into 11 ministries in an attempt to modernize the central government. It promised to open consultative provincial assemblies in October 1907 and proclaimed in August 1908 the outline of a constitution and a nine-year period of tutelage before its full implementation.

Three months later the strangely coinciding deaths of Tz'u-hsi and the Emperor were announced, and a boy who ruled as the Hsüan-t'ung emperor (1908–1911/12) was enthroned under the regency of his father, the second Prince Chün. These deaths, followed by that of Chang Chih-tung in 1909, almost emptied the Ch'ing court of prestigious members. The consultative provincial assemblies were convened in October 1910 and became the main base of the furious movement for immediate opening of a consultative national assembly, with which the court could not comply.

The gentry and wealthy merchants were the sponsors of constitutionalism; they had been striving to gain the rights held by foreigners. Started first in Hunan, the so-called rights recovery movement spread rapidly and gained noticeable success, reinforced by local officials, students returned from Japan, and the Peking government. But finally the recovery of the railroad rights ended in a clash between the court and the provincial interests.

The retrieval of the Han-k'ou–Canton line from the American China Development Company in 1905 tapped a nationwide fever for railway recovery and development. But difficulty in raising capital delayed railway construction by the Chinese year after year. The Peking court therefore decided to nationalize some important railways in order to accelerate their construction by means of foreign loans, hoping that the expected railway profits would somehow alleviate the court's inveterate financial plight. In May 1911 the court nationalized the Han-k'ou–Canton and Szechwan–Han-k'ou lines and signed a loan contract with the four-power banking consortium. This incensed the Szechwan gentry, merchants, and landlords who had invested in the latter line, and their anti-Peking remonstrance grew into a provincewide uprising. The court moved some troops into Szechwan from Hupeh; some other troops in Hupeh mutinied and suddenly occupied the capital city, Wu-ch'ang, on October 10, now the memorial day of the Chinese Revolution.

The commoners' standard of living, which had not continued to grow in the 19th century and may have begun to deteriorate, was further dislocated by the mid-century

civil wars and foreign commercial and military penetration. Paying for the wars and their indemnities certainly increased the tax burden of the peasantry, but how serious a problem this was is still an open question. The Manchu reforms and preparations for constitutionalism added a further fiscal exaction for the populace, which benefited very little from these urban-oriented developments. Rural distress, resulting from this and from natural disasters, was among the causes of local peasant uprisings in the Yangtze River region in 1910 and 1911 and of a major rice riot at Ch'ang-sha, the capital of Hunan, in 1910. But popular discontent was limited and not a major factor contributing to the revolution that ended the Ch'ing dynasty and inaugurated the republican era in China.

Burden of the tax-paying commoners

*The Chinese Revolution (1911–12).* The Chinese Revolution was not triggered by the United League itself but by the Hupeh army troops urged on by the local revolutionary bodies not incorporated in the league. An accidental exposure of a mutinous plot forced a number of junior officers to choose between arrest or revolt in Wu-han. The revolt was initially successful because of the determination of lower level officers and revolutionary troops and the cowardice of the responsible Manchu and Chinese officials. Within a day the rebels had seized the arsenal and the governor-general's offices and had gained possession of the provincial capital, Wu-ch'ang. With no nationally known revolutionary leaders on hand, the rebels coerced a colonel, Li Yüan-hung, to assume military command, although only as a figurehead. They persuaded the Hupeh provincial assembly to proclaim the establishment of the Chinese republic; T'ang Hua-lung, the assembly's chairman, was elected head of the civil government.

After this initial victory, a number of historical tendencies converged to bring about the downfall of the Ch'ing dynasty. A decade of revolutionary organization and propaganda paid off in a sequence of supportive uprisings in important centres of central and South China; these occurred in recently formed military academies and in newly created divisions and brigades, in which many cadets and junior officers were revolutionary sympathizers. Secret-society units also were quickly mobilized for local revolts. The antirevolutionary constitutionalist movement also made an important contribution; its leaders had become disillusioned with the Imperial government's unwillingness to speed the process of constitutional government, and a number of them led their respective provincial assemblies to declare their provinces independent of Peking or actually to join the new republic. T'ang Hua-lung was the first among them. As a product of the newly emerging nationalism, there was widespread hostility among Chinese toward the alien dynasty. Many had absorbed the revolutionary propaganda that blamed a weak and vacillating court for the humiliations China had suffered from foreign powers since 1895. Therefore, there was a broad sentiment in favour of ending Manchu rule. Also, as an outcome of two decades of journalizing discussion of "people's rights," there was substantial support among the urban educated for a republican form of government. Probably the most decisive development was the recall of Yüan Shih-k'ai, the architect of the elite Peiyang Army, to government service to suppress the rebellion when its seriousness became apparent.

After the collapse of the Huai Army in the Sino-Japanese War, the Ch'ing government had striven to build up a new Western-style army, among which the elite corps trained by Yüan Shih-k'ai, former governor-general of Chihli, had survived the Boxer uprising and emerged as the strongest force in China. But it was, in a sense, Yüan's private army and did not easily submit to the Manchu court. Yüan had been retired from officialdom at odds with the regent prince Chün; but on the outbreak of the revolution in 1911, the court had no choice but to recall him from retirement to take command of his new army. Instead of using force, however, he played a double game—on the one hand, he deprived the floundering court, step by step, of all its power; on the other, he started to negotiate with the revolutionaries. At the peace talks that opened at the end of the year, Yüan's emissaries and the revolutionary representatives agreed that the abdication of the Ch'ing and the appointment of Yüan to the presidency of the new republic were to be formally decided by the National Assembly. But this was renounced by Yüan, probably because he hoped to be appointed by the retiring Manchu monarch to organize a new government rather than nominated as chief of state by the National Assembly. (This is a formula of the Chinese dynastic revolution called *ch'an-jang,* which means the peaceful shift in rule from a decadent dynasty to a more virtuous one.) But events turned against him, and the presidency was given to Sun Yat-sen, who had been appointed provisional president of the republic by the National Assembly. In February 1912 the Ch'ing court proclaimed the decree of abdication, which contained a paragraph purporting that Yüan Shih-k'ai was to organize a republican government to negotiate with the revolutionists on unification of North and South; this portion was a fabrication inserted by Yüan into the last Imperial document. Thus, the Ch'ing dynasty ended its 268-year rule with no legitimate successor. (C.Su./A.Fe.)

End of the Ch'ing dynasty

## The republican period

### THE DEVELOPMENT OF THE REPUBLIC (1912–20)

The first half of the 20th century saw the gradual disintegration of the old order in China and the turbulent preparation for a new society. Foreign political philosophies undermined the traditional governmental system, nationalism became the strongest activating force, and civil wars and Japanese invasion tore the vast country and retarded its modernization. Although the revolution ushered in a republic, China had virtually no preparation for democracy. A three-way settlement ended the revolution—abdication by the dynasty; relinquishment of the provisional presidency by Sun Yat-sen in favour of Yüan Shih-k'ai, regarded as the indispensible man to restore unity; and Yüan's promise to establish a republican government. This placed at the head of state an autocrat by temperament and training, and the revolutionaries had only a minority position in the new national government.

**Early power struggles.** During the first years of the republic there was a continuing contest between Yüan and the former revolutionaries over where ultimate power should lie. The contest began with the election of parliament (National Assembly) in February 1913. The Nationalist Party (Kuomintang; KMT), made up largely of former revolutionaries, won a commanding majority of seats. Parliament was to produce a permanent constitution. Sung Chiao-jen, the main organizer of the KMT's electoral victory, advocated executive authority in a cabinet responsible to parliament rather than to the president. On March 20, 1913, Sung was assassinated; the confession of the assassin and later circumstantial evidence strongly implicated the Premier and, possibly, Yüan himself.

Yüan versus parliament

Parliament tried to block Yüan's effort to get a "reorganization loan" (face value $125,000,000) from a consortium of foreign banks, but in April Yüan concluded the negotiations and received the loan. He then dismissed three Nationalist military governors. That summer, revolutionary leaders organized a revolt against Yüan, later known as the Second Revolution, but his military followers quickly suppressed it. Sun Yat-sen, one of the principal revolutionaries, fled to Japan. Yüan then coerced parliament into electing him formally to the presidency, and he was inaugurated on October 10, the second anniversary of the outbreak of the revolution. By then his government had been recognized by most foreign powers. When parliament promulgated a constitution placing executive authority in a cabinet responsible to the legislature, Yüan revoked the credentials of the KMT members, charging them with involvement in the recent revolt. He dissolved parliament on Jan. 10, 1914, and appointed another body to prepare a constitution according to his own specifications. The presidency had become a dictatorship.

**China in World War I.** *Japanese gains.* With the outbreak of World War I, in August 1914, Japan joined the side of the Allies and seized the German leasehold around Chiao-chou Bay together with German-owned railways in Shantung. China was not permitted to interfere. Then, on Jan. 18, 1915, the Japanese government secretly pre-

sented to Yüan the Twenty-one Demands, which sought, in effect, to make China a Japanese dependency. Yüan skillfully directed the negotiations by which China tried to limit its concessions, which centred around greater access to Chinese ports and railroads and even a voice in Chinese political and police affairs. At the same time Yüan searched for foreign support. The European powers, locked in war, were in no position to restrain Japan. The United States was unwilling to intervene. The Chinese public, however, was aroused. Most of Yüan's political opponents supported his resistance to Japan's demands. Nevertheless, on May 7 Japan gave Yüan a 48-hour ultimatum, forcing him to accept the terms as they stood at that point in the negotiations.

Japan gained extensive special privileges and concessions in Manchuria and confirmation of its gains in Shantung from Germany. The Han-yeh-p'ing mining and metallurgical enterprise in the middle Yangtze Valley was to become a joint Sino-Japanese company. China promised not to alienate to any other power any harbour, bay, or island on the coast of China nor to permit any nation to construct a dockyard, coaling station, or naval base on the coast of Fukien, the province nearest to Japan's colony of Taiwan.

*Yüan's attempts to become emperor.* In the wake of the humiliation of these forced concessions, Yüan launched a movement to revive the monarchy, with some modernized features, and to place himself on the throne. The Japanese government began to "advise" against this move in October and induced its allies to join in opposing Yüan's plan. Additional opposition came from the leaders of the Nationalist and Progressive parties. In December, Ch'en Ch'i-mei and Hu Han-min, two followers of Sun Yat-sen, who was actively scheming against Yüan from his exile in Japan, began a movement against the monarchy. More significant was a military revolt in Yunnan, led by Gen. Ts'ai O, a disciple of Liang Ch'i-ch'ao, and by the governor of Yunnan, T'ang Chi-yao. Joined by Li Lieh-chün and other revolutionary generals, they established a Hu-kuo chün (National Protection Army) and demanded that Yüan cancel his plan. When he would not, the Yunnan army in early January 1916 invaded Szechwan, and subsequently Hunan and Kwangtung, hoping to bring the southwestern and southern provinces into rebellion and then induce the lower Yangtze provinces to join them. The Japanese government covertly provided funds and munitions to Sun and the Yunnan leaders. One by one military leaders in Kweichow, Kwangsi, and parts of Kwangtung declared the independence of their provinces or districts. By March the rebellion had assumed serious dimensions, and public opinion was running strongly against Yüan.

)pposition
) Yüan's
nthrone-
ient

A third source of opposition came from Yüan's direct subordinates, generals Tuan Ch'i-jui and Feng Kuo-chang, whose powers Yüan had attempted to curtail. When he called upon them for help, they both withheld support. On March 22, with the tide of battle running against his forces in the southwest, Japanese hostility increasingly open, public opposition in full cry, and his closest subordinates advising peace, Yüan announced the abolition of the new empire. His opponents, however, demanded that he give up the presidency as well. The revolt continued to spread, with more military leaders declaring the independence of their provinces. The issue became that of succession should Yüan retire. The president, however, became gravely ill; he died on June 6 at the age of 56.

Yüan's four years had serious consequences for China. The country's foreign debt was much enlarged, and a precedent had been established of borrowing for political purposes. Yüan's defiance of constitutional procedures and his dissolution of parliament also set precedents that were later repeated. There was much disillusionment with the republican experiment; China was a republic in name, but arbitrary rule based upon military power was the political reality. The country was becoming fractured into competing military satrapies—the beginning of warlordism.

Gen. Li Yüan-hung, the vice president, succeeded to the presidency, and Gen. Tuan Ch'i-jui continued as premier, a position he had accepted in April. A man of great ability and ambition, Tuan was supported by many generals of the former Peiyang Army, a powerful force based in North China that developed originally under Yüan's leadership. Tuan quickly began to gather power into his own hands. Li favoured the restoration of parliament and a return to the provisional constitution of 1912. Parliament reconvened on August 1; it confirmed Tuan as premier but elected Gen. Feng Kuo-chang, the leader of another emerging faction of the Peiyang Army, as vice-president. The presidential transition and restoration of parliament had by no means answered the underlying question of where the governing power lay.

*Conflict over entry into the war.* In February 1917 the U.S. government severed diplomatic relations with Germany and invited the neutral powers, including China, to do the same. This brought on a crisis in the Chinese government. Li opposed the step, but Tuan favoured moving toward entry into the war. Parliamentary factions and public opinion were bitterly divided. Sun Yat-sen, now in Shanghai, argued that entering the war could not benefit China and would create additional perils from Japan. Under heavy pressure, parliament voted to sever diplomatic relations with Germany, and Li was compelled by his premier to acquiesce. When the United States entered the war in April, Tuan wished China to do the same but was again opposed by the President.

Tuan and his supporters demanded that China enter the war and that Li dissolve parliament. On May 23, Li dismissed Tuan; he then called upon Gen. Chang Hsün, a power in the Peiyang clique and also a monarchist, to mediate. As a price for mediation, Chang demanded that Li dissolve parliament, which he did reluctantly on June 13. The next day Chang entered Peking with an army and set about to restore the Ch'ing dynasty. Telegrams immediately poured in from military governors and generals denouncing Chang and the coup; Li refused to sign the restoration order and called upon Tuan to bring an army to the capital to restore the republic. Li requested that Vice President Feng assume the duties of president during the crisis and then took refuge in the Japanese legation. Tuan captured Peking on July 14; Chang fled to asylum in the Legation Quarter. Thus ended a second attempt to restore the Imperial system.

Attempt at Manchu restoration

Tuan resumed the premiership, and Feng came to Peking as acting president, bringing a division as his personal guard. The two powerful rivals, each supported by an army in the capital, formed two powerful factions—the Chihli clique under Feng and the Anhwei clique under Tuan. Opposed neither by Li nor by the dissolved parliament, Tuan pushed through China's declaration of war on Germany, announced on Aug. 14, 1917.

*Formation of a rival southern government.* Meanwhile, in July Sun Yat-sen, supported by part of the Chinese navy and followed by some 100 members of parliament, attempted to organize a rival government in Canton. The initial costs of this undertaking, termed the Movement to Protect the Constitution, probably were supplied by the German consulate in Shanghai. On August 31 the rump parliament in Canton established a military government and elected Sun commander in chief. Real power, however, lay with military men, who only nominally supported Sun. The southern government declared war on Germany on September 26 and unsuccessfully sought recognition from the Allies as the legitimate government. A Hu-fa chün (Constitution Protecting Army) made up of southern troops launched a punitive campaign against the government in Peking and succeeded in pushing northward through Hunan. Szechwan also was drawn into the fight. Tuan tried to quell the southern opposition by force, while Feng advocated a peaceful solution. Tuan resigned and mustered his strength to force Feng to order military action; Gen. Ts'ao K'un was put in charge of the campaign and drove the Southerners out of Hunan by the end of April 1918. In May the southern government was reorganized under a directorate of seven, in which military men dominated. Sun therefore left Canton and returned to Shanghai. Although his first effort to establish a government in the South had been unsuccessful, it led to a protracted split between South and North.

*Wartime changes.* Despite limited participation, China

made some gains from its entry into the war, taking over the German and Austrian concessions and canceling the unpaid portions of the Boxer indemnities due its enemies. It was also assured a seat at the peace conference. Japan, however, extended its gains in China. The Peking government, dominated by Tuan after Feng's retirement, granted concessions to Japan for railway building in Shantung, Manchuria, and Mongolia. These were in exchange for the Nishihara loans, amounting to nearly $90,000,000, which went mainly to strengthen the Anhwei clique with arms and cash. Japan also made secret agreements with its allies to support its claims to the former German rights in Shantung and also induced the Peking government to consent to these. In November 1917 the United States, to adjust difficulties with Japan, entered upon the Lansing–Ishii Agreement, which recognized that because of "territorial propinquity . . . Japan has special interests in China." This seemed to underwrite Japan's wartime gains.

Important economic and social changes occurred during the first years of the republic. With the outbreak of the war, foreign economic competition with native industry abated and native-owned light industries developed markedly. By 1918 the industrial labour force numbered some 1,750,000. Modern-style Chinese banks increased in number and expanded their capital.

The rise of a new intelligentsia

**Intellectual movements.** A new intelligentsia had also emerged. The educational reforms and the ending of the governmental examination system during the final Ch'ing years enabled thousands of young people to study sciences, engineering, medicine, law, economics, education, and military skills in Japan. Others went to Europe and the United States. Upon their return they took important positions and were a modernizing force in society. Their writing and teaching became a powerful influence on upcoming generations of students. In 1915–16 there were said to be nearly 130,000 new-style schools in China with more than 4,000,000 students. This was mainly an urban phenomenon, however; rural life was barely affected except for what may have been a gradual increase in tenancy and a slow impoverishment that sent rural unemployed into cities and the armies or into banditry.

*An intellectual revolution.* An intellectual revolution took place during the first decade of the republic, sometimes referred to as the New Culture Movement. It was led by many of the new intellectuals, who held up for critical scrutiny nearly all aspects of Chinese culture and traditional ethics. Guided by concepts of individual liberty and equality, a scientific spirit of inquiry, and a pragmatic approach to the nation's problems, they sought a much more profound reform of China's institutions than had resulted from self-strengthening or the republican revolution. They directed their efforts particularly to China's educated youth.

In September 1915 Chen Duxiu (Ch'en Tu-hsiu), who had studied in Japan and France, founded *Hsin ch'ing-nien* ("New Youth") magazine to oppose Yüan's Imperial ambitions and to regenerate the nation's youth. This quickly became the most popular reform journal, and in 1917 it began to express the iconoclasm of new faculty members in Peking University (Pei-ta), which Chen had joined as dean of the College of Letters. Peking University, China's most prestigious institution of higher education, was being transformed by its new chancellor, Ts'ai Yüan-p'ei, who had spent many years in advanced study in Germany. Ts'ai made the university a centre of scholarly research and inspired teaching. The students were quickly swept into the New Culture Movement. A proposal by Hu Shih, a former student of John Dewey, that literature be written in the vernacular language (*pai-hua*) rather than in classical style won quick acceptance. By 1918 most of the contributors to *Hsin ch'ing-nien* were writing in *pai-hua,* and other journals and newspapers soon followed suit. Students at Peking University began their own reform journal, *Hsin ch'ao* ("New Tide"). A new experimental literature inspired by Western forms became highly popular, and scores of new literary journals were founded.

*Riots and protests.* On May 4, 1919, patriotic students in Peking protested the decision at the Versailles Peace Conference that Japan should retain defeated Germany's rights and possessions in Shantung. Many students were arrested in the rioting that followed. Waves of protest spread throughout the major cities of China. Merchants closed their shops, banks suspended business, and workers went on strike to pressure the government. Finally, the government was forced to release the arrested students, to dismiss some officials charged with being tools of Japan, and to refuse to sign the Treaty of Versailles. This outburst helped spread the iconoclastic and reformist ideas of the intellectual movement, which was renamed the May Fourth Movement. By the early 1920s China was launched on a new revolutionary path.

The May Fourth Movement

## THE INTERWAR YEARS (1920–37)

**Beginnings of a national revolution.** This new revolution was led by the Nationalist Party (KMT) and the Chinese Communist Party (CCP).

*The Nationalist Party.* The Nationalist Party had its origins in the earlier United League (T'ung-meng hui) against the Manchu. The name Nationalist Party was adopted in 1912. After the suppression of this expanded party by Yüan Shih-k'ai, elements from it were organized by Sun Yat-sen in 1914 into the Chinese Revolutionary Party, which failed to generate widespread support. Sun and a small group of veterans were stimulated by the patriotic upsurge of 1919 to rejuvenate this political tradition, as well as to revive the Nationalist Party name. The party's publications took on new life as the editors entered the current debates on what was needed to "save China." Socialism was popular among Sun's followers.

The formation of an effective party took several years, however. Sun returned to Canton from Shanghai late in 1920, when Gen. Ch'en Chiung-ming drove out the Kwangsi militarists. Another rump parliament elected Sun president of a new southern regime, which claimed to be the legitimate government of China. In the spring of 1922 Sun attempted to launch a northern campaign as an ally of the Manchurian warlord, Chang Tso-lin, against the Chihli clique, which by now controlled Peking. Ch'en, however, did not want the provincial revenues wasted in internecine wars. One of Ch'en's subordinates drove Sun from the presidential residence in Canton on the night of June 15–16, 1922. Sun took refuge with the southern navy, and he retired to Shanghai on August 9. He was able to return to Canton in February 1923; he then began to consolidate a base under his own control and to rebuild his party.

*The Chinese Communist Party.* The CCP grew directly from the May Fourth Movement. Its leaders and early members were professors and students who came to believe that China needed a social revolution and who began to see Soviet Russia as a model. Chinese students in Japan and France had earlier studied socialist doctrines and the ideas of Karl Marx, but the Russian Revolution of 1917 stimulated a fresh interest in keeping with the enthusiasm of the period for radical ideologies. Li Dazhao (Li Ta-chao), the librarian of Peking University, and Chen Duxiu were the CCP's cofounders.

In March 1920 word reached China of Soviet Russia's revolutionary foreign policy enunciated in the first Karakhan Manifesto, which promised to give up all special rights gained by tsarist Russia at China's expense and to return the Russian-owned Chinese Eastern Railway in Manchuria without compensation. The contrast between this promise and the Versailles award to Japan that had touched off the 1919 protest demonstrations could hardly have been more striking. Although the Soviet government later denied such a promise and attempted to regain control of the railway, the impression of this first statement and the generosity still offered in a more diplomatic second Karakhan Manifesto of September 1920 left a favourable image of Soviet foreign policy among Chinese patriots.

Russian influence

Russia set up an international Communist organization, the Comintern, in 1919 and sent Grigory N. Voytinsky to China the next year. Voytinsky met Li Dazhao in Peking and Chen Duxiu in Shanghai, and they organized a Socialist Youth League, laid plans for a Communist Party, and started recruiting young intellectuals. By the spring of 1921 there were about 50 members in various Chinese

cities and in Japan, many of them former students who had been active in the 1919 demonstrations. Mao Zedong (Mao Tse-tung), a protégé of Li Dazhao, had started one such group in Ch'ang-sha. The CCP held its First Congress in Shanghai in July 1921, with 12 or 13 attendants and with a Dutch Communist—Hendricus Sneevliet, who used his Comintern name, Maring, in China—and a Russian serving as advisers. Maring had become head of a new bureau of the Comintern in China, and he had arrived in Shanghai in June 1921. At the First Congress Chen Duxiu was chosen to head the party.

The CCP spent the next two years in recruiting, in publicizing Marxism and the need for a national revolution directed against foreign imperialism and Chinese militarism, and in organizing unions among railway and factory workers. Maring was instrumental in bringing the KMT and the CCP together in a national revolutionary movement. A number of young men were sent to Russia for training. Among the CCP members were many students who had worked and studied in France, where they had gained experience in the French labour movement and with the French Communist Party; Zhou Enlai (Chou En-lai) was one of these. Other recruits were students influenced by the Japanese Socialist movement. By 1923 the party had some 300 members, with perhaps 3,000 to 4,000 in the ancillary Socialist Youth League.

*Communist-Nationalist cooperation.* By then, however, the CCP was in serious difficulty. The railway unions had been brutally suppressed, and there were few places in China where it was safe to be a known Communist. In June 1923 the Third Congress of the CCP met in Canton, where Sun Yat-sen provided a sanctuary. After long debate this congress accepted the Comintern strategy pressed by Maring—that Communists should join the KMT and make it the centre of the national revolutionary movement. Sun had rejected a multiparty alliance but had agreed to admit Communists to his party, and several, including Chen Duxiu and Li Dazhao, had already joined the KMT. Even though Communists would enter the other party as individuals, the CCP was determined to maintain its separate identity and autonomy and to attempt to control the labour union movement. The Comintern strategy called for a period of steering the Nationalist movement and building a base among the Chinese masses, followed by a second stage—a socialist revolution in which the proletariat would seize power from the capitalist class.

By mid-1923 the Soviets had decided to renew the effort to establish diplomatic relations with the Peking government. Lev M. Karakhan, the deputy commissar for foreign affairs, was chosen as plenipotentiary for the negotiations. In addition to negotiating a treaty of mutual recognition, Karakhan was to try to regain for the Soviet Union control of the Chinese Eastern Railway. On the revolutionary front, the Soviets had decided to financially assist Sun in Canton and to send a team of military men to help train an army in Kwangtung. By June, five young Soviet officers were in Peking for language training. More importantly, the Soviet leaders selected an old Bolshevik, Mikhail M. Borodin, as their principal adviser to Sun Yat-sen. The Soviet leaders also decided to replace Maring with Voytinsky as principal adviser to the CCP, which had its headquarters in Shanghai. Thereafter three men—Karakhan in Peking, Borodin in Canton, and Voytinsky in Shanghai—were the field directors of the Soviet effort to bring China into the anti-imperialist camp of "world revolution." The offensive was aimed primarily at the positions in China of Great Britain, Japan, and the United States.

The arrival of Soviet advisers

**Reactions to warlords and foreigners.** These states, too, were moving toward a new, postwar relationship with China. At the Washington Conference (November 1921–February 1922), eight powers agreed to respect the sovereignty, independence, and territorial and administrative integrity of China, to give China opportunity to develop a stable government, to maintain the principle of equal opportunity in China for the commerce and industry of all nations, and to refrain from taking advantage of conditions in China to seek exclusive privileges. The powers also agreed to steps leading toward China's tariff autonomy and to the abolition of extraterritoriality. Japan agreed separately to return the former German holdings in Shantung, although under conditions that left Japan with valuable privileges in the province. For a few years thereafter Great Britain, Japan, the United States, and France attempted to adjust their conflicting interests in China, cooperated in assisting the Peking government, and refrained, on the whole, from aiding particular Chinese factions in the recurrent power struggles. But China was in turmoil, with regional militarism in full tide. Furthermore, a movement against the "unequal treaties" (see below) began to take shape.

*Militarism in China.* During the first years of the republic China had been fractured by rival military regimes to the extent that no one authority was able to subordinate all rivals and create a unified and centralized political structure. The South was detached from Peking's control; but even the southern provinces, and indeed districts within them, were run by different military factions (warlords). Szechwan was a world in itself, divided among several military rulers. The powerful Peiyang Army had split into two major factions whose semi-independent commanders controlled provinces in the Yangtze Valley and in the North; these factions competed for control of Peking. In Manchuria, Chang Tso-lin headed a separate Fengtien army. Shansi was controlled by Yen Hsi-shan. Each separate power group had to possess a territorial base from which to tax and recruit. Arms were produced in many scattered arsenals. Possession of an arsenal and control of ports through which foreign-made arms might be shipped were important elements of power. Most of the foreign powers had agreed in 1919 not to permit arms to be smuggled into China, but this embargo was not entirely effective.

The rival military regimes of republican China

The richer the territorial base, the greater the potential power of the controlling faction. Peking was the great prize because of its symbolic importance as the capital and because the government there regularly received revenues collected by the Maritime Customs Service, administered by foreigners and protected by the powers. Competition for bases brought on innumerable wars, alliances, and betrayals. Even within each military system there was continuous conflict over spoils. To support their armies and conduct their wars, military commanders and their subordinates taxed the people heavily. Money for education and other government services was drained away; revenues intended for the central government were retained in the provinces. Regimes printed their own currency and forced "loans" from merchants and bankers. This chaotic situation partly accounts for the unwillingness of the maritime powers to give up the protection that the treaties with China afforded their nationals.

*The foreign presence.* As a result of several wars and many treaties with China since 1842, foreign powers had acquired a variety of unusual privileges for their nationals. These were specified in the "unequal treaties," which patriotic Chinese bitterly opposed. Hong Kong, Taiwan, and vast areas in Siberia and Central Asia had been detached from China. Dependencies such as Korea, Outer Mongolia, Tibet, and Vietnam had been separated. Leaseholds on Chinese territory were granted to separate powers—such as the southern part of the Liaotung Peninsula and the territory in Shantung around Chiao-chou Bay, which Japan had seized from Germany, to Japan; the New Territories to the adjacent British Crown Colony of Hong Kong; and the Kuang-chou Bay area to France. In most major cities there were concession areas, not governed by China, for the residence of foreigners. Nationals and subjects of the "treaty powers" were protected by extraterritoriality (*i.e.*, they were subject only to the civil and criminal laws of their own countries); this extended to foreign business enterprises in China, providing a great advantage in competition with Chinese firms, which was enhanced when foreign factories or banks were located in concession areas under foreign protection. The Chinese had to compete with foreign ships in Chinese rivers and coastal waters, with foreign mining companies in the interior, and with foreign banks that circulated their own notes. Foreign trade also had a great advantage because there could be no protective tariff to favour Chinese products.

Christian missionaries operated many schools, hospitals, and other philanthropic enterprises in China, all protected by extraterritoriality. The separate school system, outside of Chinese governmental control, was a sore point for nationalists, who regarded the education of Chinese youth as a Chinese prerogative. There were bodies of foreign troops on Chinese soil and naval vessels in its rivers and ports to enforce treaty rights. Bound by a variety of interlocking treaties, the Chinese government was not fully sovereign in China. Past regimes had accumulated a vast foreign debt against which central government revenues were pledged for repayment. All this was the foreign imperialism against which the KMT launched its attack after being reorganized along Bolshevist lines.

*Reorganization of the KMT.* The KMT held its First National Congress in Canton on Jan. 20–30, 1924. Borodin, who had reached Canton in October 1923, began to advise Sun in the reorganization of his party. He prepared a constitution and helped draft a party program as a set of basic national policies. Delegates from throughout China and from overseas branches of the party adopted the program and the new constitution. The program announced goals of broad social reform and a fundamental readjustment of China's international status. Its tone was nationalistic; it identified China's enemies as imperialism and militarism. It singled out farmers and labourers as classes for special encouragement but also appealed to intellectuals, soldiers, youth, and women. It threatened the position of landlords in relation to tenants and of employers in relation to labour. Western privileges were openly menaced.

KMT goals

The constitution described a centralized organization, modeled on the Soviet Communist Party, with power concentrated in a small, elected group and with a descending hierarchy of geographical offices controlled by executive committees directed from above. Members were pledged to strict discipline and were to be organized in tight cells. Where possible they were to penetrate and try to gain control of such other organizations as labour unions, merchant associations, schools, and parliamentary bodies at all levels. Sun was designated as leader of the party and had veto rights over its decisions. The congress elected a central executive committee and a central supervisory committee to manage party affairs and confirmed Sun's decision to admit Communists, though this was opposed by numerous party veterans, who feared the KMT itself might be taken over. A few Communists, including Li Dazhao, were elected to the executive committee.

Organization of the executive committee

The executive committee set up a central headquarters in Canton. It also decided to strengthen the party throughout the country by deputizing most of its leaders to manage regional and provincial headquarters and by recruiting new members. A military academy was planned for training a corps of young officers, loyal to the party, who would become lower level commanders in a new national revolutionary army that was to be created. Borodin provided funds for party operations, and the Soviet Union promised to underwrite most of the expenses of, and to provide training officers for, the military academy. Chiang Kai-shek was chosen to be the first commandant of the academy and Liao Chung-k'ai to be the party representative, or chief political officer.

From February to November 1924 Sun and his colleagues had some success in making the KMT's influence felt nationally; they also consolidated the Canton base, although it was still dependent upon mercenary armies. The military academy was set up at Whampoa, on an island south of Canton, and the first group of some 500 cadets was trained. In September Sun began another northern campaign in alliance with Chang Tso-lin against Ts'ao K'un and Wu P'ei-fu, who now controlled Peking. The campaign was interrupted, however, when Wu's subordinate, Feng Yü-hsiang, betrayed his chief and seized Peking on October 23, while Wu was at the front facing Chang Tso-lin. Feng and his fellow plotters invited Sun to Peking to participate in the settlement of national affairs, while Feng and Chang invited Tuan Ch'i-jui to come out of retirement and take charge of the government. Sun accepted the invitation and departed for the North on November 13. Before he arrived in Peking, however, he fell gravely ill with incurable cancer of the liver. He died in Peking on March 12, 1925.

**Struggles within the two-party coalition.** After Sun's death the KMT went through a period of inner conflict, although it progressed steadily, with Russian help, in bringing the Kwangtung base under its control. The conflict was caused primarily by the radicalization of the party under the influence of the Communists. They organized labour unions and peasant associations and pushed class struggle and the anti-imperialist movement.

*Clashes with foreigners.* On May 30, 1925, patriotic students, engaged in an anti-imperialist demonstration in Shanghai, clashed with foreign police. The British captain in charge ordered the police to fire upon a crowd that he believed was about to rush his station. Some 12 Chinese were killed in the May Thirtieth Incident, including students. This aroused a nationwide protest and set off a protracted general strike in Shanghai. A second incident occurred on June 23, when French and British marines exchanged fire with Whampoa cadets who were part of an anti-imperialist parade, killing 52 Chinese, many of them civilians, and wounding at least 117; which side had fired first became a matter of dispute. This set off a strike and boycott against Britain, France, and Japan, which was later narrowed to Britain alone. The strike and boycott, led mainly by Communists, lasted for 16 months and seriously affected British trade. These incidents intensified hostility toward foreigners and their special privileges, enhanced the image of the Soviet Union, and gained support for the KMT, which promised to end the unequal treaties. By January 1926 the KMT could claim some 200,000 members. The CCP's membership grew from less than 1,000 in May 1925 to about 10,000 by the end of that year.

Strikes and boycotts against foreigners

*KMT opposition to radicals.* The two parties competed for direction of nationalist policy, control of mass organizations, and recruitment of new members. Under Comintern coaching, the Communist strategy was to try to split the KMT, drive out its conservative members, and turn it to an ever more radical course. In August 1925, KMT conservatives in Canton tried to stop the leftward trend. One of the strongest advocates of the Nationalists' Soviet orientation, Liao Chung-k'ai, was assassinated. In retaliation, Borodin, Chiang Kai-shek, and Wang Ching-wei deported various conservatives. A group of KMT veterans in the North then ordered the expulsion of Borodin and the Communists and the suspension of Wang Ching-wei; they set up a rival KMT headquarters in Shanghai. The left-wing leaders in Canton then held a Second National Congress in January 1926, confirming the radical policies and the Soviet alliance. But as the Soviet presence became increasingly overbearing, as the Canton–Hong Kong strike and boycott dragged on, and as class conflict intensified in the South, opposition to the radical trend grew stronger, particularly among military commanders.

Rise of Chiang Kai-shek

Chiang Kai-shek, now commander of the National Revolutionary Army, took steps in March to curb the Communists and to send away several Soviet officers whom he believed were scheming with Wang Ching-wei against him. In a readjustment of party affairs, Communists no longer were permitted to hold high offices in the central headquarters, and Wang Ching-wei went into retirement in France. Chiang also demanded Comintern support of a northern military campaign and the return of Gen. V.K. Blücher as his chief military adviser. Blücher, who used the pseudonym Galen in China, was a commander in the Red Army who had worked with Chiang in 1924 and 1925 in developing the Whampoa Military Academy and forming the National Revolutionary Army. Blücher returned to Canton in May and helped refine plans for the Northern Expedition, which began officially in July, with Chiang as commander in chief.

*The Northern Expedition.* In the Northern Expedition the outnumbered southern forces were infused with revolutionary spirit and fought with great élan. They were assisted by propaganda corps, which subverted enemy troops and agitated among the populace in the enemy's rear. Soviet military advisers accompanied most of the divisions, and Soviet pilots reconnoitred the enemy positions. The army was well financed at the initial stages because of

fiscal reforms in Kwangtung during the previous year, and many enemy divisions and brigades were bought over. Within two months the National Revolutionary Army gained control of Hunan and Hupeh, and by the end of the year it had taken Kiangsi and Fukien. The Nationalist government moved its central headquarters from Canton to the Wu-han cities of the Yangtze. By early spring of 1927, revolutionary forces were poised to attack Nanking and Shanghai.

The political situation, however, was unstable. Hunan and Hupeh were swept by a peasant revolt marked by violence against landlords and other rural power holders. Business in the industrial and commercial centre of the middle Yangtze, the Wu-han cities, was nearly paralyzed by a wave of strikes. Communists and KMT leftists led this social revolution. In January the British concessions in Han-k'ou and Chiu-chiang were seized by Chinese crowds. The British government had just adopted a conciliatory policy toward China, and it acquiesced in these seizures, but it was readying an expeditionary force to protect its more important position in Shanghai. Foreigners and many upper-class Chinese fled from the provinces under Nationalist control. The northern armies began to form an alliance against the Southerners.

The split in the KMT

Conservative Nationalist leaders in Shanghai mobilized against the headquarters in Wu-han. There was a deep rift within the revolutionary camp itself; the leftists at Wu-han, guided by Borodin, pitted themselves against Chiang and his more conservative military supporters, who were also laying plans against the leftists. Resolutions of the CCP's Central Committee in January 1927, showed apprehension of a counterrevolutionary tide against their party, Soviet Russia, and the revolutionary peasant and workers' movement; they feared a coalition within the KMT and its possible alliance with the imperialist powers. The central leadership resolved to check revolutionary excesses and give all support to the KMT leadership at Wu-han. Others within the CCP, notably Mao Zedong, disagreed; they believed the mass revolution should be encouraged to run its course.

*Expulsion of Communists from the KMT.* The climax of the conflict came after Nationalist armies had taken Shanghai and Nanking in March. Nanking was captured on March 23, and the following morning Nationalist troops looted foreign properties, attacked the British, U.S., and Japanese consulates, and killed seven foreigners. In Shanghai a general strike led by Communists aroused fears that Chinese might seize the International Settlement and French concession, now guarded by a large international expeditionary force. Conservative Nationalist leaders, some army commanders, and Chinese business leaders in Shanghai encouraged Chiang to expel the Communists and suppress the Shanghai General Labour Union. On April 12–13, gangsters and troops bloodily suppressed the guards of the General Labour Union, arrested many Communists, and executed large numbers. Similar suppressions were carried out in Canton, Nanking, Nan-ch'ang, Fuchou, and other cities under military forces that accepted Chiang's instructions. The KMT conservatives then established a rival Nationalist government in Nanking.

Wang Ching-wei had returned to China via the Soviet Union. Arriving in Shanghai, he refused to participate in the expulsions and went secretly to Wu-han, where he again headed the government. In July, however, the leftist Nationalist leaders in Wu-han, having learned of a directive by Joseph Stalin to Borodin to arrange for radicals to capture control of the government, decided to expel the Communists and invite the Soviet advisers to leave. The leftist government thereby lost important bases of support; furthermore, it was ringed by hostile forces and cut off from access to the seas, and it soon disintegrated.

The CCP went into revolt. Using its influence in the Cantonese army of Chang Fa-k'uei, it staged an uprising at Nan-ch'ang on August 1 and then attempted an "Autumn Harvest" uprising in several central provinces. Both efforts failed. In December Communist leaders in Canton started a revolt known as the Canton Commune. They captured the city with much bloodshed, arson, and looting; but this uprising was quickly suppressed, also with much slaughter. Between April and December 1927 the CCP lost most of its membership by death and defection. A few score leaders and some scattered military bands then began the process of creating military bases in the mountains and plains of central China, remote from centres of Nationalist power.

The now more conservative KMT resumed its Northern Expedition in the spring of 1928 with a reorganized National Revolutionary Army. In the drive on Peking it was joined by the National People's Army under Feng Yü-hsiang, part of the Kwangsi army, and the Shansi army of Yen Hsi-shan. In early June they captured Peking, from which Chang Tso-lin and the Fengtien army withdrew for Manchuria. As his train neared Mukden (modern Shenyang), Chang died in an explosion arranged by a few Japanese officers without knowledge of the Japanese government. Japan did not permit the Nationalist armies to pursue the Fengtien army into Manchuria, hoping to keep that area out of KMT control. By the end of the Northern Expedition the major warlords had been defeated by the Nationalists, whose armies now possessed the cities and railways of eastern China. On October 10 the Nationalists formally established a reorganized National Government of the Republic of China, with its capital at Nanking.

Nationalist control of eastern China

**The National Government from 1928 to 1937.** The most serious immediate problem facing the new government was the continuance of military separatism. The government had no authority over the vast area of western China, and even regions in eastern China were under the rule of independent regimes that had lately been part of the Nationalist coalition. After an unsuccessful attempt at negotiations, Chiang launched a series of civil wars against his former allies. By 1930 one militarist regime after another had been reduced to provincial proportions, and Nanking's influence was spreading. Explained in material terms, Chiang owed his success to the great financial resources of his base in Kiangsu and Chekiang and to foreign arms. Quick recognition by the foreign powers brought the National Government the revenues collected by the efficient Maritime Customs Service; when the powers granted China the right to fix its own tariff schedules, that revenue increased.

Although the aim of constitutional, representative government was asserted, the National Government at Nanking was in practice personally dominated by Chiang Kai-shek. The army and the civil bureaucracy were marked by factional divisions, which Chiang carefully balanced against one another so that ultimate decision making was kept in his own hands. The KMT was supposed to infuse all government structures and to provide leadership, but the army came to be the most powerful component of government. Chiang's regime was marked by a military orientation, which external circumstances reinforced.

Nevertheless, the Nationalists did much to create a modern government and a coherent monetary and banking system and to improve taxation. They expanded the public educational system, developed a network of transportation and communication facilities, and encouraged industry and commerce. Again it was urban China that mainly benefited; little was done to modernize agriculture or to eradicate disease, illiteracy, and underemployment in the villages, hamlets, and small towns scattered over a continental territory. With conscription and heavy taxation to support civil war, and a collapsing export market for commercial crops, rural economic conditions may have grown worse during the Nationalist decade.

The National Government during its first few years in power had some success in reasserting China's sovereignty. Several concession areas were returned to Chinese control, and the foreign powers assented to China's resumption of tariff autonomy. Yet these were merely token gains; the unequal treaties were scarcely breached. The country was in a nationalistic mood, determined to roll back foreign economic and political penetration. Manchuria was a huge and rich area of China in which Japan had extensive economic privileges, possessing part of the Liaotung Peninsula as a leasehold and controlling much of southern Manchuria's economy through the South Manchurian Railway. The Chinese began to develop Hu-lu-tao, in Liaotung, as a port to rival Dairen (modern Lü-ta) and

to plan railways to compete with Japanese lines. Chang Hsüeh-liang, Chang Tso-lin's son and successor as ruler of Manchuria, was drawing closer to Nanking and sympathized with the Nationalists' desire to rid China of foreign privilege.

Japanese interest in Manchuria

For Japan, Manchuria was regarded as vital. Many Japanese had acquired a sense of mission that Japan should lead Asia against the West. The Great Depression had hurt Japanese business, and there was deep social unrest. Such factors influenced many army officers to regard Manchuria as the area where Japan's power must be consolidated, especially officers of the Kwantung Army, which protected Japan's leasehold in the Liaotung Peninsula and the South Manchurian Railway.

*Japanese aggression.* In September 1931 a group of officers in the Kwantung Army set in motion a plot to compel the Japanese government to extend its power in Manchuria. The Japanese government was drawn, step by step, into the conquest of Manchuria and the creation of a regime known as Manchukuo. China was unable to prevent Japan from seizing this vital area. In 1934, after long negotiations, Japan acquired the Soviet interest in the Chinese Eastern Railway, thus eliminating the last legal trace of the Soviet sphere of influence there. During 1932–35 Japan seized more territory bordering on Manchuria. In 1935 it attempted to detach Hopeh and Chahar from Nanking's control and threatened Shansi, Suiyüan, and Shantung. The National Government's policy was to trade space for time in which to build military power and unify the country. Its slogan "Unity before resistance" was directed principally against the Chinese Communists.

*War between Nationalists and Communists.* In the meantime, the Communists had created 15 rural bases in central China, and they established a soviet government, the Kiangsi Soviet, on Nov. 7, 1931. Within the soviet regions the Communist leadership expropriated and redistributed land and in other ways enlisted the support of the poorer classes. The Japanese occupation of Manchuria and an ancillary localized war around Shanghai in 1932 distracted the Nationalists and gave the Communists a brief opportunity to expand and consolidate. But the Nationalists in late 1934 forced the Red armies to abandon their bases and retreat. Most of the later Communist leaders, including Mao Zedong, Zhu De (Chu Teh), Zhou Enlai, Liu Shaoqi (Liu Shao-ch'i), and Lin Biao (Lin Piao), marched and fought their way across western China. By mid-1936 the remnants of several Red armies had gathered in an impoverished area in northern Shensi, with headquarters located in the town of Yen-an, which lent its name to the subsequent period (1936–45) of CCP development.

The Long March

During the Long March Mao Zedong rose to preeminence in the CCP leadership. In the early 1930s he had engaged in bitter power struggles with other party leaders and actually had found himself in a fairly weak position at the start of the Long March campaigns; but in January 1935 a rump session of the CCP Politburo confirmed Mao in the newly created post of chairman. It was also during the Long March that the CCP began to develop a new political strategy—a united front against Japan. It was first conceived as an alliance of patriotic forces against Japan and the National Government; but as Japan's pressure on China and the pressure of the Nationalist armies against the weakened Red armies increased, the Communist leaders began to call for a united front of all Chinese against Japan alone. Virtually all classes and various local regimes supported this, and the Communists moderated their revolutionary program and terminated class warfare in their zone of control.

Chiang was determined, however, to press on with his extermination campaign. He ordered the Manchurian army under Chang Hsüeh-liang, now based in Sian, and the Northwestern army under Yang Hu-ch'eng to attack the Communist forces in northern Shensi. Many officers in these armies sympathized with the Communist slogan "Chinese don't fight Chinese"; they preferred to fight Japan, a sentiment particularly strong in the homeless Manchurian army. Chang Hsüeh-liang was conducting secret negotiations with the Communists and had suspended the civil war. In December 1936 Chiang Kai-shek flew to Sian to order Chang and Yang to renew the anti-Communist campaign. Under pressure from subordinates, Chang detained Chiang on the morning of December 12 (this became known as the Sian Incident).

The Sian Incident

*The United Front against Japan.* Fearing that, if Chiang were killed, China would be plunged into renewed disorder, the nation clamoured for his release. The Soviet Union quickly denounced the captors and insisted that Chiang be freed (the Soviet Union needed a united China opposing Japan, its potential enemy on the east). The CCP leaders also decided that Chiang's release would serve China's interests as well as their own, if he would accept their policy against Japan. Zhou Enlai and several other Communist leaders flew to Sian to try to effect this. Chang Hsüeh-liang finally agreed to free his captive, with the understanding that Chiang would call off the civil war and unite the country against the invader. On December 25 Chiang was freed.

The two Chinese parties began protracted and secret negotiations for cooperation, each making concessions. But it was not until September 1937, after the Sino-Japanese War had begun, that the National Government formally agreed to a policy of cooperation with the CCP. For its part, the CCP publicly affirmed its adherence to the realization of Sun Yat-sen's Three Principles of the People, its abandonment of armed opposition to the KMT and of the forcible confiscation of landlords' property, the substitution of democracy for its soviet government, and the reorganization of the Red Army as a component of the national army under the central government.

## THE WAR AGAINST JAPAN (1937–45)

**The Sino-Japanese War.** On July 7, 1937, a minor clash between Japanese and Chinese troops near Pei-p'ing (Peking's name under the National Government) finally led the two nations into war. The Japanese government tried for several weeks to settle the incident locally, but China's mood was highly nationalistic and public opinion clamoured for resistance to further aggression. In late July, new fighting broke out. The Japanese quickly took Pei-p'ing and captured Tientsin. On August 13 savage fighting broke out in Shanghai. By now the prestige of both nations was committed, and they were locked in a war.

*Phase one.* As never before in modern times, the Chinese united themselves against a foreign enemy. China's standing armies in 1937 numbered some 1,700,000 men, with 500,000 in reserve. Japan's naval and air superiority were unquestioned. But Japan could not commit its full strength to campaigns in China; the main concern of the Japanese Army was the Soviet Union, while for the Japanese Navy it was the United States.

Early Japanese victories

During the first year of the undeclared war, Japan won victory after victory against sometimes stubborn Chinese resistance. By late December, Shanghai and Nanking had fallen. But China had demonstrated to the world its determination to resist the invader; this gave the government time to search for foreign support. China found its major initial help from the Soviet Union. On Aug. 21, 1937, the Soviet Union and China signed a nonaggression pact, and the former quickly began sending munitions, military advisers, and hundreds of aircraft with Soviet pilots. Japanese forces continued to win important victories. By mid-1938 Japanese armies controlled the railway lines and major cities of northern China. They took Canton on October 12, stopping the railway supply line to Wu-han, the temporary Chinese capital, and captured Han-k'ou, Hanyang, and Wu-ch'ang on October 25–26. The Chinese government and military command moved to Chungking in Szechwan, farther up the Yangtze and behind a protective mountain screen.

At the end of this first phase of the war, the National Government had lost the best of its modern armies, its air force and arsenals, most of China's modern industries and railways, its major tax resources, and all the ports through which military equipment and civilian supplies might be imported. But it still held a vast though backward territory and had unlimited manpower reserves. So long as China continued to resist, Japan's control over the conquered eastern part of the country would be difficult.

*Phase two: stalemate and stagnation.* During the second stage of the war (1939–43) the battle lines changed very little, although there were many engagements of limited scale. Japan tried to bomb Free China into submission; Chungking suffered repeated air raids in which thousands of civilians were killed. In 1940 Japan set up a rival government in Nanking under Wang Ching-wei. But the Chinese would not submit. Hundreds of thousands migrated to west China to continue the struggle. Students and faculties of most eastern colleges took the overland trek to makeshift quarters in distant inland towns. Factories and skilled workers were reestablished in the west. The government rebuilt its shattered armies and tried to purchase supplies from abroad.

Allied aid to China

In 1938–40 the Soviet Union extended credits for military aid of $250,000,000, while the United States, Great Britain, and France granted some $263,500,000 for civilian purchases and currency stabilization. Free China's lines of supply were long and precarious; when war broke out in Europe, shipping space became scarce. After Germany's conquest of France in the spring of 1940, Britain bowed to Japanese demands and temporarily closed Rangoon to military supplies for China (July–September). In September 1940 Japan seized control of northern Indochina and closed the supply line to K'un-ming. The Soviet Union had provided China its most substantial military aid, but when Germany attacked the Soviet Union in June 1941, this aid virtually ceased. By then, however, the United States had sold China 100 fighter planes—the beginning of a U.S. effort to provide air protection.

In addition to bombing, the civilian population in Free China endured great hardships. Manufactured goods were scarce, and hoarding drove up prices. The government did not have the means to carry out rationing and price control, though it did supply government employees with rice. The government's sources of revenue were limited, yet it supported a large bureaucracy and an army of more than 3,000,000 conscripts. The government resorted to printing currency inadequately backed by reserves. Inflation grew until it was nearly uncontrollable. Between 1939 and 1943 the morale of the bureaucracy and military officers declined. Old abuses of the Chinese political system reasserted themselves—factional politics and corruption, in particular. The protracted war progressively weakened the Nationalist regime.

The war had the opposite effect upon the CCP. The Communist leaders had survived 10 years of civil war and had developed a unity, camaraderie, and powerful sense of mission. They had learned to mobilize the rural population and to wage guerrilla warfare. In 1937 the CCP had about 40,000 members and the poorly equipped Red Army numbered perhaps 100,000. By agreement with the National Government, the Red Army was renamed the Eighth Route Army (later the 18th Group Army); Zhu De and Peng Dehuai (P'eng Te-huai) served as commander and vice commander, and Lin Biao, Ho Lung, and Liu Bocheng (Liu Po-ch'eng) were in charge of its three divisions. The Communist base in the northwest covered parts of three provinces with a backward economy and a population of about 1,500,000. Operating within the general framework of the United Front against Japan, the leaders of the Eighth Route Army adopted a strategy that used their experience in guerrilla warfare. They sent small columns into areas of northern China that the Japanese Army had overrun but lacked the manpower to control; there they incorporated remnant troops and organized the population to supply food, recruits, and sanctuaries for guerrilla units attacking small Japanese garrisons.

Communist guerrilla warfare

Early in the period of united resistance, the government permitted the creation of the New 4th Army from remnants of Communist troops left in Kiangsi and Fukien at the time of the Long March. Commanded by Gen. Ye Ting (Yeh T'ing), with Xiang Ying (Hsiang Ying), a Communist, as chief of staff, this force of 12,000 officers and men operated behind Japanese lines near Shanghai with great success. Its strategy included guerrilla tactics, organizing resistance bases, and recruitment. This army grew to more than 100,000 in 1940; by then it operated in a wide area on both sides of the lower Yangtze.

Thus the CCP revitalized itself. It recruited rural activists and patriotic youths from the cities and systematically strengthened the ranks by continuous indoctrination and by expelling dissident and ineffective party members.

*Renewed Communist-Nationalist conflict.* There were numerous clashes between Communists and Nationalists as their military forces competed for control of enemy territory and as the Communists tried to expand their political influence in Nationalist territory through propaganda and secret organizing. Though both sides continued the war against Japan, each was fighting for its own ultimate advantage. Bitter anti-Communist sentiment in government circles found its most violent expression in the New 4th Army Incident of January 1941.

New 4th Army Incident

The government had ordered the New 4th Army to move north of the Huang Ho and understood that its commanders had agreed to do so as part of a demarcation of operational areas. But most of the army had moved into northern Kiangsu (south of the Huang) and, together with units of the 18th Group Army, was competing with government troops for control of bases there and in southern Shantung. Ye Ting and Xiang Ying stayed at the army's base south of the Yangtze. Apparently believing that Ye did not intend to move northward, government forces attacked the base on Jan. 6, 1941. The outnumbered Communists were defeated, Ye Ting and some 2,000 others were captured, Xiang Ying was killed, and both sides suffered heavy casualties. Ignoring Chiang's order to dissolve the New 4th Army, the Communist high command named Chen Yi (Ch'en I) as its new commander and Liu Shaoqi as political commissar.

The danger of renewed civil war caused widespread protest from China's civilian leaders. The People's Political Council, a multiparty advisory body formed in 1938 as an expression of united resistance, debated the issue and later tried to mediate. Neither the KMT nor the CCP was willing to push the conflict to open civil war in 1941. The government deployed many of its best divisions in positions to prevent the Communist forces from further penetration of Nationalist-held territories and to weaken the CCP through a strict economic blockade.

**The international alliance against Japan.** The United States had broken the Japanese diplomatic code, and by July 1941 it knew that Japan hoped to end the undeclared war in China and was preparing for a southward advance toward British Malaya and the Dutch East Indies, planning first to occupy southern Indochina and Thailand even at the risk of war with Britain and the United States.

*U.S. aid to China.* One U.S. response was the decision to send large amounts of arms and equipment to China, along with a military mission to advise on their use. The underlying strategy was to revitalize China's war effort as a deterrent to Japanese military and naval operations southward. The Nationalist army was ill equipped to fight the Japanese in 1941. Its arsenals were so lacking in nonferrous metals and explosives that they could not produce effectively. The maintenance of millions of ill-trained and underequipped troops was a heavy drain on the economy. There was no possibility that the United States could arm such numbers from its limited stocks while building up its own forces and assisting many other nations. And there was a formidable logistics problem in shipping supplies along the 715-mile (1,150-kilometre) Burma Road, which extended from K'un-ming to Lashio, the Burma terminus of the railway and highway leading to Rangoon.

The U.S. military mission

By December 1941 the United States had sent a military mission to China and had implicitly agreed to create a modern Chinese air force, to maintain an efficient line of communications into China, and to arm 30 divisions. Japan's bombing of Pearl Harbor brought the United States into alliance with China, and Great Britain joined the Pacific war as its colonial possessions were attacked. This widening of the Sino-Japanese conflict lifted Chinese morale, but its other early effects were harmful. With the Japanese conquest of Hong Kong on December 25, China lost its air link with the outside world and one of its principal routes for smuggling supplies. By the end of May 1942 the Japanese held most of Burma, having defeated the British, Indian, Burmese, and Chinese

defenders. China was almost completely blockaded. The United States granted China a loan of $500,000,000, and Great Britain stated its willingness to lend £50,000,000, but there was little else China's allies could do.

The solution was found in an air route from Assam, India, to K'un-ming in southwest China—the dangerous "Hump" route along the southern edge of the Himalayas. In March 1942 the China National Aviation Corporation (CNAC) began freight service over the Hump, and the United States began a transport program the next month. But shortages and other difficulties had to be overcome, and not until December 1943 were cargo planes able to equal the tonnage carried along the Burma Road by trucks two years before. This was much less than China's needs for gasoline and military equipment and supplies.

*Conflicts within the international alliance.* China's alliance with the United States and Great Britain was marked by deep conflict. Great Britain gave highest priority to the defeat of its main enemy, Germany. The U.S. Navy in the Pacific had been seriously weakened by the Japanese air attack at Pearl Harbor and required many months to rebuild. During the winter of 1941–42 the grand strategy of the United States and Great Britain called for the defeat of Germany first and then an assault across the Pacific against Japan's island empire. China was relegated to a low position in U.S. strategic planning. The United States aimed to keep China in the war and enable it to play a positive role in the final defeat of Japan on the continent. Chiang Kai-shek, on the other hand, envisaged a joint strategy by the United States, the British Commonwealth, and China over the whole Pacific area, with China playing a major role. He demanded an equal voice in Allied war planning, which he never received, although U.S. Pres. Franklin D. Roosevelt was generally solicitous. From the fundamentally different outlooks of President Chiang, British Prime Minister Winston Churchill, and President Roosevelt and because of the divergent national interests of China, the British Commonwealth, and the United States, there followed many controversies that had powerful repercussions in China and led to frustrations and suspicions among the partners.

After the fall of Burma, a controversy developed over whether the principal Chinese and U.S. effort against Japan should be devoted to building up U.S. air power based in China or to reform of the Chinese army and its training and equipment for a combat role. Chiang advocated primary reliance on U.S. air power to defeat Japan. Several high-ranking U.S. generals, on the other hand, emphasized creation of a compact and modernized Chinese ground force able to protect the airfields in China and to assist in opening an overland supply route across northern Burma. Already in India, the United States was training two Chinese divisions from remnants of the Burma campaign, plus artillery and engineering regiments (this became known as X-Force). Also in training were Chinese instructors to help retrain other divisions in China. Both air development and army modernizing were being pushed in early 1943, with a training centre created near K'un-ming to reenergize and reequip select Chinese divisions (called Y-Force), and a network of airfields was being built in southern China. This dual approach caused repeated conflict over the allocation of scarce airlift space.

Chinese-U.S. disagreement

By the end of 1943 the China-based U.S. 14th Air Force had achieved tactical parity with the Japanese over central China, was beginning to bomb Yangtze shipping, and had conducted a successful raid on Japanese airfields on Taiwan. A second training centre had been started at Kuei-lin to improve 30 more Chinese divisions (Z-Force). The campaign to open a land route across northern Burma had run into serious difficulty. At the Cairo Conference in November, Chiang met Churchill and Roosevelt for the first time. The Cairo Declaration promised the return of Manchuria, Taiwan, and the Pescadores Islands to China and the liberation of Korea. The three allies pledged themselves to "persevere in the . . . prolonged operations necessary to procure the unconditional surrender of Japan." These words, however, concealed deep differences over global strategy. U.S. planners realized that Japan might be approached successfully through the south and central Pacific and that the Soviet Union would enter the war against Japan after Germany's defeat; hence, the importance of China to U.S. grand strategy declined. Churchill was unwilling to use naval resources, needed for the forthcoming European invasion, in a seaborne invasion of Burma to help reopen China's supply line. Yet Chiang had demanded a naval invasion of Burma as a condition to committing the Y-Force to assist in opening his supply line. Shortly after Cairo, Churchill and Roosevelt agreed to set aside the seaborne invasion of Burma, and when Chiang learned of this he requested enormous amounts of money, supplies, and air support, asserting that otherwise Japan might succeed in eliminating China from the war. The United States did not accede, and Chinese–American relations began to cool.

*Phase three: approaching crisis (1944–45).* China was in crisis in 1944. Japan faced increasing pressure in the Pacific and threats to its supply bases and communications lines in China as well as to nearby shipping. Its response was twofold—first, to attack from Burma toward Assam to cut the supply lines or capture the airfields at the western end of the Hump, and, second, to capture the railway system in China from north to south and seize the eastern China airfields used by the United States.

The British and Indian army defeated the Japanese attack on Assam (March–July 1944) with help from transport planes withdrawn from the Hump. But the Japanese campaign in China, known as Ichigo, showed up the weakness, inefficiency, and poor command of the Chinese armies after nearly seven years of war. During April and May the Japanese cleared the Pei-p'ing–Han-k'ou railway between the Huang and the Yangtze rivers. Chinese armies nominally numbering several hundred thousand men were unable to put up effective resistance. Peasants in Honan attacked the collapsing Chinese armies, their recent oppressors.

Chinese military weaknesses

The second phase of the Ichigo campaign was a Japanese drive southward from Han-k'ou and northwestward from Canton to take Kuei-lin and open the communication line to the India–China border. By November the Chinese had lost Kuei-lin, Liu-chou, and Nan-ning, and the Japanese were approaching Kuei-yang on the route to Chungking and K'un-ming. This was the high watermark of Japan's war in China. Thereafter, it withdrew experienced divisions for the defense of its overextended empire, and China finally began to benefit from the well-trained X-Force when two divisions were flown in from Burma in December to defend K'un-ming.

Meanwhile, the Chinese government was involved in a crisis of relations with the United States, which contended that the Chinese Army must be reformed, particularly in its command structure, and that lend-lease supplies must be used more effectively. There were also many subsidiary problems. Gen. Joseph Stilwell, the executor of disagreeable U.S. policies in China, had developed an unconcealed disdain for Chiang, whom he nominally served as chief of staff. Stilwell was an effective troop commander, and Roosevelt requested that Chiang place Stilwell in command of all Chinese forces. In the context of Chinese politics, in which control of armies was the main source of power, President Chiang's compliance was virtually inconceivable. He declined the request and asked for Stilwell's recall. Roosevelt agreed, but thereafter his relations with Chiang were no longer cordial. Stilwell was replaced by Gen. Albert C. Wedemeyer.

*Nationalist deterioration.* The military weakness in 1944 was symptomatic of a gradual deterioration that had taken place in most aspects of Nationalist Chinese public life. Inflation began to mount alarmingly as the government pumped in large amounts of paper currency to make up its fiscal deficits. Salaries of government employees, army officers, teachers, and all those on wages fell far behind rising prices. For most, this spelled poverty amid growing war-weariness. Dissatisfaction with the government's policies spread among intellectuals. Inflation gave opportunities for some groups to profit through hoarding of needed goods, smuggling of high-value commodities, black market currency operations, and graft. Corruption spread in the bureaucracy and the armed forces. As the war dragged on,

governmental suppression of dissidence grew oppressive. Secret police activity and efforts at thought control were aimed not only against Communists but also against all influential critics of the government or the KMT.

*Communist growth.* The Communist armies were growing rapidly in 1943 and 1944. According to U.S. war correspondents visiting the Yen-an area in May 1944 and to a group of U.S. observers that established itself there in July, the Communists professed allegiance to democracy and to continued cooperation with the National Government in the war effort. There was convincing evidence that the areas under Communist control extended for hundreds of miles behind Japanese lines in northern and central China.

Communist economic policy

This situation was the result of many factors. Communist troop commanders and political officers in areas behind Japanese lines tried to mobilize the entire population against the enemy. Party members led village communities into greater participation in local government than had been the case before. They also organized and controlled peasants' associations, labour unions, youth leagues, and women's associations. They linked together the many local governments and the mass organizations and determined their policies. Because of the need for unity against Japan, the Communist organizers tended to follow reformist economic policies. The party experimented with various forms of economic cooperation to increase production; one of these was mutual-aid teams in which farmers temporarily pooled their tools and draft animals and worked the land collectively. In areas behind Japanese lines some mutual-aid teams evolved into work-and-battle teams composed of younger peasants; when danger threatened, the teams went out to fight as guerrillas under direction of the local Red army, and when the crisis passed they returned to the fields. The party recruited into its ranks the younger leaders who emerged from populist activities. Thus, it penetrated and to some extent controlled the multitude of villages in areas behind Japanese lines. As the Japanese military grip weakened, the experienced Communist armies and political organizers spread their system of government ever more widely. By the time of the CCP's Seventh Congress in Yen-an (April–May 1945), the party claimed to have an army of more than 900,000 and a militia of more than 2,000,000. It also claimed to control areas with a total population of 90,000,000. These claims were disputable, but the great strength and wide geographical spread of Communist organization was a fact.

*Efforts to prevent civil war.* Between May and September 1944, representatives of the government and the CCP carried on peace negotiations at Sian. The main issues were the disposition, size, and command of the Communist armies; the relationship between Communist-organized regional governments and the National Government; and problems of civil rights and legalization of the CCP and its activities in Nationalist areas. Suggestions for a coalition government arose for the first time. No settlement was reached, but it appeared that the antagonists were seeking a peaceful solution. U.S. Vice Pres. Henry Wallace visited Chungking in June and had several discussions with Chiang, who requested U.S. assistance in improving relations between China and the Soviet Union and in settling the Communist problem.

In September 1944, Patrick J. Hurley arrived as Roosevelt's personal representative. Hurley attempted to mediate, first in discussions in Chungking and then by flying to Yen-an on November 7 for a conference with Mao Zedong. But the positions of the two sides could not be reconciled, and the talks broke off in March 1945. Between June and August, Hurley resumed protracted discussions, both indirect and in conferences with high-level representatives from both sides. Each side distrusted the other; each sought to guarantee its own survival, but the KMT intended to continue its political dominance, while the CCP insisted upon the independence of its armies and regional governments under whatever coalition formula might be worked out.

End of the war

The Pacific war ended on Aug. 14, 1945, and the formal Japanese surrender came on September 2. China rejoiced. Yet the country faced enormously difficult problems of reunification and reconstruction and a future clouded by the dark prospect of civil war.

## CIVIL WAR (1945–49)

In a little more than four years after Japan's surrender, the CCP and the People's Liberation Army (PLA) conquered mainland China, and, on Oct. 1, 1949, the People's Republic of China was established, with its capital at Peking. The factors that brought this about were many and complex and subject to widely varying interpretation, but the basic fact was a Communist military triumph growing out of a profound and popularly based revolution. The process may be perceived in three phases: (1) from August 1945 to the end of 1946, the Nationalists and Communists raced to take over Japanese-held territories, built up their forces, and fought many limited engagements while still conducting negotiations for a peaceful settlement; (2) during 1947 and the first half of 1948, after initial Nationalist success, the strategic balance turned in favour of the Communists; (3) the Communists won smashing victories in the latter part of 1948 and 1949.

**A race for territory.** As soon as Japan's impending surrender was known, the commander of the Communist armies, Gen. Zhu De, ordered his men, on August 11, to move into Japanese-held territory and take over Japanese arms, despite Chiang's order that they stand where they were. The United States aided the Chinese government by flying many divisions from the southwest to occupy the main eastern cities, such as Pei-p'ing, Tientsin, Shanghai, and the prewar capital, Nanking. The U.S. Navy moved Chinese troops from South China to other coastal cities, and landed 53,000 marines at Tientsin and Tsingtao to assist in disarming and repatriating Japanese troops but also to serve as a counterweight to the Soviet army in southern Manchuria. Furthermore, U.S. Gen. Douglas MacArthur ordered all Japanese forces in China proper to surrender their arms only to forces of the National Government. They obeyed and thereby were occasionally engaged against Chinese Communist forces.

U.S. aid to the Nationalists

Immediately after the surrender, the Communists sent political cadres and troops into Manchuria. This had been planned long in advance. Gen. Lin Biao became commander of the forces (the Northeast Democratic Allied Army), which incorporated "puppet" troops of Manchukuo and began to recruit volunteers; it got most of its arms from Japanese stocks taken over by the Soviets.

Manchuria was a vast area, with a population of 40,000,000, the greatest concentration of heavy industry and railways in China, and enormous reserves of coal, iron, and many other minerals. The Soviet Union had promised the National Government to withdraw its occupying armies within 90 days of Japan's surrender and to return the region to China. The government was determined to control Manchuria, which was vital to China's future as a world power. But Lin Biao's army attempted to block the entry of Nationalist troops by destroying rail lines and seizing areas around ports of entry. Soon the two sides were locked in a fierce struggle for the corridors into Manchuria, although negotiations were under way in Chungking between Mao and Chiang for a peaceful settlement. The Soviet army avoided direct involvement in the struggle, but it dismantled much industrial machinery and shipped it to the Soviet Union together with hundreds of thousands of Japanese prisoners of war. By the end of 1945 the Nationalists had positioned some of their best U.S.-trained armies in southern Manchuria as far north as Mukden (modern Shen-yang), a strategic rail centre to which Nationalist troops were transported by air. The government's hold was precarious, however, because the Communist 18th Group Army and the New 4th Army had regrouped in North China, abandoning areas south of the Yangtze after a weak bid to take Shanghai. By the end of 1945 Communist forces were spread across a band of provinces from the northwest to the sea. They had a grip on great sections of all the railway lines north of the Lung-hai, vital supply lines for Nationalist armies in the Tientsin–Pei-p'ing area and in Manchuria. The National Government held vast territories in the South and west and had reestablished its authority in the rich provinces

The struggle for Manchuria

of the lower Yangtze Valley and a few important North China cities; it had also assumed civil control on Taiwan.

*Attempts to end the war.* Peace negotiations continued in Chungking between Nationalist and Communist officials after Japan's surrender. An agreement reached on Oct. 10, 1945, called for the convening of a multiparty Political Consultative Council to plan a liberalized postwar government and to draft a constitution for submission to a national congress. Still, the sides were far apart over the character of the new government, control over the Communist liberated areas, and the size and degree of autonomy of the Communist armies in a national military system. Hurley resigned his ambassadorship on November 26, and the next day U.S. Pres. Harry S. Truman appointed Gen. George C. Marshall as his special representative, with the specific mission of trying to bring about political unification and the cessation of hostilities in China.

Marshall arrived in China on December 23. The National Government proposed the formation of a committee of three, with Marshall as chairman, to end the fighting. This committee, with generals Chang Chun and Zhou Enlai as the Nationalist and Communist representatives, met on Jan. 7, 1946. It agreed on January 10 that Chiang and Mao would issue orders to cease hostilities and halt troop movements as of January 13 midnight, with the exception of government troop movements south of the Yangtze and into and within Manchuria to restore Chinese sovereignty. The agreement also called for the establishment in Pei-p'ing of an executive headquarters, equally represented by both sides, to supervise the cease-fire.

Cease-fire agreement

This agreement provided a favourable atmosphere for meetings in Chungking of the Political Consultative Council, composed of representatives of the KMT, the CCP, the Democratic League, the Young China Party, and nonparty delegates. From Jan. 10 to 31, 1946, the Council issued a series of agreed recommendations regarding governmental reorganization, peaceful national reconstruction, military reductions, a national assembly, and the drafting of a constitution. President Chiang pledged that the government would carry out these recommendations, and the political parties stated their intention to abide by them. The next step was meetings of a military subcommittee, with Marshall as adviser, to discuss troop reductions and amalgamation of forces into a single national army.

Early 1946 was the high point of conciliation. It soon became clear, however, that implementation of the various recommendations and agreements was being opposed by conservatives in the KMT, who feared the dilution of their party's control of the government, and by Nationalist generals, who objected to the reduction of their armies. The Communists attempted to prevent the extension of Nationalist military control in Manchuria. On March 17–18 a Communist army attacked and captured a strategic junction between Mukden and Ch'ang-ch'un, the former Manchukuo capital; on April 18, Communists captured Ch'ang-ch'un from a small Nationalist garrison directly after the Soviet withdrawal. On that day Marshall returned to China after a trip to Washington and resumed his efforts to stop the spreading civil war.

*Resumption of fighting.* Each side seemed convinced that it could win by war what it could not achieve by negotiation—dominance over the other. Despite the efforts of Chinese moderates and General Marshall, fighting resumed in July in Manchuria, and in North China the Nationalists attempted massive drives in Kiangsu and Shantung to break the Communist grip on the railways. The Communists launched a propaganda campaign against the United States, playing upon the nationalistic theme of liberation; they were hostile because of the extensive U.S. military and financial assistance to the KMT at the very time that Marshall was mediating. The National Government had become increasingly intransigent, confident of continued U.S. help. To exert pressure and to try to keep the United States out of the civil war, Marshall in August imposed an embargo on further shipment of U.S. arms to China. By the end of the year, however, he realized that his efforts had failed. In January 1947 he left China, issuing a statement denouncing the intransigents on both sides. All negotiations ended in March; the die was cast for war.

In the latter half of 1946, government forces made impressive gains in North China and Manchuria, capturing 165 towns from the enemy. Buoyed by these victories, the government convened a multiparty National Assembly on November 15, despite a boycott by the CCP and the Democratic League. The delegates adopted a new constitution, which was promulgated on New Year's Day. The constitution reaffirmed Sun Yat-sen's Three Principles of the People as the basic philosophy of the state; called for the fivefold division of powers among the executive, legislative, judicial, control, and examination *yüan* ("governmental bodies"); and established the four people's rights of initiation, referendum, election, and recall. The way was prepared for election of both central and local officials, upon which the period of Nationalist tutelage would end.

Economic problems of the National Government

The National Government struggled with grave economic problems. Inflation continued unabated, caused principally by government financing of military and other operations through the printing press: approximately 65 percent of the budget was met by currency expansion and only 10 percent by taxes. Government spending was uncontrolled; funds were dissipated in maintaining large and unproductive garrison forces. Much tax revenue failed to reach the treasury because of malpractices throughout the bureaucracy. Inflation inhibited exports and enhanced the demand for imports. The government had to import large amounts of grain and cotton, but in the months immediately after Japan's surrender it also permitted the import of luxury goods without effective restrictions. As an anti-inflationary measure, it sold gold on the open market. These policies permitted a large gold and U.S. currency reserve, estimated at $900,000,000 at the end of the war, to be cut in half by the end of 1946. Foreign trade was hampered by excessive regulation and corrupt practices.

The spiraling effects of inflation were somewhat curbed by large amounts of supplies imported by the United Nations Relief and Rehabilitation Administration, chiefly food and clothing, a wide variety of capital goods, and materials for the rehabilitation of agriculture, industry, and transportation. In August 1946 the United States sold to China civilian-type army and navy surplus property at less than 20 percent of the estimated procurement cost of $900,000,000. In spite of these and other forms of aid, the costs of civil war kept the budget continuously out of balance. Speculation, hoarding of goods, and black market operations as hedges against inflation continued unabated. The constant depreciation in the value of paper currency undermined morale in all classes dependent upon salaries, including troops, officers, and civilian officials.

By contrast, it appears that in their areas, mostly rural, the Communists practiced a Spartan style of life close to the common people. Morale remained high in the army and was continuously bolstered by indoctrination and effective propaganda. As they had during the war years, Communist troops tried in many ways to win support of the masses. In newly occupied areas social policy was at first reformist rather than revolutionary.

Communist gains in Manchuria

In Manchuria, Lin Biao was forging a formidable army of veteran cadres from North China and natives of Manchuria, now well equipped with Japanese weapons. By 1947 the Communists' Northeast Democratic Allied Army controlled all of Manchuria north of the Sungari River, the east, and much of the countryside in the Nationalist stronghold in the South. There the Nationalists had most of their best trained and equipped divisions; but the troops had been conscripted or recruited in China's southwest, and they garrisoned cities and railways in a distant land. Beginning in January 1947, Lin Biao launched a series of small offensives. By July the Nationalists had lost half of their territory in Manchuria and much matériel; desertions and casualties, caused by indecisive Nationalist leadership and declining troop morale, reduced their forces by half. Lin Biao was not yet strong enough to take Manchuria, but he had the Nationalist armies hemmed up in a few major cities and with only a tenuous hold on the railways leading southward.

**The tide begins to shift.** Although government forces overran Yen-an in March 1947, the strategic initiative passed to the PLA during that year. In midsummer Liu

Bocheng started moving toward the Yangtze; by late in the year the Communists had concentrated strong forces in central China. Chen Yi operated on both sides of the Lung-hai, east of K'ai-feng; Liu Bocheng was firmly established in the Ta-pieh Shan on the borders of Anhwei, Honan, and Hupeh, northeast of Han-k'ou; and Chen Geng (Ch'en Keng) had another army in Honan west of the Pei-p'ing–Han-k'ou railway. These groups cut Nationalist lines of communication, they destroyed protecting outposts along the Lung-hai and Ping-han lines, and they isolated cities.

By the end of 1947 the government forces, according to U.S. military estimates, still numbered some 2,700,000 facing 1,150,000 Communists. But the Nationalists were widely spread and on the defensive. In November, Mao Zedong established the Communist capital at Shih-chia-chuang, a railway centre leading from the Pei-p'ing–Han-k'ou railway into Shansi; this was a measure of the consolidation of the Communist position in North China. In a report to the CCP Central Committee on Dec. 25, 1947, Mao exuded confidence:

> The Chinese people's revolutionary war has now reached a turning point. . . . The main forces of the People's Liberation Army have carried the fight into the Kuomintang Area. . . . This is a turning point in history.

*A land revolution.* One reason for Communist success was the social revolution in rural China. The CCP was now unrestrained by the multiclass alliance of the United Front period. In the middle of 1946, as civil war became more certain, the party leaders launched a land revolution. They saw land redistribution as an integral part of the larger struggle; by encouraging peasants to seize the landlords' fields and other property, the party apparently expected to weaken the government's rural class base and strengthen its own support among the poor. This demanded a decisive attack upon the traditional village social structure. The party leaders believed that to crack the age-old peasant fear of the local elite and overcome traditional respect for property rights required unleashing the hatred of the oppressed. Teams of activists moved through the villages, organizing the poor in "speak bitterness" meetings to struggle against landlords and Nationalist supporters, to punish and often to kill them, and to distribute their land and property. The party tried to control the process in order not to alienate the broad middle ranks among the peasants; but land revolution had a dynamism of its own, and rural China went through a period of terror. Yet apparently the party gained from the revolutionary dynamism; morale was at fever pitch, and, for those who had benefited from land distribution, there was no turning back.

The turning point in the war

*The decisive year, 1948.* The year 1948 was the turning point. In central China, Communist armies of 500,000 men proved their ability to fight major battles on the plains and to capture, though not always hold, such important towns on the Lung-hai as Lo-yang and K'ai-feng. In North China they encircled T'ai-yüan, the capital of Shansi; took most of Chahar and Jehol, provinces on Manchuria's western flank; and recaptured Yen-an, which had been lost in March 1947. The decisive battles were fought in Shantung and Manchuria, where the forces of Chen Yi and Liu Bocheng and those under Lin Biao crushed the government's best armies. For the government it was a year of military and economic disasters.

In Shantung, despite the departure of Chen Yi's forces, Communist guerrillas gradually reduced the government's hold on the railway from Tsingtao to Chi-nan; they penned up about 60,000 government troops in the latter city, an important railway junction. Instead of withdrawing this garrison southward to Suchow, the government left it, for political reasons, to stand and fight. Then Chen Yi's forces returned to Shantung and overwhelmed the dispirited Chi-nan garrison on September 24. This opened the way for a Communist attack upon Suchow, the historic northern shield for Nanking and a vital railway centre.

Beginning in December 1947, a Communist offensive severed all railway connections into Mukden and isolated the Nationalist garrisons in Manchuria. The government armies went on the defensive in besieged cities, partly out of fear that demoralized divisions would defect in the field. Instead of withdrawing from Manchuria before it was too late, the government tried unsuccessfully to reinforce its armies and to supply the garrisons by air. With the fall of Chi-nan, Lin Biao launched his final offensive. He now had an army of 600,000, nearly twice the Nationalist force in Manchuria. He first attacked Chin-chou, the government's supply base on the railway between Chi-nan and Mukden; it fell on October 17. Ch'ang-ch'un fell three days later. The great garrison at Mukden then tried to retake Chin-chou and Ch'ang-ch'un and to open the railway line to the port of Ying-k'ou. In a series of battles, Lin Biao's columns defeated this cream of the Nationalist forces. By early November the Nationalists had lost some 400,000 men as casualties, captives, or defectors.

Lin Biao's final offensive

The government's military operations in the first part of 1948 produced ever larger budget deficits through the loss of tax receipts, dislocation of transportation and productive facilities, and increased military expenditures. Inflation was out of control. In August the government introduced a new currency, the gold yuan, to replace the old notes at the rate of 3,000,000 for one, promising drastic reforms to curtail expenditures and increase revenue. Domestic prices and foreign-exchange rates were pegged, with severe penalties threatened for black market operations. The people were required to sell their gold, silver, and foreign currency to the government at the pegged rate; large numbers did so in a desperate effort to halt the inflation. In Shanghai and some other places the government used Draconian methods to enforce its decrees against speculators, but it apparently could not control its own expenditures or stop the printing presses. Furthermore, the government's efforts to fix prices of food and commodities brought about an almost complete stagnation of economic activity, except for illicit buying and selling at prices far above the fixed levels. Some army officers and government officials were themselves engaged in smuggling, speculation, and other forms of corruption. Then came the loss of Chi-nan and knowledge of the threat in Manchuria. During October the final effort to halt inflation collapsed, with shattering effect to morale in Nationalist-held cities. Prices started rocketing upward once more.

**Communist victory.** Between early November 1948 and early January 1949, the two sides battled for the control of Suchow. Gen. Zhu De concentrated 600,000 men under Chen Yi, Liu Bocheng, and Chen Geng near that strategic centre, which was defended by Nationalist forces of similar size. Both armies were well equipped, but the Nationalists had a superiority in armour and were unopposed in the air. Yet poor morale, inept command, and a defensive psychology brought another disaster to the National Government. One after another its armies were surrounded and defeated in the field. When the 65-day battle was over on January 10, the Nationalists had lost some 500,000 men and their equipment. The capital at Nanking would soon lie exposed.

With Manchuria and most of the eastern region south to the Yangtze in Communist hands, the fate of Tientsin and Pei-p'ing was sealed. The railway corridor between Tientsin and Chang-chia-k'ou was hopelessly isolated. Tientsin fell on January 15 after a brief siege, and Fu Tso-i surrendered Pei-p'ing on the 23rd, allowing a peaceful turnover of China's historic capital and centre of culture.

Thus, during the last half of 1948, the Communist armies had gained control over Manchuria and northeastern China nearly to the Yangtze, except for pockets of resistance. They had a numerical superiority and had captured such huge stocks of rifles, artillery, and armour that they were better equipped than the Nationalists.

Political shifts in 1949

Great political shifts occurred in 1949. Chiang Kai-shek retired temporarily in January, turning over to the vice president, Gen. Li Tsung-jen, the problem of holding together a government and trying to negotiate a peace with Mao Zedong. But Li's peace negotiations (February–April) proved hopeless. The Nationalists were not prepared to surrender; they still claimed to govern more than half of China and still had a large army. General Li tried to secure U.S. support in the peace negotiations and in the military defense of South China, but the U.S. government,

attempting to extricate itself from its entanglement with the collapsing forces of the National Government, pursued a policy of noninvolvement.

Final Communist victories

When peace negotiations broke down, Communist armies crossed the Yangtze virtually unopposed; the National Government abandoned its undefensible capital on April 23 and moved to Canton. In succession, Communist forces occupied Nanking (April 24), Han-k'ou (May 16–17), and Shanghai (May 25). The Nationalists' last hope lay in the South and west. But Sian, a long-time Nationalist bastion and the gateway to the northwest, had fallen to Gen. Peng Dehuai on May 20. During the last half of 1949 powerful Communist armies succeeded in taking the provinces of south and west China. By the end of the year only the islands of Hai-nan, Taiwan, and a few offshore positions were still in Nationalist hands, and only scattered pockets of resistance remained on the mainland. The defeated National Government reestablished itself on Taiwan, to which Chiang had withdrawn early in the year, taking most of the government's gold reserves and the Nationalist Air Force and Navy. On October 1, with most of the mainland held by the PLA, Mao proclaimed the establishment in Peking of the government of the People's Republic of China. (C.M.Wi./E.P.Y.)

## The People's Republic of China

The Communist victory in 1949 brought to power a peasant party that had learned its techniques in the countryside but had adopted Marxist ideology and believed in class struggle and rapid industrial development. Extensive experience in running base areas and waging war before 1949 had given the Chinese Communist Party (CCP) deeply ingrained operational habits and proclivities. The long civil war that created the new nation, however, had been one of peasants triumphing over urban dwellers and had involved the destruction of the old ruling classes. In addition, the party leaders recognized that they had no experience in overseeing the transitions to socialism and industrialism that would occur in China's huge urban centres. For this, they turned to the only government with such experience—the Soviet Union. Western hostility against the People's Republic of China, sharpened by the Korean War, contributed to the intensity of the ensuing Sino-Soviet relationship.

### ESTABLISHMENT OF THE PEOPLE'S REPUBLIC

When the CCP proclaimed the People's Republic, most Chinese understood that the new leadership would be preoccupied with industrialization. A priority goal of the Communist political system was to raise China to the status of a great power. While pursuing this goal, the "centre of gravity" of Communist policy shifted from the countryside to the city, but Chairman Mao Zedong (Mao Tse-tung) insisted that the revolutionary vision forged in the rural struggle would continue to guide the party.

In a series of speeches in 1949 Chairman Mao stated that his aim was to create a socialist society and, eventually, world Communism. These objectives, he said, required transforming consumer cities into producer cities to set the basis on which "the people's political power could be consolidated." He advocated forming a four-class coalition of elements of the urban middle class—the petty bourgeoisie and national bourgeoisie—with workers and peasants, under the leadership of the CCP. The people's state would exercise a dictatorship "for the oppression of antagonistic classes" made up of opponents of the regime.

The four-class coalition

The authoritative legal statement of this "people's democratic dictatorship" was given in the 1949 Organic Law for the Chinese People's Political Consultative Conference, and at its first session the conference adopted a Common Program that formally sanctioned the organization of state power under the coalition. Following the Communist victory, a widespread urge to return to normality helped the new leadership restore the economy. Police and party cadres in each locality, backed up by army units, began to crack down on criminal activities associated with economic breakdown. Soon it was possible to speak of longer-term developmental plans.

The cost of restoring order and building up integrated political institutions at all levels throughout the nation proved important in setting China's course for the next two decades. Revolutionary priorities had to be made consonant with other needs. Land reform did proceed in the countryside: landlords were virtually eliminated as a class, land was redistributed, and, after some false starts, China's countryside was placed on the path toward collectivization. In the cities, however, a temporary accommodation was reached with non-Communist elements; many former bureaucrats and capitalists were retained in positions of authority in factories, businesses, schools, and governmental organizations. The leadership recognized that such compromises endangered their aim of perpetuating revolutionary values in an industrializing society, yet out of necessity they accepted the lower priority for Communist revolutionary goals and a higher place for organizational control and enforced public order.

Once in power, Communist cadres could no longer condone what they had once sponsored, and inevitably they adopted a more rigid and bureaucratic attitude toward popular participation in politics. Many Communists, however, considered these changes a betrayal of the revolution; their responses gradually became more intense, and the issue eventually began to divide the once cohesive revolutionary elite. This development is central to an understanding of China's political history since 1949.

Scholars tend to divide China's history into periods after the formation of the People's Republic in 1949. These postwar periods, which will be discussed here briefly, must be used with caution, however, because their definition depends so much on which events are selected for emphasis.

**Reconstruction and consolidation, 1949–52.** During this period, the CCP made great strides toward bringing the country through three critical transitions: from economic prostration to economic growth, from political disintegration to political strength, and from military rule to civilian rule. The determination and capabilities demonstrated during these first years—and the respectable showing (after a century of military humiliations) that Chinese troops made against UN forces on the Korean Peninsula in 1950–53—provided the CCP with a reservoir of popular support that would be a major political resource for years.

Liberation Army troops—called Chinese People's Volunteers—entered the Korean War against United Nations forces in October 1950. Peking had felt threatened by the northward thrust of UN units and had attempted to halt them by its threats to intervene. These threats were ignored by U.S. Gen. Douglas MacArthur, however, and when UN troops under his command reached the Chinese border, Peking acted. By the war's end in July 1953 approximately two-thirds of China's combat divisions had seen service in Korea.

Intervention in the Korean War

In the three years of war, a "Resist America, Aid Korea" campaign translated the atmosphere of external threat into a spirit of sacrifice and enforced patriotic emergency at home. Regulations for the Suppression of Counterrevolutionaries (Feb. 20, 1951) authorized police action against dissident individuals and suspected groups. A campaign against anti-Communist holdouts, bandits, and political opponents was also pressed. Greatest publicity attended Peking's dispatch of troops to Tibet at about the same time that it intervened in Korea. The distinctiveness and world reputation of the Tibetan culture was to make this a severe test of Communist efforts to complete the consolidation of their power. In 1959, after a period of sporadic clashes with the Chinese, the Tibetans rose in rebellion, to which Peking responded with force.

Under the Agrarian Reform Law (June 1950) the property of rural landlords was confiscated and redistributed, which fulfilled a promise to the peasants and smashed a class identified as feudal or semifeudal. The property of traitors, "bureaucrat capitalists" (especially the "four big families" of the Nationalist Party [KMT]—the K'ungs, Soongs, Chiangs, and Ch'ens), and selected foreign nationals was also confiscated, helping end the power of many industrialists and providing an economic basis for industrialization. Programs were begun to increase production and to lay the basis for long-term socialization.

These programs coincided with a massive effort to win over the population to the leadership. Such acts as a marriage law (May 1950) and a trade-union law (June 1950) symbolized the break with the old society, while mass organizations and the regime's "campaign style" dramatized the new.

During 1949–50, policy toward the cities focused on restoring order, rehabilitating the economy, and—above all—wringing disastrous inflation out of the urban economy. To accomplish these tasks, the CCP tried to discipline the labour force, win over the confidence of the capitalists, and implement drastic fiscal policies so as to undercut inflation. These policies brought such remarkable successes that by late 1950 many urban Chinese viewed the CCP leadership as needed reformers. Indeed, numerous capitalists believed them to be "good for business."

Beginnings of the economic and social revolution

But beginning in 1951, the revolutionary agenda of the Communists began to be felt in the cities. A Suppression of Counterrevolutionaries campaign dealt violently with many former leaders of secret societies, religious associations, and the KMT in early 1951. In late 1951 and early 1952, three major political campaigns brought the revolutionary essence of the CCP home to key urban groups. The Three Antis campaign targeted Communist cadres who had become too close to China's capitalists. The Five Antis campaign aimed at the capitalists themselves and brought them into line on charges of bribery, tax evasion, theft of state property and economic information, and cheating on government contracts. And the Thought Reform campaign humbled university professors and marked a turning point in the move from Western to Soviet influence in structuring China's university curriculum.

The pressures toward national political consolidation and the costly struggle in Korea produced significant consequences. In the several provinces of Manchuria (now called the Northeast) there was a growing concentration of industrial and military presence, as well as an increased presence of Soviet economic advisers and key elements of China's tiny corps of technicians and specialists. This was a natural development in view of the extensive economic infrastructure left behind by the Japanese in this region and its proximity to Korea. Additionally, Northeast China had long been an area of Soviet interest.

Gao Gang (Kao Kang) headed Northeast China, and, in addition to his authoritative regional position, Gao also influenced decisions in Peking. He planned the Three-Anti campaign and took the lead in adapting Soviet techniques to Chinese factory management and economic planning. He promoted these techniques on a national basis when he moved to Peking in late 1952 to set up the State Planning Commission. Working closely with the head of the party's Organization Department and other senior officials, Gao subsequently allegedly tried to reduce drastically the authority of his potential competitors, notably Liu Shaoqi (Liu Shao-ch'i) and Zhou Enlai (Chou En-lai), both leading members of party and state organs. The ensuing power struggle lasted more than a year, reflecting an underlying fissure in the CCP. Gao himself had long been a man of the rural base areas, while Liu and Zhou were associated far more with the pre-1949 work in the "white areas" (areas outside CCP control). After 1949, base area veterans believed that they received fewer high positions than their struggles in the wilderness had warranted. Within weeks after the National Conference of the party (March 1955) had proclaimed the defeat of the Gao clique, Peking approved a long-delayed First Five-Year Plan (technically covering the years 1953–57). That summer, active programs for agricultural collectivization and the socialization of industry and commerce were adopted.

Changes in Soviet influence

The period 1949–52 was marked by changes in Soviet influence in China. The officially sanctioned terms of that influence had been worked out in a visit by Mao to Moscow from mid-December 1949 until the following March and were formalized in a Treaty of Friendship, Alliance, and Mutual Assistance (signed Feb. 14, 1950). Years later the Chinese charged that Moscow had failed to give Peking adequate support under that treaty and had left the Chinese to face UN forces virtually alone in Korea. The seeds of doubt concerning Soviet willingness to help China had been sown. Moreover, one of the errors purportedly committed by Gao Gang was his zealousness in using Soviet advisers and promoting the Soviet economic model for management. After the purge of pro-Gao elements, steps were taken to reduce direct Soviet control in China. These steps included reaching agreement on the final withdrawal of Soviet troops from Port Arthur (Lüshun) by mid-1955. Moscow proved amenable to these changes, as Stalin's death had produced new Soviet efforts to end tensions with the Chinese. The applicability of the Soviet model to China and the degree to which its use might become a pretext for Soviet manipulation of China began to be questioned.

Nevertheless, these potential reductions in Soviet influence were counterbalanced by growing Soviet activity in other fields. The Chinese army was reorganized along Soviet lines, with a greater emphasis on heavy firepower and mobility. Soviet texts and propaganda materials flooded the country. The Soviet Union had earlier extended $300,000,000 credit (used up by 1953); this was followed by a smaller developmental loan in 1954 (used up by 1956). Under these aid programs the Soviets supplied the equipment and technical aid for a large number of industrial projects. The Soviet Union also played a major role in Chinese foreign policy, and it appears that China accepted Moscow's leadership in the international Communist movement. Coordinating with Stalin, Peking supported revolutionary activity throughout Asia and opposed compromise with neutralist regimes.

**The transition to socialism, 1953–57.** The period 1953–57, corresponding to the First Five-Year Plan, began China's rapid industrialization. This period is still regarded as having been enormously successful. A strong central governmental apparatus proved able to channel scarce resources into the rapid development of heavy industry. Despite some serious policy issues and problems, the Communist leadership seemed to have the overall situation well in hand. Public order improved and many saw a stronger China taking form. The march to socialism seemed to go along reasonably well with the dictates of industrial development. The determination and fundamental optimism of the Communist leaders appeared justified, especially in view of the decades of invasion, disintegration, self-doubt, and humiliation that had been the lot of the Chinese people before 1949.

Beginning of rapid industrialization

The First Five-Year Plan was explicitly modeled on Soviet experience, and the Soviet Union provided both material aid and extensive technical advice on its planning and execution. During 1952–54 the Chinese established a central planning apparatus and a set of central ministries and other government institutions that were close copies of their Soviet counterparts. These actions were officially ratified by the first meeting of the National People's Congress in September 1954, which formally established the Central People's Government and adopted the first constitution of the People's Republic of China. The plan adopted Stalinist economic priorities. In a country where more than 80 percent of the population lived in rural areas, about 80 percent of all government investment was channeled into the urban economy. The vast majority of this investment went to heavy industry, leaving agriculture relatively starved for resources. The plan provided for substantial income differentials to motivate the labour force in the state sector, and it established a "top down" system in which a highly centralized government apparatus exercised detailed control over economic policy through enormous ministries in Peking. These developments differed substantially from the priorities and proclivities of the Chinese Communist movement in the decades before 1949. Nevertheless, the First Five-Year Plan was linked with the transition of China's rural and urban economy to collective forms.

*Rural collectivization.* This transition was most obvious in the countryside. After land reform had been carried out, mutual aid teams allowed the Communists to experiment with voluntary forms of agricultural collectivization. Starting in late 1953, a campaign was launched to organize into small collectives, called lower level agricultural producers' cooperatives, averaging 20 to 30 households.

Vehement debate soon broke out within the CCP concerning how quickly to move to higher stages of cooperative production in the countryside. This debate was symptomatic of the larger tensions within the party regarding urban and rural development, Soviet influence, and the development of huge government ministries in Peking. The strengths of Mao Zedong lay in agricultural policy, social change, and foreign relations, and in the mid-1950s he began to shift the national agenda more in the direction of his own expertise.

In July 1955 Mao, against the wishes of most of his colleagues in the CCP leadership, called for an acceleration of the transition to lower-level, and then to higher-level, agricultural producers' cooperatives in the countryside. The key difference between these two forms concerned the middle class of peasants, farmers able to live off their own land. The advanced cooperative was particularly disadvantageous to the wealthier peasants because it invested the cooperative itself with title to the land, granting no right of withdrawal, and because wages were based on labour performed, not land contributed. This also made middle-level peasants resent landless peasants, whom the party was recruiting into the new cooperatives. Also, the advanced form, modeled on the Soviet *kolkhoz,* brought with it the outside political controls that were necessary to extract the agricultural surpluses required to pay for China's capital equipment in its industrialization and to feed the workers moving into the cities to man the growing industries. Many middle-level peasants actively resisted these changes and the measures for enforcing them, particularly grain rationing, compulsory purchase quotas, and stricter regulations on savings and wage rates. Nevertheless, Chinese agricultural organization in 1956 reached the approximate level of collectivization achieved in the Soviet Union—a peasant owned his house, some domestic animals, a garden plot, and his personal savings; by the end of 1956, 88 percent of China's peasant households were organized into advanced cooperatives.

Higher level cooperatives

*Urban socialist changes.* Mao combined this massive transformation of the agricultural sector with a call for the "socialist transformation" of industry and commerce, in which the government would become, in effect, the major partner. In Chinese Communist fashion, this change was not simply decreed from above. Rather, extreme pressures were put on private merchants and capitalists in late 1955 to "volunteer" their enterprises for transformation into "joint state–private" firms. The results were sometimes extraordinary. For example, all the capitalists in a given trade (such as textiles) would parade together to CCP headquarters to the beat of gongs and the sound of firecrackers. Once there, they would present a petition to the government, asking that the major interest in their firms be bought out at the rate that the government deemed appropriate. The government would graciously agree.

Such actions can be understood against the background of the experiences of the capitalists in the previous few years. The Five-Anti campaign of 1952 had terrorized many of them and had left most of them deeply in debt to the government, owing purported back taxes and financial penalties. In any case, the state sector of the economy and state controls over banking had increased to such a degree that the capitalists relied heavily on the government for the contracts and business necessary to keep from bankruptcy. After the Five-Anti campaign, the government extended the reach of its trade unions into the larger capitalist enterprises, and the "joint labour–management" committees set up under government pressure in those firms usurped much of the power that the capitalists formerly had exercised. Thus, many Chinese capitalists saw the socialist transformation of 1955–56 as an almost welcome development since it secured their position with the government while costing them little in money or power.

*Political developments.* The socialist transformation of agriculture, industry, and commerce thus went relatively smoothly. Nevertheless, changes of this sort could not take place without considerable tensions. Many peasants streamed into the cities in 1956–57 to escape the new cooperatives and to seek employment in the rapidly expanding state-run factories, where government policy kept wages rising rapidly. China's urban population mushroomed from 77,000,000 in 1953 to 99,500,000 by 1957.

Three major economic problems

Several problems also became increasingly pressing. First, CCP leaders found that the agricultural sector was not growing fast enough to provide additional capital for its own development and to feed the workers of the cities. Until then, agricultural policy had attempted to wring large production increases out of changes in organization and land ownership, with little capital investment. By 1956–57 this policy was shown to be inadequate.

Second, Soviet assistance had been made available to China as loans, not grants. After 1956 China had to repay more each year than it borrowed in new funds. Thus, the Chinese could no longer count on Moscow for net capital accumulation in its industrialization drive.

Third, the vastly expanded governmental responsibility for managing the country's urban firms and commerce required far more experts than before. For this, the leadership tried to resolve the increasingly severe strains that had characterized the relationship between the country's intellectuals (including technical specialists) and the CCP.

The leadership's policies in the past had reflected much ambivalence toward the intelligentsia—on the one hand it had required their services and prestige, but on the other it had suspected that many were untrustworthy, coming from urban and bourgeois backgrounds and often having close family and other personal ties with the KMT. After 1949 and particularly during the first part of the Korean War, the Central Committee launched a major campaign to reeducate teachers and scientists and to discredit Western-oriented scholarship. In 1951 the emphasis shifted from general campaigns to self-reform; in 1955 it shifted once again to an intensive thought-reform movement, following the purge of Hu Feng, until then the party's leading spokesman on art and literature. This latter movement coincided with the denunciation of a scholarly study of the *Dream of the Red Chamber,* an 18th-century novel of tragic love and declining fortunes in a Chinese family. Literature without a clear class moral received blistering criticism, as did any hint that the party should not command art and literature—a theme identified with the ousted Hu Feng—and "Hu Feng elements" were exposed among intellectuals in schools, factories, and cooperatives.

Hundred Flowers Campaign

The intensity of these attacks slackened in early 1956. Party leaders publicly discussed the role of intellectuals in the new tasks of national construction and adopted the line of "Let a hundred flowers blossom, a hundred schools of thought contend." Because intellectuals in China included high school graduates as well as those with college or advanced professional training, the policy affected a vast number of people. The "hundred flowers" line explicitly encouraged "free-ranging" discussion and inquiry, with the explicit assumption that this would prove the superiority of Marxism-Leninism and speed the conversion of intellectuals to Communism. Their response was gradual and cautious to the party's invitation for free discussion and criticism. Instead of embracing Marxism, moreover, many used the opportunity to translate and discuss Western works and ideas and blithely debated "reactionary" doctrines at the very moment Hungarian intellectuals were triggering a wave of anti-Communist sentiment in Budapest.

Following this initial phase of the Hundred Flowers Campaign, Mao Zedong issued what was perhaps his most famous post-1949 speech "On the Correct Handling of Contradictions Among the People" (Feb. 27, 1957). Its essential message was ambiguous. He stressed the importance of resolving "nonantagonistic contradictions" by methods of persuasion, but he stated that "democratic" methods of resolution would have to be consistent with centralism and discipline. He left it unclear when a contradiction might become an "antagonistic" and no-holds-barred struggle. The final authoritative version of this speech contained explicit limits on the conduct of debate that had been absent in the original. According to this version, the party would judge words and actions to be correct only if they united the populace, were beneficial to socialism, strengthened the state dictatorship, consolidated organizations, especially the party, and generally helped

strengthen international Communism. These textual manipulations, moreover, led to an unresolved controversy concerning the initial intent of Mao's speech.

The leadership's explanation was that Mao had set out to trap the dangerous elements among the intellectuals by encouraging their criticism of the party and government. An alternative view was that the leaders used the metaphor of the trap to rationalize their reaction to the unanticipated criticism, popular demonstrations, and general antiparty sentiments expressed in the late spring, when the term "hundred flowers" gained international currency. Whatever the correct explanation for these significant textual changes, the Communist leaders had encouraged free criticism of the party and its programs, and they had then turned on their critics as rightists and counterrevolutionaries. In June non-Communists who had thrown caution to the winds reaped the full fury of retaliation in an antirightist campaign. The intellectuals who had responded to Mao's call for open criticism were the first victims, but the movement quickly spread beyond that group to engulf many specialists in the government bureaucracy and state-run firms. By the fall the fury of the campaign began to turn toward the countryside, and those, especially among the rural cadres, who had remained unenthusiastic about the "high tide" of agricultural change came under fire and were removed. The spreading antirightist campaign then inspired fear in those who wanted a slower, more pragmatic approach to development and shifted the initiative to others who, like Mao, believed that the solutions to China's core problems lay in a major break with the incrementalist Soviet strategy and in a bold new set of distinctly Chinese ideas. International events dovetailed with this basic thrust by the winter of 1957–58.

*Foreign policy.* While taking their principal cues in their foreign policies from domestic developments and generally adhering to the initial pro-Soviet line, the Chinese began to act—on the basis of several important lessons gained during the Korean struggle—to reduce Peking's militant and isolationist attitudes in international affairs. Peking had recognized that the great costs of the war, the questionable reliability of Soviet military backing, and the danger of direct U.S. retaliation against China had come close to threatening its very existence. Although in preserving North Korea as a Communist state China had attained its principal strategic objective, its leaders understood the costs and risks involved and were determined to exercise a greater caution in their international dealings. Another lesson was that the neutralist states in Asia and Africa were not Western puppets, and it was politically profitable to promote friendly relations with them. These lessons, as reinforced by domestic considerations, led China to take a conciliatory role in the Geneva Conference on Indochina in 1954 and to try to normalize its foreign relations.

Lessons from the Korean War

Premier Zhou Enlai symbolized China's more active diplomatic role at the Asian-African Conference held at Bandung, Indon., in April 1955. His slogan was "unity with all," according to the line of peaceful coexistence. This "Bandung line" associated with Zhou gained worldwide attention when he told the Bandung delegates that his government was fully prepared to achieve normal relations with all countries, including the United States. As a result of his initiative, ambassadorial talks between China and the United States were begun.

Between 1955 and 1957, however, changes in Soviet and U.S. policies caused Chinese leaders to doubt the validity of this more cautious and conciliatory foreign policy. At the 20th Congress of the Soviet Communist Party in 1956, Party Secretary Nikita Khrushchev announced a de-Stalinization policy. This development angered Mao Zedong for two reasons: he thought, correctly, that it would undermine Soviet prestige, with potentially dangerous consequences in eastern Europe, and he chafed at Khrushchev's warning to other Communist parties not to let a willful leader have his way unchecked. Thus, a new situation in Sino-Soviet relations began to emerge, in which antagonisms based on different national traditions, revolutionary experiences, and levels of development that had previously been glossed over broke through to the surface.

Chinese leaders—Mao foremost among them but by no means alone—now began to question the wisdom of closely following the Soviet model. Economic difficulties provided a major set of reasons for moving away from that model, and increasing mutual distrust exacerbated the situation. Nevertheless, at the end of 1957 the Soviet Union evidently agreed to provide China with the technical assistance needed to make an atomic bomb, and during 1958 the Soviet Union increased its level of aid to China. In the final analysis, however, the spiral of deterioration in Sino-Soviet relations proved impossible to reverse.

China adopted a new, more militant foreign policy that can be traced most clearly to Mao's statement during a Moscow trip in November 1957 that the "East wind prevails over the West wind." This implied a return to militant struggle. According to some estimates, the change in line was necessitated by the U.S. buildup of anti-Communist regimes to encircle China and by the lack of major gains in peaceful coexistence with Third World neutrals. Other analysts argue that Mao regarded the launching of a Soviet space vehicle (October 1957) and the Sino-Soviet nuclear-sharing agreement as indications that the balance of world forces had changed in favour of Communism.

Return to a militant struggle

**New directions in national policy, 1958–61.** The pressures behind the dramatic inauguration in 1958 of "Three Red Banners"—the general line of socialist construction, the Great Leap Forward, and the rural people's communes—are still not fully known. Undoubtedly, a complex mixture of forces came into play. Mao personally felt increasingly uncomfortable with the alliance with the Soviet Union and with the social and political ramifications of the Soviet model of development. On ideological grounds and because it shifted policy away from his personal political strengths, Mao disliked the Soviet system of centralized control by large government ministries, substantial social stratification, and strong urban bias. In addition, the Soviet model assumed that agricultural surplus need only be captured by the government and made to serve urban development. This was true for the Soviet Union in the late 1920s, when the model was developed, but the situation in China was different. Chinese policy had to devise a way first to create an agricultural surplus and then to take a large part of it to serve urban growth. The Soviet model also rested on implicit assumptions about the energy and transportation sectors that were not compatible with the Chinese realities of the 1950s.

To some extent, obscure political battles also became caught up in the debates over Chinese development strategies. In the spring of 1958, for example, Mao Zedong elevated Marshal Lin Biao (Lin Piao) to a higher position in the CCP than that held by Defense Minister Peng Dehuai (P'eng Te-huai). At the same time, Mao initiated a critique of China's slavish copying of Soviet military strategy. Peng bore the brunt of this criticism because he had advocated close military ties with the Soviet Union.

Overall, the radicalization of policy that led to the Great Leap Forward can be traced back to the antirightist campaign of 1957 and a major meeting of China's leaders at the resort city of Tsingtao in October of that year. By the time of another central meeting—this one in Nan-ning in January 1958—Mao felt confident enough to launch a blistering critique of the domination of economic policy by the State Council and its subordinate ministries. The best available evidence suggests that almost all of the top leaders supported Mao as he developed a series of initiatives that eventually produced the Great Leap strategy and the people's communes. The only major exceptions appear to have been Zhou Enlai and Chen Yun (Ch'en Yün), a force in Chinese economic policy; both faded from the public eye in 1958 only to be brought back into active roles as the Great Leap faltered in 1959.

The general line of socialist construction and the Great Leap Forward were announced at the second session of the Eighth Party Congress (May 1958), which concentrated as much on political slogans as on specific objectives. Special emphasis was placed on political guidance by party cadres of the country's scientists and technicians, who were viewed as potentially dangerous unless they would become fully "Red and expert." The progressive indoctrination

Great Leap Forward

of experts would be paralleled by introductory technical training for cadres, thereby in theory transforming the entire elite into political-technical generalists. The Congress of 1958 called for a bold form of ideological leadership that could unleash a "leap forward" in technical innovation and economic output. To link the new generalist leaders and the masses, emphasis fell on sending cadres to the lower levels (*hsia-fang*) for firsthand experience and manual labour and for practical political indoctrination.

The Great Leap Forward involved an enormous amount of experimentation. It had no detailed blueprint, but there were some underlying strategic principles. There was a general reliance on a combination of ideological and organizational techniques to overcome seemingly insuperable obstacles that was focused on the countryside and that drew from policies of the 1930s and '40s. The basic idea was to convert the massive labour surplus in China's hinterlands into a huge production force through a radical reorganization of rural production. The search for the best organizational form to achieve this result led in August 1958 to popularization of the "people's commune," a huge rural unit that pooled the labour of tens of thousands of peasants from different villages in order to increase agricultural production, engage in local industrial production, enhance the availability of rural schooling, and organize a local militia force in accordance with Mao's preferred national military strategy of combining the deterrence of an atomic bomb with guerrilla warfare.

Mao believed that through these radical organizational changes, combined with adequate political mobilization techniques, the Chinese countryside could be made to provide the resources both for its own development and for the continuing rapid development of the heavy industrial sector in the cities. Through this strategy of "walking on two legs," China could obtain the simultaneous development of industry and agriculture and, within the urban sector, of both large- and small-scale industry. If it worked, this would resolve the dilemma of an agricultural bottleneck that had seemed to loom large on the horizon as of 1957. It would, however, involve a major departure from the Soviet model, which would predictably lead to increased tensions between Peking and Moscow.

Largely because of unusually good weather, 1958 was an exceptionally good year for agricultural output. But, by the end of that year, the top CCP leadership sensed that some major problems demanded immediate attention. Initial optimism had led peasants in many areas to eat far more than they usually would have, and stocks of grain for the winter and spring months threatened to fall dangerously low. In addition, reports of sporadic peasant unrest cast some doubt on the rosy picture being presented to the leaders by their own statistical system, the accuracy of which, in turn, came into question.

The fall harvest of 1958 had not been as large as expected, and in February and March 1959 Mao Zedong began to call for appropriate adjustments to make policies more realistic without abandoning the Great Leap as a whole. Mao emerged as one of the most forceful advocates of scaling back the Great Leap in order to avert a potential disaster. He faced substantial resistance from provincial CCP leaders whose powers had been greatly increased as part of the Great Leap strategy. A meeting at Lu-shan in the summer of 1959 produced an unanticipated and ultimately highly destructive outcome. Defense Minister Peng Dehuai raised a range of criticisms of the Great Leap, based in large part on his own investigations. He summed these up in a letter that he sent to Mao during the conference. Mao waited eight days to respond to the letter and then attacked Peng for "right deviationism" and demanded the purge of Peng and all his followers.

The Lu-shan Conference resulted in several major decisions: Peng Dehuai was replaced with Lin Biao, who would later be marked for succession to Mao's position of CCP chairman; the Great Leap Forward was scaled back; and a political campaign was launched to identify and remove all "rightist" elements. The third decision effectively canceled the second, as party officials refused to scale back the Great Leap for fear of being labeled as "rightists." The net effect was to produce a "second leap"—a new radical upsurge in policy that was not corrected until it produced results so disastrous that they called into question the very viability of the Communist system.

The CCP celebrated the 10th anniversary of national victory in October 1959 in a state of near euphoria. The weather turned in 1959, however, and during the next two years China experienced a severe combination of floods and drought. Although the economy was in serious trouble by mid-1960, the Chinese leaders sharpened their debate with Moscow. In April 1960, on the occasion of Lenin's 90th birthday, for example, Peking published an article that contained a slightly veiled critique of Soviet foreign policy, arguing that the Soviets had become soft on imperialism. Khrushchev reacted with a rapid withdrawal of all Soviet technicians and assistance that July. (When he quietly offered to return them that November, his offer was refused.)

Failure of the Great Leap

Despite the importance of these difficulties, China's worst problem was bad policy. The people's communes were too large to be effective, they ignored age-old marketing patterns in the countryside, and they required administrative and transport resources that did not exist. Their structure and means of allocating resources removed almost all incentive to work, and the breakdown in the statistical system meant that the top leaders had grossly erroneous ideas about what was occurring. Thus, even after massive starvation had beset many rural areas, the orders from above continued to demand large-scale procurement of foodstuffs. The rural cadres were so afraid of being branded rightist that they followed these unrealistic orders, thus deepening the famine. By 1961 the rural disaster caught up with the cities, and urban industrial output plummeted by more than 25 percent. As an emergency measure, nearly 30,000,000 urban residents were sent back to the countryside because they could no longer be fed in the cities. The Great Leap Forward had run its course and the system was in crisis.

**Readjustment and reaction, 1961–65.** The years 1961–65 did not resemble the three previous ones, despite the persistence of radical labels and slogans. The Chinese themselves were loath to acknowledge the end of the Great Leap period, declaring the validity of the general line of socialist construction and its international revolutionary corollary for one and all.

Reality can be seen, however, in the increasing role of the Chinese military and security personnel. At a top-level meeting of the Military Affairs Committee in October 1960 and at one of the rare plenary sessions of the party's Central Committee the following January, the elite gave the highest priority to the restoration of security and national order. Party recruitment procedures were tightened, and a major thought-reform movement was launched within the cadres' ranks. The Central Committee also established six supraprovincial regional bureaus charged with enforcing obedience to Peking and bringing the new procedures for control into line with local conditions. The army, now firmly under Lin Biao, took the lead, beginning with a "purification" movement against dissidents within its own ranks. Throughout 1961 and most of 1962 the central officials worked to consolidate their power and to restore faith in their leadership and goals.

Rise of Liu Shaoqi and Deng Xiaoping

By January 1962 Mao had, as he later put it, moved to the "second line" to concentrate "on dealing with questions of the direction, policy, and line of the party and the state." The "first-line" administrative and day-by-day direction of the state had been given to Liu Shaoqi, who had assumed the chairmanship of the People's Republic of China in 1959 (though Mao retained his position of party chairman); additional responsibilities in the first line were given to Deng Xiaoping (Teng Hsiao-p'ing), another tough-minded organizer who, as general secretary, was the party's top administrator. By 1962 Mao had apparently begun to conclude that the techniques used by these comrades in the first line not only violated the basic thrust of the revolutionary tradition but also formed a pattern of error that mirrored what he viewed as the "modern revisionism" of the Soviet Union.

Under Liu and Deng, the CCP during 1960–61 developed a series of documents in major policy areas to try to

bring the country out of the rapidly growing crisis. In most instances, the drafting of these documents was done with the assistance of experts who had been reviled during the Great Leap Forward. These documents marked a major retreat from Great Leap radicalism. The communes were to be reduced on the average by about two-thirds so as to make them small enough to link peasants' efforts more clearly with their remuneration. Indeed, by 1962 in many areas of rural China the collective system in agriculture had broken down completely and individual farming was revived. Policy toward literature, art, and motion pictures permitted a "thaw" involving treatment of a far broader range of subjects and revival of many older, prerevolutionary artistic forms. The new program in industry strengthened the hands of managers and made a worker's efforts more closely attuned to his rewards. Similar policies were adopted in other areas. In general, over the years 1961–65 a remarkable job of reviving the economy succeeded in at least regaining the level of output of 1957 in almost all sectors.

These policies raised basic questions about the future direction of the revolution. While almost all top CCP leaders had supported the launching of the Great Leap, there was disagreement over the lessons to be learned from the movement's dramatic failure. The Great Leap had been intended both as a means of accelerating economic development and as a vehicle for achieving a mass ideological transformation. All leaders agreed in its aftermath that a mobilization approach to economic development was no longer appropriate to China's conditions. Most also concluded that the age of mass political campaigns as an instrument to remold the thinking of the public was past. Mao and a few of his supporters, however, still viewed class struggle and mass mobilization as core ingredients in keeping the revolutionary vision alive.

Mao personally lost considerable prestige over the failure of the Great Leap—and the party's political and organizational apparatus was damaged—but he remained the most powerful individual in China. He proved able time and again to enforce his will on the issues that he deemed to be of top priority. Claims made during the Cultural Revolution that Mao had been pushed aside and ignored during 1962–65 are not supported by the evidence.

Mao was in fact deeply troubled as he contemplated China's situation during 1961–65. He perceived the Soviet Socialist revolution in the years after Stalin's death in 1953 to have degenerated into "social imperialism." Mao evidently had been shocked by these developments in the Soviet Union, and the revelation made him look at events in China from a new vantage point. Mao became convinced that China, too, was headed down the road toward revisionism. He used class struggle and ideological campaigns, as well as concrete policies in various areas, to try to prevent and reverse this slide into revolutionary purgatory. Mao's nightmare about revisionism played an increasing role in structuring politics in the mid-1960s.

Fears of revisionism

Mao was not the only leader who harboured doubts about the trends in the recovery effort of 1961–65. Others gathered around him and tried to use their closeness to Mao as a vehicle for enhancing their political power. The key individuals involved were Mao's political assistant of many years, Chen Boda (Ch'en Po-ta), who was an expert in the realm of ideology; Mao's wife, Jiang Qing (Chiang Ch'ing), who had strong policy views in the cultural sphere; Kang Sheng (K'ang Sheng), whose strength lay both in his understanding of Soviet ideology and in his mastery of Soviet-style secret police techniques; and Lin Biao, who headed the military and tried to make it an ideal type of Maoist organization that combined effectiveness with ideological purity. Each of these people, in turn, had personal networks and resources to bring to a coalition. While their goals and interests did not entirely coincide, they all could unite on two efforts—enhancing Mao's power and upsetting Mao's relations with Liu Shaoqi (then the likely successor to Mao), Deng Xiaoping, and most of the remainder of the party leadership.

Mao took a number of initiatives in domestic and foreign policy during the mid-1960s. At a major Central Committee plenum in September 1962 he insisted that "class struggle" remain high on the Chinese agenda, even as enormous efforts continued to be made to revive the economy. He also called for a campaign of "socialist education," aimed primarily at reviving the demoralized party apparatus in the countryside. By 1964 he began to press hard to make the Chinese educational system less elitist by organizing "part-work, part-study" schools that would provide more vocational training. Throughout this period, foreign observers noted what appeared to be some tension between a continuing thread of radicalism in China's propaganda and a strong pragmatic streak in the country's actual domestic policies.

The most important set of measures Mao took concerned the People's Liberation Army (PLA), which he and Lin Biao tried to make into a model organization. Events on the Sino-Indian border in the fall of 1962 helped the PLA reestablish discipline and its image. From 1959 to 1962 both India and China, initially as a by-product of the uprising in Tibet, resorted to military force along their disputed border. On Oct. 12, 1962, a week before the Chinese moved troops into disputed border territories, Indian Prime Minister Jawaharlal Nehru stated that the army was to free all Indian territory of "Chinese intruders." In the conflict that followed, Peking's regiments defeated the Indian Army in the border region, penetrating well beyond it. The Chinese then withdrew from most of the invaded area and established a demilitarized zone on either side of the line of control. Most significantly, the leadership seized on the army's victory and began to experiment with the possibility of using army heroes as the ideal types for popular emulation.

Dispute with India

Increasingly preoccupied with indoctrinating its heirs and harking back to revolutionary days, Peking's leaders closest in outlook to Mao Zedong and Lin Biao viewed the soldier-Communist as the most suitable candidate for the second- and third-generation leadership. Army uniformity and discipline, it was seen, could transcend the divided classes, and all army men could be made to comply with the rigorous political standards set by Mao's leadership.

Lin Biao developed a simplified and dogmatized version of Mao's thought, eventually published in the form of the "little red book," *Quotations from Chairman Mao,* to popularize Maoist ideology among the relatively uneducated military recruits. As the military forces under Lin increasingly showed that they could combine ideological purity with technical virtuosity, Mao tried to expand the PLA's organizational authority and its political role. Beginning in 1963, Mao called on all Chinese to "learn from the PLA." Then, starting in 1964, Mao insisted that political departments modeled on those in the PLA be established in all major government bureaucracies. In many cases, political workers from the PLA itself staffed these new bodies, thus effectively penetrating the civilian government apparatus. Other efforts, such as a national propaganda campaign to learn from a purported army hero, Lei Feng, also contributed to enhancement of the PLA's prestige.

The militancy of subsequent campaigns to learn from army heroes, or from the PLA as a whole, was echoed in international politics. In a tour of Africa in late 1963 and early 1964, Zhou Enlai startled his hosts by calling for revolution in newly independent states and openly challenged the Soviet Union for the leadership of the Third World. Simultaneously, China challenged the U.S. system of alliances by establishing formal relations with France and challenged the Soviet Union's system by forming closer ties with Albania.

Peking's main target was Moscow. A Soviet-U.S. crisis in Cuba (October 1962) had coincided with the Sino-Indian struggle, and in both cases the Chinese believed the Soviet Union had acted unreliably and had become "capitulators" of the worst sort. For the next months polemicists in Peking and Moscow publicly engaged in barbed exchanges. When the Soviet Union signed a partial Nuclear Test-Ban Treaty with the United States and Great Britain in July 1963, Chinese articles accused the Soviets of joining an anti-Chinese conspiracy. Confronted by this new strategic situation, the Chinese shifted their priorities to support an antiforeign line and to promote the country's "self-reliance." Mao's calls for "revolutionization" acquired a

The end of Sino-Soviet unity

more nationalistic aspect, and the PLA assumed an even larger place in Chinese political life.

These many-sided trends seemed to collide in 1963 and 1964. With the split in the international Communist movement, the party in late 1963 called on intellectuals, including those in the cultural sphere, to undertake a major reformulation of their academic disciplines to support China's new international role. The initial assignment for this reformulation fell to Zhou Yang (Chou Yang), a party intellectual and deputy director of the Central Committee's Propaganda Department, who tried to enlist China's intellectuals in the ideological war against Soviet revisionism and in the struggle for rigidly pure political standards. (Less than three years later, however, Zhou Yang was purged as a revisionist, and many intellectuals were condemned as Mao Zedong's opponents.)

Closely connected with the concerns of the intellectuals were those relating to the party and the Communist Youth League. A drive began to cultivate what one author called "newborn forces," and by mid-1964 young urban intellectuals were embroiled in a major effort by the Central Committee to promote those forces within the party and league; meanwhile their rural cousins were buffeted by moves to keep the socialist education campaign under the party's organizational control through the use of "work teams" and a cadre-rectification movement.

Mao in the summer of 1964 wrote a document entitled "On Khrushchev's Phony Communism and Its Historical Lessons for the World," which summarized most of Mao's doctrinal principles on contradiction, class struggle, and political structure and operation. This summary provided the basis for the reeducation ("revolutionization") of all youth hoping to succeed to the revolutionary cause. This high tide of revolutionization lasted until early August, when U.S. air strikes on North Vietnam raised the spectre of war on China's southern border. A yearlong debate followed on the wisdom of conducting disruptive political campaigns during periods of external threat.

This period has come to be interpreted as a time of major decision within China. One ingredient of the debate was whether to prepare rapidly for conventional war against the United States or to continue the revolutionization of Chinese society, which in Mao's view had fundamental, long-term importance for China's security. Those who argued for a postponement of the internal political struggle supported more conventional strategies for economic development and took seriously Soviet calls for "united action" in Vietnam and the establishment of closer Sino-Soviet ties. Their position, it was later alleged, received the backing of the general staff. With the dispatch of about 50,000 logistic personnel to Vietnam after February 1965, factional lines began to divide the military forces according to ideological or national-security preferences.

Effects of the war in Indochina

Meanwhile, some members tried to restore rigid domestic controls. Where Mao in May 1963 had called for an upsurge in revolutionary struggle, other leaders in the following September circumscribed the area of cadre initiative and permitted a free-market system and private ownership of rural plots to flourish. A stifling of the revolutionary upsurge was supposedly evident in regulations of June 1964 for the organization of poor and lower-middle-peasant associations, and by early 1965 Mao could point to bureaucratic tendencies throughout the rural areas. In a famous document on problems arising in the course of the socialist education campaign, usually referred to as the "Twenty-three Articles," Mao in January 1965 stated, for the first time, that the principal enemy was to be found within the party and once more proclaimed the urgency of class struggle and mass-line politics.

It was in this period of emphasis on self-reliant struggle that China acquired nuclear weapons. Although the Soviet Union supported Chinese nuclear aims for a time, this effort was taken over completely by the Chinese after June 1959. By 1964 the costs of the program had forced a substantial reduction in other defense costs. China's first atomic explosion (Oct. 16, 1964) affected the debate by appearing to support Mao's contention that domestic revolutionization would in no way jeopardize long-term power aspirations and defense capabilities.

China's first atomic explosion

Mao's military thinking, a product of his own civil war experiences and an essential component of his ideology, stressed the importance of people's war during the transition to nuclear status. He felt that preparation for such a war could turn China's weaknesses into military assets and reduce its vulnerability. Mao's view of people's war belittled the might of modern advanced weapons as "paper tigers" but recognized that China's strategic inferiority subjected it to dangers largely beyond its control. His reasoning thus made a virtue out of necessity in the short run, when China would have to depend on its superior numbers and the morale of its people to defeat any invader. In the long run, however, he held that China would have to have nuclear weapons to deprive the superpowers of their blackmail potential and to deter their aggression against smaller states.

Lin Biao repeated Mao's position on people's war, further arguing that popular insurrections against non-Communist governments could succeed only if they took place without substantial foreign assistance. To the extent that indigenous rebels came to depend on outside support, inevitably their bonds with the local populace would be weakened. When this happened, the rebellion would wither for lack of support. On the other hand, the hardships imposed by relying on indigenous resources would stimulate the comradeship and the ingenuity of the insurgents. Equally important, Lin's statement also indicated a high-level decision for China to remain on the defensive.

Lin's speech coincided with yet another secret working conference of the Central Committee, in which the Maoist group reissued its call for cultural revolutionization, this time convinced that the effort of 1964 had been deliberately sabotaged by senior party and military officials. Initiated by Mao Zedong and Lin Biao, the purge first struck dissident army leaders, especially the chief of staff; and as the power struggle began, China turned its back on the war in Vietnam and other external affairs. The September meeting may be taken as a clear harbinger of the Great Proletarian Cultural Revolution.

### THE CULTURAL REVOLUTION, 1966–76

As the clash over issues in the autumn of 1965 became polarized, the army initially provided the battleground. The issues concerned differences over policy directions and their implications for the organization of power and the qualifications of senior officials to lead. Much of the struggle went on behind the scenes; in public it took the form of personal vilification and ritualized exposés of divergent worldviews or, inevitably, "two lines" of policy. Lin Biao, in calling for the creative study and application of Mao's thought in November and at a meeting of military commissars the following January, consistently placed the army's mission in the context of the national ideological and power struggle. In these critical months the base of operations for Mao and Lin was the large eastern Chinese city of Shanghai; and newspapers published in that city, especially the *Liberation Army Daily,* carried the public attacks on the targets selected.

**Attacks on cultural figures.** The first target was the historian Wu Han, who doubled as the deputy mayor of Peking. In a play, Wu supposedly had used allegorical devices to lampoon Mao and laud the deposed former minister of defense, Peng Dehuai. The denunciation of Wu and his play on Nov. 10, 1965, constituted the opening volley in an assault on cultural figures and their thoughts.

As the Cultural Revolution gained momentum, Mao turned for support to the youth as well as the army. In seeking to create a new system of education that would eliminate differences between town and country, workers and peasants, and mental and manual labour, Mao struck a responsive chord within the youth; it was their response that later provided him with his best shock troops. As a principal purpose, the Cultural Revolution was launched to revitalize revolutionary values for the successor generation of Chinese young people.

Mao on education

During the spring of 1966 the attack against authors, scholars, and propagandists emphasized the cultural dimension of the Cultural Revolution. Increasingly it was hinted that behind the visible targets lay a sinister "black

gang" in the fields of education and propaganda and high up in party circles. Removal of Peng Zhen (P'eng Chen) and Lu Dingyi (Lu Ting-yi) and subsequently of Zhou Yang, then tsar of the arts and literature, indicated that this was to be a thoroughgoing purge. Clearly, a second purpose of the Cultural Revolution would be the elimination of leading cadres whom Mao held responsible for past ideological sins and alleged errors in judgment.

**Attacks on party members.** Gradual transference of the revolution to top echelons of the party was managed by a group centred on Mao Zedong, Lin Biao, Jiang Qing, Kang Sheng, and Chen Boda. In May 1966 Mao secretly assigned major responsibilities to the army in cultural and educational affairs. Another purpose of the Cultural Revolution, as then conceived, would be a "revolution in the superstructure": a transformation from a bureaucratically run machine to a more popularly based system led personally by Mao and a simplified administration under his control.

Following the May instructions, the educational system received priority. "Big-character posters," or large wall newspapers (*ta-tzu-pao*), spread from the principal campuses in Peking throughout the land. University officials and professors were singled out for criticism, while their students, encouraged by the central authorities, held mass meetings and began to organize. In June the government dropped examinations for university admissions and called for a reform of entrance procedures and a delay in reopening the campuses. Party officials and their wives circulated among the campuses to gain favour and to obstruct their opponents. Intrigue and political maneuvering dominated, although political lines were not at first sharply drawn or even well understood. The centres of this activity were Peking's schools and the inner councils of the Central Committee; the students were the activists in a game they did not fully comprehend.

This phase of the Cultural Revolution ended in August 1966 with the convening of a plenary session of the Central Committee. Mao issued his own big-character poster to "Bombard the Headquarters," a call for the denunciation and removal of senior officials, and a 16-point Central Committee decision was issued, in which the broad outlines for the Cultural Revolution were laid down and supporters were rallied to the revolutionary banner. The immediate aim was to seize power from "bourgeois" authorities. The locus of the struggle would be their urban strongholds. Now more than ever, Mao's thought became the "compass for action." (J.W.Le./K.G.L.)

Evidently fearing that China would develop along the lines of the Soviet revolution, and concerned about his own place in history, Mao threw China's cities into turmoil in a gigantic effort to reverse the historic processes then under way. He ultimately failed in his quest, but his efforts generated problems with which his successors would have to struggle for decades. Mao adopted four goals for his Cultural Revolution: to replace his designated successors with leaders more faithful to his current thinking, to rectify the CCP, to provide China's youth with a revolutionary experience, and to achieve specific policy changes to make the educational, health-care, and cultural systems less elitist. He initially pursued those goals through a massive mobilization of the country's urban youths—organized in groups called the Red Guards—while ordering the CCP and the PLA not to suppress the movement.

The Red Guards

When Mao formally launched the Cultural Revolution in August 1966, he had already shut down the schools. During the following months he encouraged the Red Guards to attack all traditional values and "bourgeois" things and to put CCP officials to the test by publicly criticizing them. These attacks were known at the time as struggles against the Four Olds (*i.e.*, old ideas, customs, culture, and habits of mind), and the movement quickly escalated to the committing of outrages. Many elderly people and intellectuals were physically abused, and many died. Nonetheless, Mao believed that this mobilization of urban youths would be beneficial for them and that the CCP cadres they attacked would be better for the experience.

**Seizure of power.** The period from mid-1966 to early 1969 constituted the Red Guard phase of the Cultural Revolution, and these years in turn included several important turning points. The latter half of 1966 witnessed not only the Red Guard mobilization (including Red Guard reviews of more than 1,000,000 youths at a time by Mao Zedong and Lin Biao in Peking) but also the removal from power of key Politburo leaders, most notably Pres. Liu Shaoqi and CCP General Secretary Deng Xiaoping. In October 1966 both Liu and Deng engaged in public self-criticism. Mao, however, rejected both acts as inadequate. At the same meeting, Mao heard bitter complaints from provincial party leaders about the chaos of the political campaign. While acknowledging the validity of much of what was said, Mao nevertheless declared that it would do more good than harm to let the Cultural Revolution continue for several more months.

In January 1967 the movement began to produce the actual overthrow of provincial CCP committees and initial attempts to construct new organs of political power to replace them. The first such "power seizure" took place in Shanghai and was followed by temporary confusion as to just what kind of new political structure should be established to replace the discredited municipal CCP and government apparatuses. The final form adopted was called a "revolutionary committee," and that appellation was given to Chinese government committees until the late 1970s.

The chaos involved in the overthrow of the Shanghai authorities combined with political outrages throughout the country to lead many remaining top CCP leaders to call in February 1967 for a halt to the Cultural Revolution. During this attempt to beat back radicalism, more conservative forces clamped down on Red Guard activism in numerous cities. The movement, dubbed the "February adverse current," was quickly defeated and a new radical upsurge began. Indeed, by the summer of 1967 large armed clashes occurred throughout urban China, and even Chinese embassies abroad experienced takeovers by their own Red Guards. The Red Guards splintered into zealous factions, each purporting to be the "true" representative of the thought of Mao Zedong. Mao's own personality cult, encouraged so as to provide momentum to the movement, assumed religious proportions. The resulting anarchy, terror, and paralysis threw the urban economy into a tailspin. Industrial production for 1968 dipped 12 percent below that of 1966.

Intervention of the PLA

During 1967 Mao called on the PLA under Lin Biao to step in on behalf of the Maoist Red Guards, but this politico-military task produced more division within the military than unified support for radical youths. Tensions surfaced in the summer, when Chen Zaidao (Ch'en Tsai-tao), a military commander in the key city of Wu-han, arrested two key radical CCP leaders. Faced with possible widespread revolt among local military commanders, Mao tilted toward the reestablishment of some order.

In 1968 Mao decided to rebuild the CCP and bring things under greater control. The military dispatched officers and soldiers to take over schools, factories, and government agencies. The army simultaneously forced millions of urban Red Guards to move to the hinterlands to live, thereby removing the most disruptive force from the cities. These drastic measures reflected Mao's disillusionment with the Red Guards' inability to overcome factional differences. The Soviet invasion of Czechoslovakia in August 1968, which greatly heightened China's fears for its security, gave these measures added urgency.

**The end of the radical period.** Thus in 1968 the society began to return to business, though not as usual. China's regular schools began to reopen, although the number of students in higher institutions represented only a small percentage of those three years before. In July yet another of Mao's "latest instructions" approved science and engineering education and called for the "return to production" of all graduates. In October 1968 a plenary session of the Central Committee met to call for the convening of a party congress and the rebuilding of the CCP apparatus. From that point on, the issue of who would inherit political power as the Cultural Revolution wound down became a central question of Chinese politics. (The answer came only with a coup against the radicals a month after Mao Zedong's death on Sept. 9, 1976.)

Fears of Soviet invasion

China's actions following the meeting of October 1968 suggested the degree to which fear of a Soviet invasion contributed to the closing down of the Cultural Revolution's most radical phase. Almost immediately after the meeting, China called on the United States to resume ambassadorial-level talks in Warsaw. Peking also renewed its conventional diplomacy—it had reduced its level of ambassadorial representation abroad to a single ambassador in Egypt—and quickly sought to expand the range of countries with which it enjoyed diplomatic relations.

China's concern stemmed partly from the Soviet leadership's articulation of the Brezhnev Doctrine in the wake of the invasion of Czechoslovakia. That doctrine explained the invasion in terms of the obligation of the Soviet Union and other Socialist countries to set things right if "scientific socialism" became threatened in any country in which a Communist Party had held power. To Peking's horror, even Hanoi came out in support of this threatening posture. Moscow had long made clear that it believed that a "military-bureaucratic dictatorship" had seized power from the "true Communists" in Peking. To add to Peking's concern, since 1966 the Soviet Union had been building up a sizable military force along the formerly demilitarized Sino-Soviet border. While the forces deployed as of late 1968 were not adequate for a full-scale invasion of China, they certainly posed a serious menace, especially given the political division and social chaos that still prevailed in much of the country.

When the Party Congress convened in April 1969, it did so in the wake of two bloody Sino-Soviet border clashes that had occurred in early and mid-March. Written into the new party constitution was an unprecedented step—Defense Minister Lin Biao was named as Mao's successor—and the military tightened its grip on the entire society. Both the Central Committee and the new party committees being established throughout the country were dominated by military men. Indeed, less than 30 percent of the Eighth Central Committee members elected in 1956 were reelected in 1969, and more than 40 percent of the members of the Ninth Central Committee chosen in 1969 held military posts.

Premier Zhou Enlai tried to cut back Lin Biao's power and to relieve some of the threat to China's security by engaging the Soviets in direct negotiations on the border dispute. A series of serious military clashes along the border, culminating in a limited but sanguinary Soviet thrust several miles into the Uighur Autonomous Region of Sinkiang, heightened tensions. Zhou briefly met with Soviet Premier Aleksey Kosygin at the Peking airport in early September, and the two agreed to hold formal talks. Nevertheless, Lin Biao declared martial law and used it to rid himself of some of his potential rivals. Several leaders who had been purged during 1966–68, including Liu Shaoqi, died under the martial law regime of 1969, and many others suffered severely.

Lin quickly encountered opposition, however. Mao became wary of a successor who seemed to want to assume power too quickly and began to maneuver against Lin. Premier Zhou Enlai joined forces with Mao in that effort, as possibly did Mao's wife, Jiang Qing. Mao's assistant, Chen Boda, decided to support Lin's cause, however. Therefore, while in 1970–71 many measures were undertaken to bring order and normalcy back to society, increasingly severe strains split the top leadership.

**Social changes.** By 1970 many of the stated goals of the Cultural Revolution had been translated into at least somewhat operational programs. These included initiatives designed to reduce what were termed the "three major differences"—those separating intellectual from manual labour, worker from peasant, and urban from rural.

Many measures had been taken to make the educational system less elitist. The number of years at each level of schooling was shortened, and admission to a university became based on the recommendations of a student's work unit rather than on competitive examination. All youths were required to engage in at least several years of manual labour before attending a university. Within schools, formal scholarship yielded in large measure to the study of politics and to vocational training. Examinations of the traditional type were abolished, and stress was placed on collective study. The authority of teachers in the classroom was seriously eroded. These trends reached their most extreme form when a student in the Northeast was made a national hero by the radicals because he turned in a blank examination paper and criticized his teacher for having asked him the examination questions in the first place.

The rustication program

Many bureaucrats were forced to leave the relative comfort of their offices for a stint in "May 7 cadre schools," usually farms run by a major urban unit. People from the urban unit had to live on the farm, typically in quite primitive conditions, for varying periods of time. (For some, this amounted to a number of years, although by about 1973 the time periods in general had been held to about six months to one year.) While on the farm the urban cadre would both engage in rigorous manual labour and undertake intensive, supervised study of ideology. The object was to reduce bureaucratic "airs."

Millions of Chinese youths were also sent to the countryside during these years. Initially, these were primarily Red Guard activists, but the program soon achieved a more general character, and it became expected that most middle-school graduates would head to the countryside. While in the hinterlands, these young people were instructed to "learn from the poor and lower middle peasants." Quite a few of these people were merely sent to the counties immediately adjacent to the city from which they came. Others, however, were sent over very long distances. Large groups from Shanghai, for instance, were made to settle in Sinkiang. This rustication was, in theory, permanent, although the vast majority of these people managed to stream back to the cities in the late 1970s, after Mao's death and the purge of his radical followers.

The system of medical care was also revamped. Serious efforts were made to force urban-based medical staffs to devote more effort to serving the needs of the peasants. This involved both the reassignment of medical personnel to rural areas and, more importantly, a major attempt to provide short-term training to rural medical personnel called "barefoot doctors." This latter initiative placed at least a minimal level of medical competence in many Chinese villages, and ideally the referral of more serious matters would be made to higher levels. Another prong of the effort in the medical arena was to place relatively greater stress on the use of Chinese traditional medicine, which relied more heavily on locally available herbs and on such low-cost methods as acupuncture. Western medicine was simply too expensive and specialized to be used effectively throughout China's vast hinterlands.

Traditional medicine

The Cultural Revolution was primarily an urban political phenomenon, and thus it had a very uneven effect on the peasants. Some villages, especially those near major cities, became caught up in the turmoil, but many peasants living in more remote areas experienced less interference from higher-level bureaucratic authorities than would normally have been the case.

Nevertheless, there were two dimensions of the Cultural Revolution that did seriously affect peasants' lives. First, the country adopted a policy of encouraging local rural self-sufficiency in foodstuffs. This policy stemmed from ideological and security considerations, and it had begun before the onset of the Cultural Revolution. Its major consequence was a stress on grain production so great that a quite irrational and uneconomical cropping pattern emerged. Second, great stress was placed on separating income from the amount of work performed by a peasant. Pressure was applied to raise the unit of income distribution to the brigade rather than the team (the former was several times larger than the latter), and an increasing share of the collective income was to be distributed on the basis of welfare and political criteria rather than on the basis of the amount of work performed.

**Struggle for the premiership.** As these programmatic aspects of the Cultural Revolution were being put into place and regularized, the political battle to determine who would inherit power at the top continued and intensified. Tensions first surfaced at a meeting of the Central Committee in the summer of 1970, when Chen Boda, Lin Biao, and their supporters made a series of remarks

that angered Mao Zedong. Mao then purged Chen as a warning to Lin. At the end of 1970 Mao also initiated a criticism of Lin's top supporters in the military forces, calling them to task for their arrogance and unwillingness to listen to civilian authority. The situation intensified during the spring of 1971 until Lin Biao's son, Lin Liguo (Lin Li-kuo), evidently began to put together plans for a possible coup against Mao should this prove the only way to save his father's position.

During this period, Zhou Enlai engaged in extremely delicate and secret diplomatic exchanges with the United States, and Mao agreed to a secret visit to Peking by the U.S. national security adviser Henry Kissinger in July 1971. That visit was one of the most dramatic events of the postwar international arena. At a time when the Vietnam War continued to blaze, China and the United States took major steps toward reducing their mutual antagonism in the face of the Soviet threat. Lin Biao strongly opposed this opening to the United States—probably in part because it would strengthen the political hand of its key architect in China, Zhou Enlai—and the Kissinger visit thus amounted to a major defeat for Lin.

Death of Lin Biao

Finally, in September 1971 Lin was killed in what the Chinese assert was an attempt to flee to the Soviet Union after an abortive assassination plot against Mao. Virtually the entire Chinese high military command was purged in the weeks following Lin's death.

Lin's demise had a profoundly disillusioning effect on many people who had supported Mao during the Cultural Revolution. Lin had been the high priest of the Mao cult, and millions had gone through tortuous struggles to elevate this chosen successor to power and throw out his "revisionist" challengers. They had in this quest attacked and tortured respected teachers, abused elderly citizens, humiliated old revolutionaries, and, in many cases, battled former friends in bloody confrontations. The sordid details of Lin's purported assassination plot and subsequent flight cast all this in the light of traditional, unprincipled power struggles, and untold millions concluded that they had simply been manipulated for personal political purposes.

Initially, Zhou Enlai was the major beneficiary of Lin's death, and from late 1971 through mid-1973 he tried to nudge the system back toward stability. He encouraged a revival and improvement of educational standards and brought numerous people back into office. China began again to increase its trade and other links with the outside world, while the domestic economy continued the forward momentum that had begun to build in 1969. Mao blessed these general moves but remained wary lest they call into question the basic value of having launched the Cultural Revolution in the first place. In Maoist thought it had always been possible for formerly wayward individuals to reform under pressure and again assume power.

During 1972 Mao suffered a serious stroke, and Zhou learned that he had a fatal cancer. These developments highlighted the continued uncertainty over the succession. In early 1973 Zhou and Mao brought Deng Xiaoping back to power in the hope of grooming him as a successor. But Deng had been the second most important victim purged by the radicals during the Cultural Revolution, and his reemergence made Jiang Qing, by then head of the radi-

Rivalry for succession

cals, and her followers desperate to return things to a more radical path. From mid-1973 Chinese politics shifted back and forth between Jiang and her followers—later dubbed the Gang of Four—and the supporters of Zhou and Deng. The former group favoured political mobilization, class struggle, anti-intellectualism, egalitarianism, and xenophobia, while the latter promoted economic growth, stability, educational progress, and a pragmatic foreign policy. Mao tried unsuccessfully to maintain a balance among these different forces while continuing in vain to search for a suitable successor.

The balance tipped back and forth—nudged by Mao first this way, then that—between the two groups. The radicals gained the upper hand from mid-1973 until mid-1974, during which time they whipped up a campaign that used criticism of Lin Biao and of Confucius as an allegorical vehicle for attacking Zhou and his policies. By July 1974, however, economic decline and increasing chaos made Mao shift back toward Zhou and Deng. With Zhou hospitalized, Deng assumed increasing power from the summer of 1974 through the late fall of 1975. During this time Deng sought (with Zhou's full support) to put the Four Modernizations (of agriculture, industry, science and technology, and defense) at the top of the country's agenda. To further this effort Deng continued to rehabilitate victims of the Cultural Revolution, and he commissioned the drafting of an important group of documents much like those developed in 1960–62. They laid out the basic principles for work in the party, industry, and science and technology. Their core elements were anathema to the radicals, who used their power in the mass media and the propaganda apparatus to attack Deng's efforts.

The radicals finally convinced Mao that Deng's policies would lead eventually to a repudiation of the Cultural Revolution and even of Mao himself. Mao therefore sanctioned criticism of these policies in the wall posters that were a favourite propaganda tool of the radicals. Zhou died in January 1976 and Deng delivered his eulogy. Deng then disappeared from public view and was formally purged (with Mao's backing) in April. The immediate reason for Deng's downfall was a group of massive demonstrations in Peking and other cities that took advantage of the traditional Ch'ing-ming festival to pay homage to Zhou's memory and thereby challenge the radicals.

In the immediate wake of Deng's purge, many of his followers also fell from power, and a political campaign was launched to "criticize Deng Xiaoping and his right deviationist attempt to reverse correct verdicts [on people during the Cultural Revolution]." Only Mao's death in September and the purge of the Gang of Four by a coalition of political, police, and military leaders in October 1976 brought this effort to vilify Deng to a close. Although it was officially ended by the 11th Party Congress in August 1977, the Cultural Revolution had in fact concluded with Mao's death and the purge of the Gang of Four.

**Consequences of the Cultural Revolution.** Although the Cultural Revolution largely bypassed the vast majority of the people, who lived in the rural areas, it had very serious consequences for the Chinese system as a whole. In the short run, of course, the political instability and zigzags in economic policy produced slower economic growth and a decline in the capacity of the government to deliver goods and services. Officials at all levels of the political system had learned that future shifts in policy would jeopardize those who had aggressively implemented previous policy. The result was bureaucratic timidity. In addition, with the death of Mao and the end of the Cultural Revolution, nearly 3,000,000 CCP members and other citizens awaited reinstatement after having been wrongfully purged.

Bold actions in the late 1970s went far toward coping with these immediate problems, but the Cultural Revolution also left more serious, longer-term legacies. First, a severe generation gap had been created in which young adults had been denied an education and had been taught to redress grievances by taking to the streets. Second, there was corruption within the CCP and the government, as the terror and accompanying scarcities of goods during the Cultural Revolution had forced people to fall back on traditional personal relationships and on extortion in order to get things done. Third, the CCP leadership and the system itself suffered a loss of legitimacy when millions of urban Chinese became disillusioned by the obvious power plays that took place in the name of political principle in the early and mid-1970s. Fourth, bitter factionalism was rampant as members of rival Cultural Revolution factions shared the same work unit, each still looking for ways to undermine the power of the other.

Four long-term legacies

## CHINA AFTER THE DEATH OF MAO

Perhaps never before in human history had a political leader unleashed such massive forces against the system that he had created. The resulting damage to that system was profound, and the goals that Mao Zedong sought to achieve ultimately remained elusive. The agenda he left behind for his successors was extraordinarily challenging.

**Readjustment and recovery.** Mao's death and the purge of the Gang of Four left Hua Guofeng (Hua Kuo-feng), a

compromise candidate elevated by Mao in the wake of the purge of Deng Xiaoping, as the official leader of China. Hua tried to consolidate his position by stressing his ties to Mao and his fidelity to Mao's basic ideas, but many others in the top leadership wanted to move away from precisely these issues, and Hua's position eroded over the remainder of the decade.

The ambivalent legacies of the Cultural Revolution were reflected in the members of the Politburo chosen just after the 11th Party Congress had convened in August 1977. Like Hua Guofeng, almost half of the members were individuals whose careers had gained from the Cultural Revolution; the other half were, like Deng Xiaoping, the Cultural Revolution's victims. While this balance would be reached only after a period of years, the tide quickly began to turn decidedly in favour of the latter group.

*Economic policy changes.* In the late fall of 1976 the CCP leadership tried to bring some order to the country by convening a series of national conferences designed to gain a grasp on the actual state of affairs. They moved quickly to appeal to workers' interests by reinstating wage bonuses. The economy had stagnated in 1976 largely because of political turmoil, and Mao's successors were anxious to start things moving again. Despite some uncertainty, Deng was rehabilitated and formally brought back into his previous offices in the summer of 1977.

Lacking detailed information on the economy, in early 1978 the leaders adopted an unrealistically ambitious 10-year plan and used the government's resources to the limit throughout that year to raise the rate of investment and achieve rapid economic growth. Much of that growth amounted to bringing back into play capacity that had lain idle due to political disruption. Future growth would be harder to achieve, and long-term trends in such matters as capital–output ratios made it increasingly clear that the old strategies would prove less and less equal to the task.

Resumption of foreign trade

One of the major changes of 1978 was China's sharp turn toward participation in the international economy. While in the 1970s there had been a resumption of the foreign trade that had been largely halted in the late 1960s, along with a far more active and Western-oriented diplomatic policy, the changes that occurred during and after 1978 were fundamental. China's leaders became convinced that large amounts of capital could be acquired from abroad to speed up the country's modernization, a change in attitude that elicited an almost frenetic response from foreign bankers and entrepreneurs.

These several strands came together in late 1978 at a major meeting of the CCP leadership, when China formally agreed to establish full diplomatic relations with the United States. China's leaders also formally adopted the Four Modernizations as the country's highest priority, with all other tasks to be subordinated to that of economic development. This set of priorities differed so fundamentally from those pursued during the Cultural Revolution that the implications for future policy and for the interests of various sectors of the population were profound. These decisions were sealed in December 1978.

The opening of China's economy to the outside world proceeded apace. In the late 1970s the country adopted a joint-venture law, and it subsequently enacted numerous other laws (such as one governing patents) to create an attractive environment for foreign capital. An initial experiment with the creation of several "special economic zones" along the southern coast in the late 1970s led in 1984 to a decision to open 14 cities to more intense engagement of the international economy. The idea was to move toward making ever larger sections of the country open to extensive foreign trade and investment.

Within the domestic economy, numerous experiments were undertaken with regard to finance, banking, planning, urban economic management, and rural policy. Of these, by far the most thorough and important were the series of measures taken toward the nearly 80 percent of the population that lived in the countryside. Prices paid for farm products were sharply increased as of 1979, thus pumping significant additional resources into the agricultural sector. The collective farming system was gradually dismantled in favour of a return to family farming. At first, families were allowed to contract for the use of collective land for a limited period of time. Subsequently, the period of those contracts was extended to "at least" 15 years, and subcontracting (essentially, allowing one family to accumulate large amounts of land) was permitted.

End of the communes

Related measures in the countryside allowed peasants far greater choice in what crops to plant and encouraged many to abandon farming altogether in favour of establishing small-scale industries, transport companies, and other services. Thus, patterns of work, ownership, and wealth changed markedly after 1978, and the rural situation continued to evolve rapidly. Exceptionally good weather during the early 1980s contributed to record harvests.

The reforms in the urban economy had more mixed results, largely because the economic system in the cities was so much more complex. Those reforms sought to provide material incentives for greater efficiency and to increase the use of market forces in the allocation of resources. Problems arose because of the relatively irrational price system, continuing managerial timidity, and the unwillingness of government officials to give up their power over economic decisions, among other difficulties. In the urban as well as the rural economy, the reformers tackled some of the fundamental building blocks of the Soviet system that had been imported during the 1950s.

*Political policy changes.* The reformers led by Deng Xiaoping tried after 1978 to reduce the level of political coercion in Chinese society. Millions of victims of past political campaigns were released from labour camps (especially during 1978–80), and bad "class labels" were removed from those stigmatized by them. This dramatically improved the career and social opportunities available to millions of former political pariahs. To a considerable extent, moreover, the range of things considered political was narrowed, so that such mundane elements as style of dress and grooming and preferences in music and hobbies were no longer considered of political significance. More importantly, the rejection of policy no longer resulted in political retaliation against that policy's advocates. Overall, the role of the Public Security (police) forces was cut back substantially, and the bureaucratic base of the police apparatus was divided up and weakened.

Rehabilitation of CCP cadres

The reformers also tried to make preparations for their own political succession. This involved first the rehabilitation of cadres who had been purged during the Cultural Revolution, most of which was accomplished in the late 1970s. These cadres in many cases were old and no longer fully able to meet the demands being made on them, and they were encouraged to retire. Younger, better educated people committed to reform were then brought into prominent positions. Deng proved masterful at maintaining a viable coalition among the diverse forces at the top. By the end of 1981 he had succeeded in nudging Hua Guofeng and others of the more rigid Maoists out of high-level positions. Although he refused to take the top position for himself, Deng saw his supporters become premier and general secretary of the CCP, and he worked hard to try to consolidate their hold on power.

During early 1982 the CCP leadership made a concerted attempt to restructure the leading bodies in both the government and the party, and much reorganization took place, with the appointment of many new officials. This general effort continued, with the focus increasingly on the bloated military establishment, but progress slowed considerably after the initial burst of organizational reformism.

Throughout 1982–85 the CCP carried out a "rectification" campaign designed to restore morals to its membership and weed out those who did not support reform. This campaign, along with similar efforts in the government, highlighted the increasing difficulties inherent in maintaining discipline and limiting corruption at a time of rapid change, when materialistic values were being officially propagated.

By the mid-1980s China was in transition, with core elements of the previous system called into question while the ultimate balance that would be struck remained unclear even to the top participants. The reform movement began to sour in 1985. Economically, financial decentralization and the two-price system combined with other

factors to produce inflation and encourage corruption. China's grain harvest for the decade peaked in 1984. In addition, the population, increasingly exposed to foreign ideas and standards of living through television and other means, put pressure on the government to speed the rate of change within the country.

These forces produced open unrest in late 1986 and again on a much larger scale in the spring of 1989. By 1989 popular disaffection with the CCP and the government had become widespread. Students—eventually joined by many others—took to the streets in dozens of cities from April to June to demand greater freedom and other changes. Government leaders, after initial hesitation, used the army to suppress this unrest in early June, with substantial loss of life. China's elderly revolutionaries then reverted to more conservative economic, political, and cultural policies in an attempt to reestablish firm control. Thus, at the close of the decade the resulting uncertainties left China vulnerable to the possibility of further political and social instability.

*Educational and cultural policy changes.* In the educational system, the reformers gave top priority to training technical, scientific, and scholarly talent to world-class standards. This involved the re-creation of a highly selective and elitist higher educational system, with admission based on competitive academic examination. Graduate study programs were introduced, and thousands of Chinese were sent abroad for advanced study. Foreign scholars were also used in large numbers to help upgrade the educational system. Somewhat ironically, the value the reformers attached to making money had the unintended consequence of encouraging many intelligent people to forgo intellectual careers in favour of more lucrative undertakings. In the cultural arena the reform effort encountered bitter opposition from many officials. But, on balance, the range of cultural fare available broadened greatly and new limits were constantly tested. Few groups had suffered so bitterly as did China's writers and artists, and policies in the 1980s reflected the ongoing battle between cultural liberals and more orthodox officials.

**International affairs.** China's foreign policy since 1978 generally has reflected the country's preoccupation with domestic economic development and its desire to promote a peaceful and stable environment in which to achieve these domestic goals. Except for its disagreement with Vietnam over that country's invasion of Cambodia (Kampuchea) in 1978, China generally has avoided disputes and encouraged the peaceful evolution of events in Asia. China adopted a policy of "one country, two systems" to provide a framework for the successful negotiation with Great Britain of the return of Hong Kong and adjacent territories, scheduled for 1997. China also became an advocate of arms control and assumed a more constructive, less combative stance in many international organizations.

Bilateral relations

In 1984 Chinese Premier Zhao Ziyang (Chao Tzu-yang) and U.S. Pres. Ronald Reagan exchanged state visits that underscored the continuing attempts of both nations to improve relations between them. China's various ties with Japan were also consolidated and expanded in the mid-1980s, and the two countries—in conjunction with the United States—attempted to ease tensions between North Korea and South Korea. Although at times the progress was halting, China's negotiations with the Soviet Union to increase bilateral trade and other exchanges were also successful; and in 1989 full normalized relations were restored between the two nations. (K.G.L.)

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

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 931, 932, 96/10, and 975, and the *Index.*

# NORTHEAST CHINA

## Heilungkiang

Heilungkiang (Hei-lung-chiang in Wade–Giles romanization, Heilongjiang in Pinyin), the northernmost province of China's Northeast region, is bounded on the north and east by the Soviet Union along the Amur River (Hei-lung Chiang) and the Ussuri (Wu-su-li) River, on the south by the Chinese province of Kirin, and on the west by the Inner Mongolia Autonomous Region of China. The province has an area of about 175,000 square miles (453,300 square kilometres). The capital is Harbin. Heilungkiang occupies about three-fifths of the area of the three Northeast provinces that formerly made up Manchuria and has more than one-third of the region's population. The long international border with the Soviet Union has been the scene of Sino-Russian clashes since the 17th century.

### PHYSICAL AND HUMAN GEOGRAPHY

Mountain ranges

**The land.** *Relief.* The province of Heilungkiang occupies more than half of the huge Manchurian Plain, surrounded on three sides by old mountain ranges of medium elevation. Its central part is the Sungari–Nen river plain, delimited by the Greater Khingan (Ta-hsing-an) Range on the west, the Lesser Khingan (Hsiao-hsing-an) Range on the north, and the Chang-kuang-ts'ai and Lao-yeh ranges on the east. The general elevation of the southern part of the Greater Khingan Range is around 3,000 feet (900 metres) above sea level; the elevation gradually drops toward the northeast to about 1,600 feet at its junction with the Lesser Khingan Range.

The Greater Khingan Range is composed mainly of igneous rocks resistant to erosion and weathering. Glaciers toward the end of the Tertiary Period (10,000,000 to 2,500,000 years ago) dug U-shaped valleys through which the tributaries of the Amur run. The structure of the Lesser Khingan Range is more complex. Its northern part is composed of granite, volcanic basalt, and other metamorphic rocks. The average elevation is about 2,300 feet; the granite peaks near I-ch'un rise to about 3,770 feet. The western slope facing the Nen (Nonni) River is gentle, while the eastern slope is steep. The southern end of the Lesser Khingan is composed of archlike, folded, stratified rock. The highest peaks may reach 3,280 feet or more, but the hills are generally lower. The valleys of the foreland are often broad and smooth, dotted with swamps. The rolling Sungari–Nen plain, at an elevation of 490 to 600 feet, has many bogs and swamps. In contrast, sand dunes occur in the drier western part of the plain.

*Drainage.* The Amur is the longest stream in the province. Its upper and middle sections serve as the international boundary for a distance of 1,180 miles (1,900 kilometres). Its average frozen period is 178 days at Ai-hui and 183 days at Mo-ho. Its chief tributary, the Sungari (Sung-hua) River, is the main waterway of the province, however. Most of the Sungari drainage system lies within the province. The Ussuri River forms the Sino-Soviet boundary on the east. It flows along a longitudinal valley between mountains. It is a broad, slow-moving river and has a tributary linking it with Lake Khanka (Hsing-k'ai Hu), the largest freshwater lake in East Asia. Only one-quarter of the lake, which is on the Sino-Soviet border, is in China.

*Soils.* The soils in the province are complex. In both the Greater and Lesser Khingan mountains soils differ with altitude. Black earths are prevalent in the foothills and mountain gray forest soils higher up. Still higher the cold, wet soils are podzolized; *i.e.,* the soluble salts along with soluble organic matter are leached out of the topsoil and deposited in an underlying subsoil. Such soils are of low fertility, and their cultivation causes erosion. The famous chernozem, or black soil, that covers one-fourth of the province is found in the Sungari–Nen river plain. Its eastern part has the best soils, yielding crops for years without fertilization. The chernozem lands form the main agricultural region of the province.

*Climate.* The province has severe winters, lasting five

to eight months. Summer is short but coincides with the rainy season, making it possible to raise temperate crops in most areas. There are considerable regional differences in climate. The northwest has a cold, wet, temperate climate with very cold winters; summer thaw is only superficial. Hu-ma, on the Amur River, has a mean temperature of −17.7° F (−27.6° C) in January. The July mean temperature is 74.5° F (23.6° C). There are only four months with mean temperatures over 50° F (10° C).

A temperate, wet climate prevails in the eastern section, in the drainage basin of the Ussuri River and the lower Sungari River. In the central core of the province the climate is temperate, with a deficiency of water and very severe winters. Nen-chiang, in the northern Manchurian Plain, has mean temperatures of −15.9° F (−26.6° C) in January and 70° F (21.1° C) in July. The mean annual precipitation is 20 inches (508 millimetres), most of which falls from June to September.

The southern part of the province is also very cold in winter but enjoys a warmer summer and a longer growing period. Harbin has mean temperatures of −2° F (−19° C) in January and 73° F (23° C) in July. Its mean annual precipitation is 20.6 inches.

*Plant and animal life.* The original vegetation of the province was forest-prairie, but it has been largely destroyed by cultivation; the remaining trees are predominantly poplars. There are many species of herbaceous plants, pasture grasses, and sorghums. The central part of the plain was originally prairie-steppe; the western part of the plain is a drier steppe.

The province has two distinct fauna districts. The first is the Manchurian Plain, which constitutes the larger part of the province. It has a predominance of temperate mixed forest fauna, with a significant admixture of elements of the Eurasian taiga. Among the district's representative animals are the Manchurian hare, the eastern field vole, the rat hamster, the Far Eastern finches, the buteo hawk, the needle-footed owl, and some species of flycatchers. Insects include the duckling beetle, the ground beetle, and the bumblebee. The region's fauna yield valuable fur and pelts, including the sable, panther, fox, chipmunk, Manchurian hare, and light-coloured polecat.

The second district, the Greater Khingan taiga, has a fauna more akin to that of the boreal forests of Europe and Siberia. The more common wildlife includes the brown bear, squirrels, chipmunks, some forest voles, the kolinsky (or Asiatic) mink, the wood hen, the crossbill, and the Siberian frog. Among the insects may be mentioned long-horned beetles, the ground beetle, and the Siberian silkworm. During the long, cold winter the birds migrate to warmer regions as far south as the Malay Peninsula.

Ethnic groups

**The people.** Although the population is predominantly Han (Chinese), there are other significant ethnic groups: Manchu, Koreans, Mongols, O-lun-chun, O-wen-k'o, Hoche (Nanai), and Chi-erh-chi-szu. Minor groups include Tibetans, Russians, and Yakuts. After the establishment of the Communist government, a number of autonomous counties and autonomous villages were created in areas inhabited by ethnic minorities. The Manchu form the largest minority group, distributed in the southern part of the province. They have been culturally assimilated by the Han majority. Most of them farm; their way of life is similar to that of the Han, and intermarriage is common, especially among the former nobility and the educated.

Korean immigration started in the mid-19th century. After the Japanese annexation of their country in 1910, a large number of Koreans emigrated to Heilungkiang and Kirin provinces, where they converted large areas of swampy wasteland into rice paddies. They live mostly in southeastern Heilungkiang, where many autonomous Korean villages have been established. The Hui (Chinese Muslims) live and work mostly in the bigger cities as merchants, handicraftsmen, and proprietors of beef and mutton restaurants. Those in An-ta and Chao-tung raise goats and dairy cattle. Mongolians live in the drier western part of the province, where they engage in farming and animal husbandry. One-third of them live in a Mongolian autonomous county.

The Daghur (Ta-hu-erh) Mongols live mostly in the upper Nen Valley, on the eastern foreland of the Greater Khingan Range. They are believed to have come from the north side of the Amur River during the 15th and 16th centuries. Hunters originally, they became the earliest farmers of Heilungkiang. Probably the O-lun-ch'un also came from north of the Amur River, later to settle in the Khingan ranges as farmers and hunters. They had domesticated the deer and were once known as the "deer riders." The O-lun-ch'un were among the earliest inhabitants of the upper and middle Amur. The O-wen-k'o tribesmen moved into the province in the 1st century AD. They are believed to be descendants of the Su-shen (Evenk) tribes of the Chou dynasty. They now live in the Amur River Valley, near Ai-hui. Originally hunters, they have learned to farm since 1949.

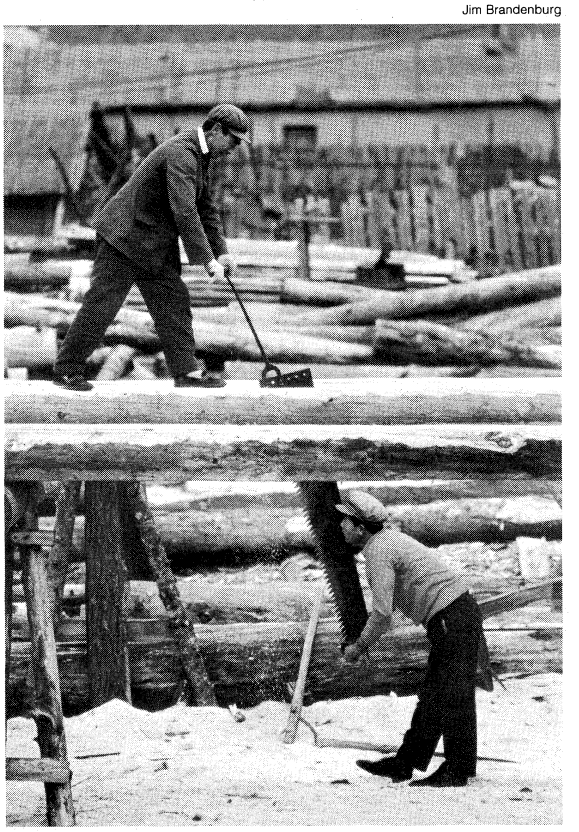
Russians entered the province at the end of the 19th century and during the early 20th. A great number of émigrés arrived after the Bolshevik Revolution. Those who became Chinese citizens have remained, many of them women who married Chinese and live mostly in Harbin and in the towns along the Amur River.

**The economy.** Despite its mineral and agricultural potential, the Heilungkiang economy was relatively underdeveloped until the mid-20th century. The process of economic growth began in the 1920s and '30s with the arrival of railroads and concomitant mineral exploitation. By the 1950s the provincial industrial output per capita was well above the national average.

*Agriculture and forestry.* Since 1949 large tracts of low-lying alluvial land have been reclaimed between the Sungari and Ussuri rivers. Large-scale state farms were established there, and millions of acres were brought under cultivation to produce sugar beets, soybeans, corn (maize), and wheat. Although cultivation in the region is highly mechanized, there is relatively little use of irrigation or chemical fertilizers. Heilungkiang is one of China's major grazing areas, its plains supporting millions of head of livestock. The province is also one of China's largest producers of raw timber.

*Industry.* Much of Heilungkiang's industry is based on mineral resources. During the 1950s emphasis was placed

Jim Brandenburg



Sawing timber at a lumber camp near I-ch'un in the forested region of the Lesser Khingan Range, Heilungkiang Province.

The Ta-ch'ing oil field

on the development of coal mining and thermal and hydroelectric power generation. The Ta-ch'ing oil field began operation in 1960 and subsequently developed into China's major inland field. Heilungkiang is the nation's largest oil producer, accounting for almost half of China's output. The oil is either exported or used in Heilungkiang's petrochemical industry.

The city of Chia-mu-ssu, on the Sungari River, was built up as a military and air base during the Japanese occupation (1931–45). Four strategic railways were completed linking the city with T'u-men, on the North Korean border; Sui-hua, to the west; Ho-kang, to the north; and Shuang-ya-shan, to the east. Rapid industrial development began under the Chinese Communists. The city began to produce machinery and electrical and telecommunication equipment. The Chia-mu-ssu paper mill is one of the largest in China, producing both for domestic needs and for export, and there is also a food-processing industry.

Another burgeoning industrial city is Shuang-ya-shan, east of Chia-mu-ssu. Its development began after World War II. The city has a number of large plants engaged in metal and food processing and in the production of lumber and construction materials. Tsitsihar (Ch'i-ch'i-ha-erh), the second largest city and former capital of the province, also grew phenomenally in the 1950s.

Harbin, the largest city and capital of the province, grew in 1898 as a construction base for the Chinese Eastern Railway across northern Manchuria. It soon became the major transportation hub and communications centre of northern Manchuria, with direct rail links to the Russian railroad network and to the Sea of Japan; through the South Manchurian Railway, it is linked with the Chinese and Korean rail networks and the Pacific. Numerous handicraft industries and small oil-pressing and flour mills are located there. By the 1950s the Harbin area had become one of China's primary industrial development centres, with an emphasis on heavy industry. It produces a variety of machinery and has chemical and fertilizer industries. The city is also a food-processing centre, as well as a producer of textiles, lumber, and construction materials.

*Transportation.* The five trunk rail lines are the Harbin–Lü-ta (Dairen), Harbin–Man-chou-li, Harbin–Chia-mu-ssu, Harbin–Sui-fen-ho, and Tsitsihar–Ssu-p'ing lines. There are also secondary lines. Inland waterways are not important, except for the Sungari River during the ice-free months. Most freight is carried by railroads or highway.

**Administration and social conditions.** The boundaries of Heilungkiang fluctuated during the early 20th century. From 1950 to 1954 the province was under the jurisdiction of the Northeast Military Administrative Commission based in Shen-yang. In 1954 Heilungkiang was placed directly under the central government, and its boundaries were expanded across the Sungari River to the Ussuri River frontier with the Soviet Union. During the Cultural Revolution, Heilungkiang was a pioneer in developing the Revolutionary Committees that became the dominant institutional form. Since the early 1980s the former system of provincial congresses has been restored. The province (*sheng*) is divided administratively into six prefectures (*ti-ch'ü*) and 10 prefecture-level municipalities (*shih*). These districts are further divided into counties (*hsien*) and county-level municipalities (*shih*).

Education

Harbin is an important educational centre, especially in engineering and applied science. The Harbin Polytechnical University was founded in 1920 to train technical personnel for the Chinese Eastern Railway. It offers specialized programs in departments of engineering and technology as well as a graduate school. Heilungkiang has numerous post-secondary educational institutions and a large corps of scientists and technicians. Almost all school-age children are enrolled in schools, and the province's literacy rate is substantially above the national norm.

HISTORY

The prehistoric population of the region appears to have consisted of people who bred pigs and horses; known as Tungids, they occupied much of northeastern Asia. Stone Age fishermen, the Sibirids and Ainoids, lived along the rivers and coast.

Heilungkiang was long sparsely inhabited by hunters and fishermen who used canoes, dogsleds, skis, and reindeer as transport. The town of San-hsing (now I-lan) was the home in the early 15th century AD of the ancestors of Nurhachi, the Manchu tribal leader who rose to power in the late 1500s through struggles with rival tribes and alliances with Manchu-related groups. Nurhachi's son, Prince Dorgon, ruled as regent during the reign of the first Ch'ing, or Manchu, emperor of China, the Shun-chih emperor.

Nurhachi's rise to power

During the 17th century the region became a zone of competition between Russia and China. Bands of musket-bearing Cossacks had been exacting tribute in furs from the tribes living along the Amur River, and in 1650 a Russian fort was built at Albazino on the river's north bank. The Ch'ing dynasty appointed a military governor to administer the region in 1683. The fort at Albazino was destroyed, and Russian retaliation was firmly opposed. By the Treaty of Nerchinsk (1689), the Russian government recognized Chinese suzerainty over the lands lying on both sides of the Amur River.

The tribes of Heilungkiang failed to recover their numerical strength after the Manchu rise to power and even after Manchu culture declined, despite the intentions of Ch'ing emperors to maintain their native language and way of life. Although large areas of Heilungkiang are fertile, agricultural development proceeded there very slowly because of the reluctance of Ch'ing rulers to allow the establishment of farms in their traditionally pastoral homeland. The region remained sparsely settled because access was difficult before the building of the railroads, and it was therefore highly vulnerable to Russian and Japanese expansion during the 19th century.

In 1858 Russia annexed the region north of the Amur River to its mouth and two years later the region east of the Ussuri River to the Sea of Japan, including the important seaport of Vladivostok (Hai-sheng-wei) and the Ussuri-Amur river port of Khabarovsk (Pai-li). The Russians occupied Heilungkiang from 1900 to 1905 and maintained their domination—despite their defeat by Japan in the Far East in 1904–05—through control of the strategic Chinese Eastern Railway, running through the region from west to east. After the Russian Revolution of 1917 the Bolsheviks renounced special privileges in northern Manchuria as a friendly gesture toward China. The province remained under Chinese control until Japan invaded Manchuria in September 1931. Heilungkiang became a part of the Japanese puppet state of Manchukuo from 1932 to 1945. On Aug. 15, 1945, just before Japan's unconditional surrender, Soviet troops entered Manchuria, but they evacuated it later to make way for Chinese Communist troops. After the Sino-Soviet rift in 1960, border clashes occurred repeatedly along the international border. (F.Hu./V.C.F.)

## Kirin

A province of the Northeast region of China, Kirin (Chi-lin in Wade–Giles romanization, Jilin in Pinyin) borders the Primorsky Kray (Maritime Territory) of the Soviet Union to the east, North Korea to the southeast, the Chinese provinces of Liaoning to the south and Heilungkiang to the north, and the Inner Mongolia Autonomous Region to the west. It has an area of about 72,400 square miles (187,400 square kilometres). The capital (since 1954) of the province is Ch'ang-ch'un.

PHYSICAL AND HUMAN GEOGRAPHY

**The land.** *Relief.* The province may be divided into three parts: the eastern mountains, the western plains, and a transitional zone of rolling hills between them. Elevation decreases from the highlands in the southeast toward the Manchurian Plain in the northwestern part of the province. The mountains of eastern Kirin take the form of parallel ranges with the Cathaysian or Sinian trend and are separated by broad valleys. The most famous of the ranges is the Ch'ang-pai Mountains close to the Korean border. One of its snow-covered summits is 9,000 feet (2,744 metres) above sea level—the highest peak in northeastern China. The summit is formed by a volcanic crater occupied by a lake. The range is the source of three

Mountains

important rivers: the Sungari, the Yalu, and the Tumen. The middle section of the Manchurian Plain forms the northwestern part of the province and constitutes three-eighths of its area. It has a rolling topography, with an average elevation of about 650 feet above sea level.

*Drainage.* The Yalu and Tumen rivers flow in opposite directions along the Sino-Korean border. The Yalu runs southwest to Korea Bay, the Tumen down the Ch'ang-pai range northeastward to the Sea of Japan. The two rivers are of great strategic importance, guarding the land approaches to northeastern China from the Korean peninsula. The Sungari River is the major stream of Kirin. It flows for almost 500 miles (800 kilometres) within the province, draining an area of more than 30,000 square miles. Its upper course runs northwest in a series of rapids through heavily forested mountains before it enters the Sungari Reservoir, a man-made lake. Emerging from the reservoir, the Sungari flows past Chi-lin (Kirin city), situated at the head of navigation of the Sungari River and at the geographical centre of the province. The river enters the Manchurian Plain and is shortly afterward joined by its chief tributary, the Nen River, which is in fact larger than the Sungari. It then turns sharply east to run along the provincial boundary for a short distance before it leaves Kirin Province.

*Soils.* There are two main types of soil in the province: podzols in the eastern mountainous region and black earth in the western plains. The podzols occur in several forms and are of both high and low fertility. Central and western Kirin are the areas of the black earths of the Manchurian Plain. Of high fertility and containing a high percentage of organic matter, they form good arable land. The young alluvial soils along the Sungari and its tributaries also provide excellent land for cultivation. In western Kirin are saline soils, alkaline soils, and high-alkaline soils of low agricultural value.

*Climate.* Kirin Province forms a transitional climatic zone between the northern and southern portions of China's Northeast. The winter is cold and long, and rivers are frozen for about five months; the ice on the Sungari is thick enough to support mule carts. Ch'ang-ch'un, near the centre of China's Northeast, has mean temperatures of 2° F (−17° C) for January and 74° F (23° C) for July. It has a mean annual precipitation of about 25 inches (630 millimetres), more than 80 percent of it during the five warm months from May to September. Precipitation increases southeastward to more than 40 inches in the Ch'ang-pai Mountains area but decreases westward; the Manchurian Plain receives only 16 inches.

Temperature and precipitation

*Plant and animal life.* The natural vegetation is prairie grass in the western plains and mixed conifer and broad-leaved deciduous forest in the eastern mountainous area. The vegetation in the eastern mountains includes tree species such as the Japanese red pine, Manchurian ash, fish-scale pine, larch, birch, oak, willow, elm, and the Manchurian walnut. In the deep mountain interior, virgin forest has been preserved. Tree types are distributed in distinct belts depending mainly on altitude: between 800 feet and 1,600 feet of elevation is the deciduous broad-leaved belt, mainly mountain willow and thumb; between 1,600 feet and 3,000 feet is found mixed coniferous and broad-leaved forest; between 3,000 feet and 5,900 feet occurs coniferous forest; and from 5,900 feet to 6,900 feet is found mountain birch.

Tree types

Many valuable wild animals and medicinal plants are found in the forested mountain areas. The Manchurian hare, valued for its fur, and some species of rodent such as the rat hamster and the eastern field vole are believed to be peculiar to the Manchurian forest. Among birds, finches, the buteo hawk, the needle-footed owl, the black and white barrier, and certain species of flycatcher are typical. Among semiaquatic animals, the lungless newts are notable. Certain species of snakes, such as the Schrenk racer, found in the inhabited areas of the Northeast and Korea, live in a semidomesticated state and are used to eliminate harmful rodents in orchards and gardens. The European wild boar, the common hedgehog, the Asian red deer, the harvest mouse, and the field mouse are among the more common Eurasian species. Valuable pelts include fox, chipmunk, the light-coloured polecat, the Manchurian hare, and the sable. The sable population, however, has become very small.

**The people.** Han (Chinese) predominate throughout the province, except in an autonomous prefecture for Koreans that is contiguous with North Korea. Most of the Manchu live in the central part of the province, in Chi-lin Municipality and T'ung-hua Prefecture. A few Hui (Chinese Muslims) are distributed in the cities and towns of the province, and some Mongolians are to be found in the Pai-ch'eng area in northwestern Kirin.

**The economy.** *Agriculture and forestry.* Kirin is a significant producer of soybeans, corn (maize), sugar beets, and oil-bearing crops, and its farmers earn an income well above the national average. This is in part due to the high land–labour ratio but also to the relatively high yields produced with the use of agricultural machinery and irrigation. In the eastern uplands, rice and millet are the staple produce of the Korean population.

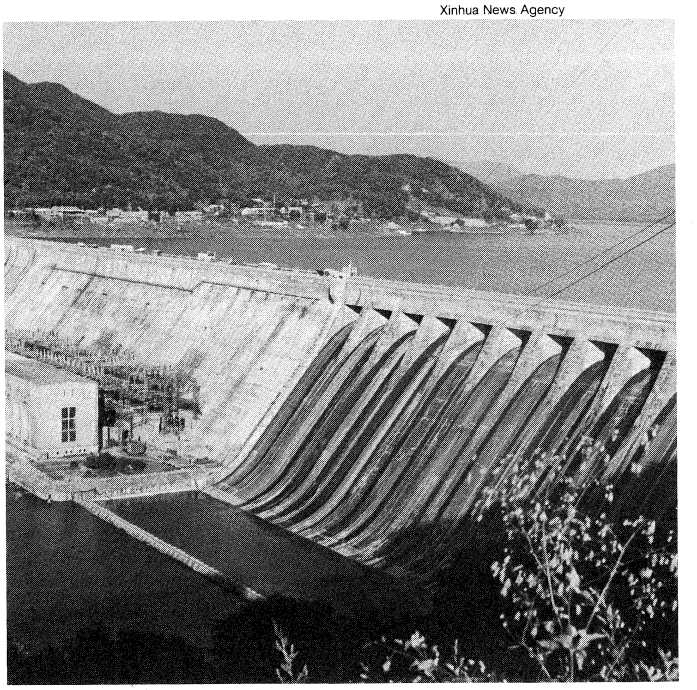
In the upper Sungari Basin, timber production for construction and milling is a major economic activity. In the eastern sections of the province, much of the forestry activity centres on pulp and paper production.

*Mining and power resources.* The major minerals of the province include coal, iron ore, copper, zinc, and gold. Coal is found in the southeast, near the Yalu River border with North Korea. Many smaller local mines also supply provincial needs. The major hydroelectric power installation, the Feng-man station on the Sungari River southeast of Chi-lin, was built by the Japanese during World War II and rebuilt by the Soviets in the 1950s.

*Industry.* Kirin is relatively highly industrialized and is a major producer of chemicals, machine tools, power, and forest products. Originally a lumbering and food-processing centre, the province acquired a heavy industrial base during the Japanese occupation of 1931–45. It was a major beneficiary of Soviet investment in the mid-1950s, acquiring an automotive industry and metals and fabrication industries. In the 1960s the Feng-man hydroelectric power station made possible the development of chemical and ferro-alloy industries. Most industry is concentrated in the two largest cities in the province—Ch'ang-ch'un and Chi-lin.

*Transportation.* Most of the municipalities and counties in the province have direct access to a rail line. The Sungari is the main artery of the inland navigation network. Its tributary the Hui-fa River and the Tumen River are

Xinhua News Agency



The Feng-man hydroelectric power station on the Sungari River in Kirin Province.

both navigable by wooden vessels. The Yalu is navigable by steamers up to Yü-shu-lin-tzu and by wooden vessels above that point. The highway network has regional centres at Ch'ang-ch'un, Chi-lin, Yen-chi, and T'ung-hua.

**Administration and social conditions.** Kirin was one of the Manchurian provinces until the Communists subordinated the region to the Northeast Military Administrative Commission in 1950. In 1954 the province was enlarged through the addition of a strip of territory annexed from northern Liaoning, including the cities of Ssu-p'ing, Liao-yüan, and T'ung-hua and a portion of Heilungkiang's steppe district near Pai-ch'eng. After 1954, with the abolition of the regional government, the province came under direct administration of the central government. Kirin is divided into two prefectures (*ti'ch'ü*), one autonomous prefecture (*tzu-chih-chou*), and five prefecture-level municipalities (*shih*). The province is further divided into counties, some of which are under the control of the municipalities or are autonomous counties established for minority nationalities.

Education

Kirin's educational facilities are well developed, with more than 30 universities and other post-secondary institutions. Overall literacy rates are significantly higher than the national average, as is the proportion of the population with at least a primary level education. Medical services are provided by hospitals and clinics staffed by medical workers, including doctors and practitioners of Chinese medicine.

HISTORY

In early modern times the Kirin region was inhabited by groups of steppe and forest dwellers and was at times loosely united politically by leaders who presented tribute of furs, ginseng, and pearls at the court of the Ming emperors of China. In the late 16th century the Hurka tribe dominated the region before being defeated by the Manchu leader Nurhachi. After the establishment of the Ch'ing, or Manchu, dynasty in 1644, the region was at first directly administered by a military governor posted in the town of Chi-lin, and the region was thereafter referred to as Kirin.

Despite the Ch'ing government policy of discouraging agricultural settlement in the Manchu homeland, large numbers of settlers from North China established farms in the region during the 18th century, a period of rapid population expansion in China proper. In 1799 the transition to an agricultural economy was officially recognized with the establishment of a prefectural government at Ch'ang-ch'un to administer the new settlements. In the late 19th century, economic development accelerated in Kirin with the building of railways and industries processing agricultural products. This development encouraged a new influx of Chinese settlers and led to conflict between Russia and Japan over economic interests in the area.

Kirin was created a province in 1907, near the end of the Ch'ing dynasty, and was occupied by the Japanese Army in 1931, becoming a part of the puppet state of Manchukuo (1932–45). Just before Japan's surrender to the Allies on Aug. 15, 1945, Soviet forces entered the region, dismantled key industrial installations, and removed them to the Soviet Union. Following the withdrawal of Soviet troops, Chinese Nationalists moved in, but by 1948 they had been driven out by Chinese Communist forces.

(F.A.L./V.C.F.)

## Liaoning

Liaoning (Liao-ning in Wade–Giles romanization, Liaoning in Pinyin) is the southernmost of the three Chinese provinces of the Northeast region (formerly called Manchuria). With an area of 56,300 square miles (145,700 square kilometres), Liaoning is bounded on the northeast by the province of Kirin, on the east by North Korea, on the south by the Yellow Sea, on the southwest by the province of Hopeh, and on the northwest by the Inner Mongolia Autonomous Region. The provincial capital is Shen-yang (formerly Mukden).

The area, a region of early Chinese settlement in the Northeast, was known as Sheng-ching in Ch'ing, or Manchu, times (1644–1911/12). The area was redefined in 1903 and named Fengtien; in 1928 the boundaries were altered and it was renamed Liaoning. From 1947 to 1954 the territory was divided into a western province, Liaosi, and an eastern province, Liaotung. In 1954, however, a northern zone was detached and it was reestablished as a single province. It achieved its present form in 1956, when the former province of Jehol was partitioned and a portion added to Liaoning. Liaoning, Liaosi, and Liaotung all take their names from the Liao River. Precedents for the names go back to Han times.

PHYSICAL AND HUMAN GEOGRAPHY

**The land.** Liaoning consists essentially of a central lowland, with Shen-yang at its centre, flanked by mountain masses to east and west. A southward extension of the eastern highlands forms the Liaotung Peninsula. There are four main topographical regions: the central plains, the Liaotung Peninsula, the western highlands, and the eastern mountain zone.

*The four main regions.* The central plains are the most important area in the province. Structurally, the depression that it occupies is continuous with that of the North China Plain, but, topographically, the Liaoning plains are erosional rather than depositional in character. The relief of the plain is undulating but low, and natural drainage is inadequate in many places, creating swamps, some of which have been redrained. Most of the landscape of the central plains consists of cultivated fields. Undeveloped areas include swamps and sand formations. The soils of the middle of the Liao lowland are of the calcareous alluvial type; those of the peripheries to east and west, of brown-forest types; those of the northern peripheries, red earths. The swamps have gley soils (having a sticky layer of clay under the waterlogged surface). Wild animals are scarce, apart from rodents. Locusts are the most destructive pest.

The Liaotung Peninsula is a rugged, mountainous area with a rocky coast. The usual height of the country is 1,000 to 1,500 feet (300 to 450 metres) above sea level. The rock types are very mixed, a fact that tends to create a complex and varied topography. Structurally, the peninsula represents a part of the same fold system as Shantung. The coastline is experiencing submergence. The soils of the peninsula, like the rock types and the topography, are very mixed and varied. Most of the best soils are of brown-forest type or of red or yellow loess (an unstratified wind-borne loamy deposit). There has been serious soil erosion, and skeletal soils occur on the steeper slopes. The natural vegetation is not well preserved because of the extent of cultivation and settlement. The forests that remain, mostly on the eastern sides of the hills, contain birch, lime, elm, and pine, together with typical Manchurian trees—oak, apple, and ash. On the western sides, trees are scarce. Wild animal life is now meagre, being almost limited to rodents.

Western Liaoning, fringing the northern shore of Liaotung Bay between Shan-hai-kuan and Chin-chou, is predominantly a highland area. These highlands comprise the broken and eroded fringe of the Mongolian Plateau. They rise in Liaoning to general heights of about 1,500 feet. Toward the sea the mountains have been intensely eroded by fast-flowing rivers, so that a complex mass of valleys and ridges has been formed. Vegetation is very mixed and includes oaks, birches, pines, limes, and spruces. In former times, especially between 1911 and 1948, there was much indiscriminate cutting and thinning of the forests, so that many areas now have scattered woodland where formerly thick forest stood. The animal life of the western highlands is impoverished by the extent of both forest clearance and human settlement, but it includes the wolf, fox, marmot, and some kinds of deer.

The eastern mountain zone lies to the east of Shen-yang. The least developed part of the province, it consists of a complex mountain mass, extending northward into Kirin Province, with elevations averaging about 1,500 feet. Natural vegetation is predominantly a mixed coniferous and broad-leaved forest.

*Climate.* Temperature extremes and precipitation amounts vary with proximity to the coast. At Lü-ta (Dairen), at the southern tip of the Liaotung Peninsula,

the January mean temperature is 23° F (−5° C) and that for July is 74° F (23° C); for Shen-yang, in central Liaoning, the respective mean temperatures are 10° F (−12° C) and 77° F (25° C). At Lü-ta there are about 200 frost-free days, while at Shen-yang there are between 160 and 180 frost-free days per year. Rainfall in Liaoning as a whole diminishes consistently from southeast to northwest. Average annual precipitation is about 29 inches (740 millimetres), three-quarters of it falling in the months of June, July, August, and September, and almost none during December, January, and February. The summer rainfall is often torrential, but everywhere the scarcity of spring precipitation tends to leave crops short of water.

Ethnic composition and distribution

**The people.** In Liaoning almost all of the population is recorded as Han (Chinese). The bulk of the national minority population is Manchu, located mainly in the Liao Valley and around Shen-yang, in the southeast around Tan-tung, and in the southwest around Chin-chou. The second significant minority is that of the Mongols, who are located toward the frontier of the Inner Mongolia Autonomous Region to the west. Broadly speaking, the Hui (Chinese Muslim) minority follows the Manchu in its distribution. There are two autonomous counties representing the Mongolian minority nationality. One is centred on the coal town of Fu-hsin, and the other is in the southwest. A small Korean minority is located near the Korean frontier.

Apart from the registered minority populations, many of the Han people of modern Liaoning have origins that are wholly or partly non-Han, usually Mongol or Manchu. Many of them are now totally assimilated into the Chinese sector of the population, in language and custom as well as in the adoption of contemporary Han life-styles.

All of the large cities are industrial, and some experienced spectacular growth in the 1950s. They include the following: Shen-yang (earlier called Mukden, Feng-t'ien, or Sheng-ching); Lü-ta (comprises Lu-shun, formerly called Port Arthur, and Dairen); Fu-shun; and An-shan.

Industrial importance

**The economy.** The economy of Liaoning is by far the strongest in the Northeast and is one of the strongest provincial economies in China. Liaoning is one of the country's principal industrial provinces. One reason for the high level of development in Liaoning is the level of capitalization—very high by Chinese standards—which is based both on investments made under the government since 1949 and on important foreign investments made between 1896 and 1945, mainly by the Japanese.

*Resources.* Liaoning Province is rich in mineral resources, especially iron ore and coal. Most of the iron ores of Liaoning are concentrated in a triangular area to the south of Shen-yang. These ores are generally easy to mine but are of relatively poor quality; ores of better quality occur in the northeastern part of the province. Coal is more widespread, and its distribution partly overlaps that of iron. Coal is exploited in three main areas to the north, east and southeast, and west of Shen-yang. Fu-shun, east of Shen-yang, and Fu-hsin, to the west, have two of the most important collieries in China. Both were exploited under the Japanese but have been expanded since the Communists came to power. Apart from its use as fuel and in smelting, coal is used in Liaoning to produce synthetic petroleum. Petroleum is also produced from oil shale, which occurs in the Fu-shun area and in western Liaoning, generally overlying coal seams. The Liao River oil field, first developed in the late 1960s, has become China's fourth largest onshore producer.

Mineral resources and mining

Rich reserves of manganese ore occur in western Liaoning and in the southeast. In the eastern mountain area there are substantial deposits of copper, lead, and zinc; smaller similar deposits occur in the west, together with an important deposit of molybdenum. Important concentrations of magnesium ore are found southwest of Shen-yang. There are also reserves of other minerals, including bauxite, gold, and diamonds, and sea salt is produced.

*Agriculture, forestry, and fishing.* Agricultural advances in Liaoning have been less spectacular than industrial development. There are several reasons for this. Investment has always been much heavier in industry than in farming. The province's inheritance from the Japanese phase was much less valuable in agriculture than in industry. Liaoning also suffers from both natural calamities, such as spring droughts, and from backward cultivation methods in many places, which result in low yields. Exceptional opportunities for employment in industry also tend to deprive agriculture of much of the best labour, in spite of policies designed to prevent this. Yet, in much of Liaoning, topography and soils, and even climate, are quite favourable to agriculture; and the degree of farm mechanization, irrigation, and chemical fertilization is very high by Chinese standards. Liaoning is, nevertheless, an importer of food. It must depend partly on food from Kirin and Heilungkiang to the north and partly on imported grain.

Agricultural problems

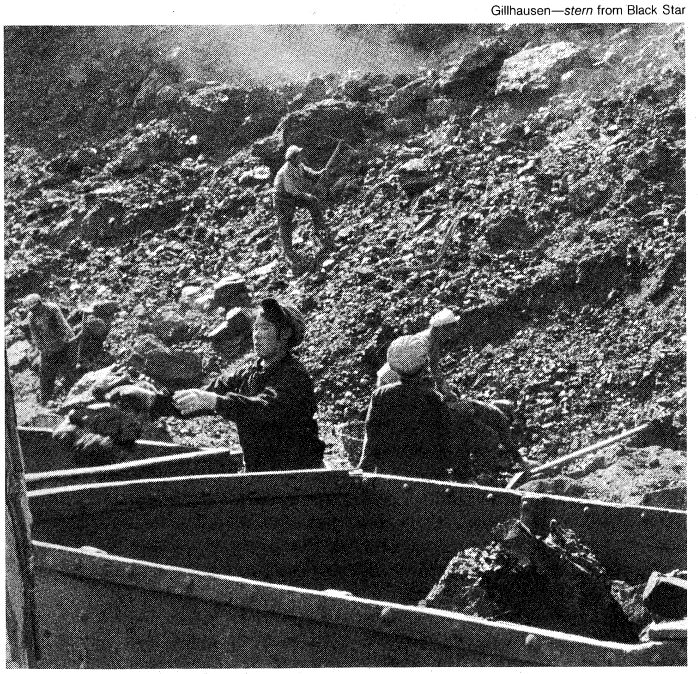
The summer in Liaoning is not long. Few places, consequently, produce two crops per year. The central plain is the best farming area, and the Liaotung Peninsula, with its shorter winter, has a diversified agriculture. Peanuts (groundnuts), sugar beets, and pears are among the province's major crops. Part of the cultivated area is used for industrial crops (primarily cotton and tobacco) or for export crops like apples; the rest is used for grain crops, vegetables, and soybeans. Higher yielding corn and rice, formerly grown mainly in the east and the southeast, have tended to supplant millet and kaoliang (a variety of grain sorghum) in the plains. The chestnut-leaved oak feeds the tussah silkworm; Liaoning is China's major producer of tussah silk. The forests support commercial lumbering, but the supply of mature trees is limited, due to previous overexploitation, and the output of lumber is low.

Livestock raising is of minor importance. Pigs are bred mainly in the south and central parts, and in the west other animals are raised. Fishing in the Yellow Sea, on the other hand, is a major income-earning activity.

*Industry.* During the mid-1950s considerable capital investment was made in Liaoning, primarily in heavy industry. In heavy industrial production, Liaoning ranks first among the provinces of China, producing, in addition to pig iron and steel, a substantial part of China's cement, crude oil, and electrical power. Apart from iron and steel, the industries of Liaoning include nonferrous-metal processing, machinery manufacture, and chemical manufacture, as well as such light industries as those producing textiles, foodstuffs, and paper.

The primary focus of investment in Liaoning has been the industrial network centred on the An-shan iron and steel complex. An-shan, south of Shen-yang, is the industrial heart of Liaoning and is China's principal steel

Gillhausen—*stern* from Black Star



Strip-mining a coalfield in Fu-shun, Liaoning Province.

Principal industrial areas

centre; taken as a whole, it is the biggest single enterprise in industrial investment ever established in China. Shen-yang, also a key industrial city, has been granted provincial-level powers in economic planning. It has a wide and varied range of heavy and light industries. Fu-shun is part of the Shen-yang complex, but its industries are all based on coal. Liao-yang, directly south of Shen-yang, is a textile centre. Lü-ta, near the tip of the Liaotung Peninsula, is of obvious strategic importance. It has the best harbour in the Northeast and is a major Chinese port. Modern engineering industries have been developed there, including shipbuilding and locomotive production. Lü-ta also has provincial-level economic authority and is one of China's coastal cities open to foreign investment. Western Liaoning, centred on Chin-chou, is less advanced than the Shen-yang–An-shan area or Lü-ta, but it has valuable mineral resources and some industries.

*Transportation.* The rail transportation facilities of Liaoning are the best in China, and the tonnage transported is also the second highest for any province. The backbone of the transportation system is the Ch'ang-ch'un–Lü-ta railway (formerly called the South Manchurian Railway), which passes through Shen-yang and which was double-tracked in 1954. Rail traffic primarily comprises either industrial freight or food products in bulk.

Highways in the province are extensive, but many are of poor quality. Many of the goods transported by road are carried in carts by animals in the traditional style.

Shipping

Navigation by sea or river carries almost none of the internal traffic of Liaoning, but sea navigation is of great importance for transport to other parts of China. Lü-ta is the largest port, followed by Ying-k'ou.

**Administration and social conditions.** *Government.* From 1950 to 1954 Liaoning, divided into Liaotung and Liaosi, was under the jurisdiction of the Northeast Military Administrative Commission. In 1954 the province was reunited, and Liaoning was brought under direct central government rule. During the Cultural Revolution the province was in the forefront of radical reform; it continued to support many of the revolution's initiatives after the death of Mao Zedong (Mao Tse-tung) in 1976, although Liaoning subsequently led the national effort to modernize its industrial plants and open its doors to trade with the outside world. Liaoning Province (*sheng*) is now divided into two prefectures (*ti-ch'ü*) and 10 municipalities (*shih*) at the prefecture level. All of the province's counties (*hsien*) are under municipal administration.

*Education.* Liaoning's economically advanced society has developed significant educational and scientific resources. Only the three province-level cities of Shanghai, Tientsin, and Peking have a greater proportion of population educated at least to the primary level. Liaoning is second only to Peking in adult literacy and has more scientists and technical staff than Shanghai. There are numerous universities and other post-secondary educational institutions in the province.

**Culture.** Though Sinitic in tradition, Liaoning's culture has been shaped by a kind of "outsider" perspective. Long periods of non-Han rule and the late onset of significant migration to the area have given Liaoning a frontier character. Many of the clan-centred traditions of central and South China have been attenuated in this still mobile society, where roots are shallow and the nuclear family predominates. In addition, as an arena of competition for influence by the Japanese and the Russians, the province has a degree of cosmopolitanism lacking in many other areas of China.

## HISTORY

Most of the present province of Liaoning fell within the confines of the earliest Great Wall of China, built during the reign of the first Ch'in emperor (221–210/209 BC), and hence formed part of China from early times. The environment and traditional Chinese civilization of the central plain in Liaoning continue into the North China Plain and into Shantung to the south. Political power in the region often passed to nonagricultural peoples, such as the Khitan, who invaded the area in the 10th century AD and established the Liao dynasty, and the Juchen, who founded the Ch'in dynasty in the 12th century. During the Yüan dynasty, Chinese (Mongol) power was reasserted over the region, and ties remained relatively close through the Ming dynasty. Chinese immigration from the south also has a long history, but until the last years of the 19th century it was on a modest scale. The traditional economy of Liaoning was one of Chinese peasant farm production in the plains; on the peripheries the economy was based in various places on herding, forestry, fishing, mining, and estate farming. Neither the aristocratic holders of the landed estates nor the people who depended on animals or on the products of the forests or rivers were generally Chinese; as elsewhere in the Northeast, they were of Mongol or Manchu stock.

Under the Ch'ing dynasty (1644–1911/12), whose own origins lay in the Manchu frontier aristocracy, official efforts were made to protect the Northeast from Chinese encroachment with the exception of the old Chinese-settled area in the Liao Valley. This policy was gradually abandoned, partly because of the pressure of Russian influence in the north.

Toward the end of the 19th century and into the 20th, two intersecting forces together created a radically novel situation in Liaoning. One was foreign interference—Russian and, later, Japanese. The other was rapidly increasing Chinese immigration. These two forces resulted in the expansion of the economy of Liaoning at a wholly unprecedented rate. Immigrants traveled to Liaoning both by the Shan-hai-kuan land corridor in the southwest and from Shantung by sea to the Liaotung Peninsula in the east. The second group were the more numerous and have been the more successful, both because of the relative ease of travel by sea and because of the better opportunities for both work and settlement in Liaotung. During the first half of the 20th century, the population of the province rose to a size proportionate to its present figure. The population movement was based on seasonal migration for farm work, with about half of the migrants remaining in Manchuria after the harvests each year. Many strong Shantung communities grew up in Liaotung, with traditions of mutual dependence and help.

Patterns of Chinese immigration

The all-important South Manchurian Railway was constructed by the Russians between 1896 and 1903. This railway linked the new Liaotung port of Lü-ta with Ch'ang-ch'un, in Kirin Province, as well as with Harbin in Heilungkiang Province, and with the then new Chinese Eastern Railway branch of the Trans-Siberian Railway. The South Manchurian Railway passed close to Shen-yang, replaced navigation on the Liao and much of the old cart transport, and bypassed the old port of Ying-k'ou. The foundations of the modern geography of Liaoning were laid by this railway. In 1907 the Russian railway, port, and territorial privileges were transferred to Japan at the conclusion of the Russo-Japanese War of 1904–05. From that time, Japan continually strengthened its hold on the economic life of Liaoning and all of Manchuria, partly through physical control but also through an active and successful policy of investment and economic expansion. From 1932 to 1945 Liaoning was part of the Japanese-dominated, "independent" state of Manchukuo. Throughout this period, which included the Sino-Japanese War and World War II (1937–45), Japanese policy aimed to develop the resources of Liaoning in a manner that would complement the economic strength of Japan. Heavy industry was particularly developed. This concentration on heavy industry at the expense of light industry and agriculture has been criticized; but, in the face of China's general lack of heavy industrial capacity, these installations have stood her in good stead since 1949.

The period of Japanese domination

Shen-yang fell to the Chinese Communists in 1948. The industrial installations of Liaoning had suffered heavily from war damage and from Soviet seizures of stockpiles and machinery. The new government made the restoration of the Northeast one of its first priorities. After the defeat of the Japanese in 1945, the Soviet Union took over from Japan and retained residual rights to share with China the use of the naval base facilities at Lü-shun and some rights on the former South Manchurian Railway. These rights were given up in 1955. (F.A.L./V.C.F./Ed.)

# NORTH CHINA

## Honan

Honan (Ho-nan in Wade–Giles romanization, Henan in Pinyin), a small province in the north central part of China, has an area of 64,700 square miles (167,700 square kilometres). The province stretches some 300 miles (500 kilometres) from north to south and 350 miles east to west at its widest point. It is bounded on the north by Shansi and Hopeh; on the east by Shantung and Anhwei; on the west by Shensi; and on the south by Hupeh. The Huang Ho (Yellow River) divides the province into two unequal parts—one-sixth north and five-sixths south of the river—and thus to some extent belies the name Honan ("South of the River"). K'ai-feng, the former capital, has been superseded by Cheng-chou, where the Peking–Han-k'ou railway crosses the Huang Ho and meets the Lung-hai Railway running from east to west.

### PHYSICAL AND HUMAN GEOGRAPHY

**The land.** *Relief.* Honan can be divided topographically into two parts, the western highlands and the eastern plain. In the northwest the rugged T'ai-hang and Chung-t'iao mountains form the steep eastern edge of the Shansi Plateau, rising in places above 5,000 feet (1,524 metres). They are part of the T'ai-hang fold system of Permian times and have a general northeast to southwest trend. They mark the northern border of the province.

The western highlands

South of the Huang Ho there is a broad stretch of upland comprising a number of moderately high mountain basins, the main ranges being the Hsiung-erh and Fu-niu. These mountains, which have an east–west trend, are the eastern extension of the Tsinling Mountains axis that divides China geologically and geographically into North and South. The T'ung-pai and Ta-pieh ranges form a further extension of this axis, running in a southeasterly direction and marking the border between Honan and Hupeh. The T'ung-pai range is separated from the Fu-niu by a gap some 75 to 100 miles wide cut by the T'ang and T'ao rivers, which are tributaries of the Han River. This gap gives easy access from the Honan Plain to the central basin of the Yangtze, a route much used from Han times onward in Chinese expansion southward.

To the east lie the plains. Until fairly recent geological times the mountains in the west (the western extension of the present Po Hai [Gulf of Chihli] and Yellow Sea) formed the coast of a sea. That sea, now filled with silt brought down by the rivers and by the wind from the Loess Plateau, forms the North China Plain and the Huai Basin. It is part of the great Neo-Cathaysian Geosyncline (downward-sloping part of the Earth's crust), which extends from Heilungkiang to Kiangsi provinces. The floor of this geosyncline is sinking at a rate equal to that of deposition; it is estimated that the sediment of the plain is now 2,800 feet deep in places.

*Drainage.* Honan has three river systems: the Huang Ho in the north and northeast, the Huai River in the east and southeast, and the T'ang and T'ao rivers in the southwest.

The Huang Ho

The Huang Ho—known in Chinese literature simply as the Ho ("River")—immediately after its confluence with the Wei River, at the Shansi provincial border, turns eastward at T'ung-kuan to flow directly across the north of Honan. Near T'ung-kuan it enters the San-men Gorge of some 80 miles, thence issuing onto the plain. It is remarkable that from T'ung-kuan to the sea, a distance of some 600 miles, the Huang Ho receives only two comparatively small tributaries: the right-bank Lo River, on which Lo-yang stands, and the left-bank Ch'in River.

The Ho is subject to very great changes in summer and winter flow. In time of maximum flow (summer) the river carries an enormous load of silt, gathered mainly in its course through the Loess Plateau of Shensi and Shansi provinces. There is a Chinese saying that "if you fall into the Huang Ho you never get clean again." While the river is fast flowing in the San-men Gorge, it is able to carry its load of silt, but when it issues onto the plain its pace is checked. It can no longer carry the silt, and flooding occurs. Throughout historical times this menace has been met by building levees to contain the waters. Generally these dikes were built five to eight miles apart, parallel to the river's banks, to give the river plenty of room in time of spate (flooding), but instead the load of silt has been slowly spread, building up the riverbed through the centuries, until today it lies above the surrounding countryside. Dikes have been built higher and higher, and when they failed to hold—as has happened in some part of the province almost every year—the river descended onto the plain, causing disastrous floods, the waters of which could not return to the high streambed when the river's flow slackened. The result was waterlogging of the soil, crop destruction, and famine. Because the watershed between the Huang and Huai rivers is almost imperceptible, the Huang Ho has radically changed its course several times in the last three millennia, flowing to the sea, first south, then north, of the Shantung Peninsula. The diversion has always been in northern Honan between Cheng-chou and K'ai-feng. In 1938 in an attempt to arrest the advance of the invading Japanese Army, the Huang Ho was deliberately diverted by blowing up the dikes near Cheng-chou and flooding 21,000 square miles of land, at an estimated cost of 900,000 lives. The river was restored to its former northern course in 1947. Under the People's Republic, work along the river has included continued strengthening of the dikes and construction of the 30-mile-long People's Victory Canal, which diverts Huang Ho water to the Wei River. A dam near the city of San-men-hsia near the Shansi border was begun in 1956 as part of an extensive flood-control and hydroelectric project. After completion of the dam in the 1970s, silt accumulation cut its generating capacity to one-quarter of planned output, hampering much of the province's industrial development, but the dam proved valuable in taming the flood stages of the river.

The Huai itself and all its major left-bank tributaries have their sources in the mountains of western Honan. They flow eastward onto the Anhwei Plain, subjecting it to disastrous floods. In 1949 the Huai Basin became the Communist regime's first large water-conservancy program. Six dams were quickly built in the upper reaches of Huai tributaries in Honan. Since 1957 three very large dams at Lung-shan, Mei-shan, and Fo-tzu-ling have been built. Dikes were strengthened, with the result that no serious disaster has since occurred.

Huai Basin water-conservancy program

*Soils.* Honan's soils are made up mainly of calcium carbonate (lime) in hardened layers of alluvium. Because of the comparatively low rainfall, there is little leaching. The higher land of the west is mainly mountain yellow-brown earth, better drained than the plains. The more fertile areas fringing the plain were the sites of early civilization. Alluvium is spread throughout the plain; it is yellowish and gray, porous, granular, and poor in organic matter. Since the bed of the Huang Ho lies above the surrounding land, much of the low-lying land on either side is waterlogged. Consequently, soil salinity and alkalinity affect the whole area. There are large areas of bleak, white saline sands. Since 1949 there has been much experimentation aimed at bringing these alkaline lands into production. Between 1954 and 1964 one-fourth of the saline land between K'ai-feng and Cheng-chou reportedly was transformed into fertile farmland, and reclamation of saline and alkaline land has continued.

Salinity of the Huang Ho flood basin

*Climate.* Climatically, Honan lies in a transitional zone between the North China Plain and the Yangtze Valley. Although protected in some degree from the Mongolian winds by the T'ai-hang Mountains, Honan has very cold winters; summers are hot and humid. Average January temperature in the north is 28° F (−2° C) and in the south 36° F (2° C). Average July temperature over the lowlands is 82° F (28° C), while in the western mountains

it is a degree or two lower. There are 210 frostless days annually in the north and 250 in the south.

Rainfall is distributed more evenly throughout the year than it is in the rest of North China, although there is a marked spring–summer maximum. K'ai-feng has an average rainfall of 23 inches (580 millimetres), of which only three inches fall in the autumn and winter months. There is a steady decrease in total rainfall from southeast to northwest and a marked increase in variability. Honan is therefore more subject to years of alternating heavy rain and drought than the provinces of the Yangtze Valley. In the past it has suffered from severe famine. It also experiences spring cloudbursts and occasional hailstorms, both of which can be very destructive. In times of drought, summer dust storms are worse even than those of winter.

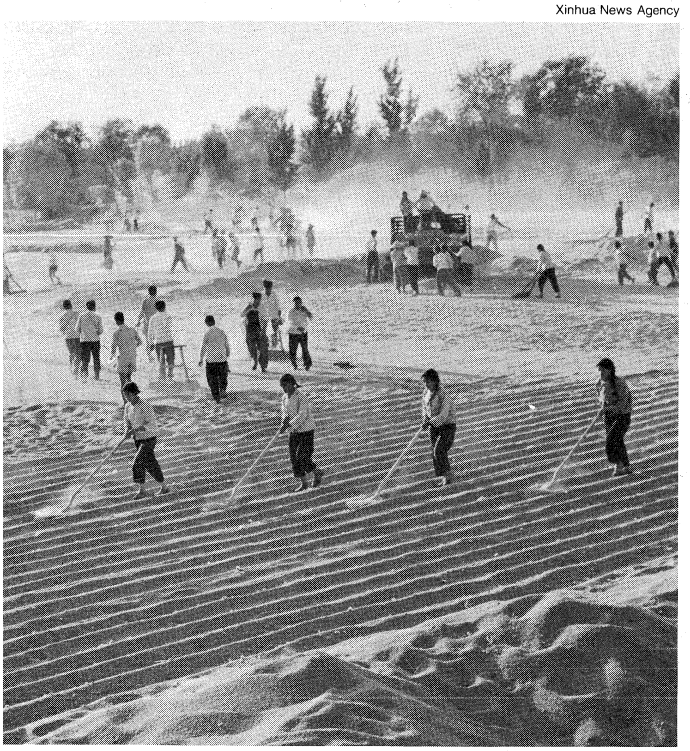
*Plant life.* The natural vegetation of Honan is deciduous forest and woodland over the plains, and deciduous and coniferous forest in the western highlands. Intensive settlement of the plains has long since led to the clearance of the trees to make way for cultivation. The mountains, however, retain some of their woodland. Since 1949 major efforts have been made in planting trees for shelter, timber, and other uses.

**The people.** Honan is China's second most populous province, with the overwhelming majority of the population living in rural areas. The greatest concentration of rural population is in the eastern plain. Nearly as great densities are found in the I and Lo river basins and in the plain around Nan-yang, but in the more mountainous west and south they are considerably less. On the eastern plain, villages are fairly close together, usually about one mile apart. In the mountains they are smaller and more widely dispersed. Houses are made mainly of mud-plastered walls and thatched roofs. There was considerable movement of rural people of the plains to towns in the west in 1958–59, during agricultural collectivization and the Great Leap Forward.

Ethnic homogeneity of Honan

The vast majority of the people of Honan is Han (Chinese). There are no autonomous minority groups such as are found in the western provinces, the small number of Hui (Chinese Muslims) being integrated into the broader population. Mongol and Manchu invaders were absorbed and Sinicized. In the 12th century, when K'ai-feng was the Imperial capital of the Sung dynasty, Jews, originally from India or Persia, became an important part of the community. They retained their identity until the 19th century but have since been absorbed.

Xinhua News Agency



Farm workers spreading harvested wheat to dry on the plain of the Huang Ho (Yellow River) in northern Honan Province.

**The economy.** *Agriculture.* Honan's economy is essentially agricultural. Most of the total cultivated area lies in the plains to the east of the Peking–Han-k'ou railway. The only idle land is found in the mountains and in the saline lands of the northeast. Main food crops include winter wheat, millet, kaoliang (a variety of grain sorghum), soybeans, barley, corn (maize), sweet potatoes, rice, and green lentils. Wheat is by far the most important, in both acreage and production, Honan ranking first in China's output. Rice occupies only a small percentage of the crop area; its yield per acre, however, is almost three times as great as that of wheat. Fruit growing has received considerable impetus in recent years, partly for its own sake and partly for soil conservation, particularly in the idle sandy lands of the northeast and on mountain slopes. Dates, persimmons, apples, and pears are the main fruits, with walnuts and chestnuts also grown. Honan produces draft animals of good quality, particularly yellow oxen and donkeys. Hogs are the most important food animals, and goats and sheep are raised in the western mountains.

The chief industrial crops are cotton, tobacco, vegetable oils, and silk. Cotton is widely grown on about half the acreage, with its main concentration north of the Huang Ho around An-yang and Hsin-hsiang. Tobacco growing, introduced in Honan in 1916, increased enormously after 1949; within 40 years Honan became China's largest tobacco producer. Vegetable oils are important, with Honan one of China's largest producers of sesame, grown mainly in the east and south. Ramie, the most important of the leafy fibres, is grown in east Honan in the Huai Valley. Honan is one of the oldest centres of sericulture (silkworm raising) in China. The industry dates back to the Tung (Eastern) Han dynasty (AD 25–220). Both mulberry-leaf culture and oak-tree culture for silkworms flourished between the two world wars until suffering severely during the Sino-Japanese War. After 1949 there was a revival on the slopes of the Fu-niu Mountains, and the province became an important exporter of silk.

Sericulture

Honan suffers very severely at times from locusts, which winter in the arid, sandy alkaline soils beside the Huang Ho. Extended and improved cultivation in these areas has helped control the pest.

*Industry.* Although before 1949 there was little industrial development in Honan, subsequent industrialization was both rapid and extensive. Much of the development tapped Honan's rich coal seams in the northwest. Both bituminous and anthracite coal are found along the slopes of the T'ai-hang Mountains, and big reserves of good coking coal in thick, easily mined seams are found in the Fu-niu Mountains between Hsü-ch'ang and P'ing-ting-shan. Iron ore is found at Ju-yang on the Ju River in the Hsiung-erh Mountains, as well as some pyrite, bauxite, and mica. Large coal mines at Chiao-tso supply the fast-growing industries of Lo-yang, Cheng-chou, K'ai-feng, and Hsin-hsiang but are still inadequate. The vast coalfield at P'ing-ting-shan has been worked since 1964.

Honan is a significant producer of energy, with thermal plants in Cheng-chou, Lo-yang, K'ai-feng, and Hsin-hsiang linked by a power line. An ultrahigh-voltage transmission system, one of the largest in China, began transporting electricity from the P'ing-ting-shan coal-mining area to Wu-ch'ang in the early 1980s. There are large proven reserves of low-sulfur petroleum and natural gas at the Chung-yüan complex of oil fields.

China's first tractor factory

Lo-yang was chosen as the site for China's first tractor factory, opened in 1958. Since then its output has burgeoned, and Lo-yang has become a heavy-industry centre. Cheng-chou lies in the heart of the cotton-growing area and is now the centre of the textile industry. K'ai-feng, Imperial capital of the Sung emperors, declined after the 11th century—especially when the Huang Ho dikes were broken and the region was ruined in 1642. A large chemical-fertilizer works and a tractor-accessories plant have led to its revival. Hsin-hsiang, the most important city of north Honan, is the centre of the railway network of the area. Emphasis has shifted over the years from development of heavy to development of light industry. Thus,

there has been a growth in the production of consumer goods such as cigarettes, electronic products, bicycles, household appliances, textiles, and tableware. Tourism is a major earner of foreign exchange.

*Transportation.* Although the Huang Ho flows through north Honan, it serves it poorly as a line of communication. Within the province it was navigable only in the Sanmen Gorge until the construction of the dam there. Even now it is useful over the plain only for small rivercraft. The Huai and its tributaries flowing down from the western mountains are rapid in their upper courses and silted in their lower, so that they, too, serve only small craft. The Wei, flowing north into the Hai system, has been joined by the People's Victory Canal to the Huang Ho. In 1964–65 it was successfully dredged in an experiment aimed at deepening the riverbed and so increasing flow and reducing waterlogging. Cheng-chou is the junction of China's two greatest trunk railways, the Peking–Han-k'ou–Canton line and the Lung-hai line, which runs from the east coast to Sinkiang, in the far west. Local railroads have also been developed, and most of the province's goods are now carried by rail. The first modern roads in Honan date from the famine of 1920–21, when the American Red Cross built earth tracks to bring relief to the stricken provinces. Since 1959 the great bulk of road building has been done with little modern technology; some roads penetrate the more remote mountain region, as, for example, a road in the T'ai-hang Mountains between Hui-hsien and Ling-ch'uan. Most of the highways have all-weather surfaces.

**Administration and social conditions.** *Government.* On the victory of the Communists in 1949. Honan, together with Hupeh, Hunan, Kiangsi, Kwangtung, and Kwangsi, formed the Central South greater administrative region. In 1954 provincial government was established, and for local governmental purposes Honan has been subsequently divided into eight prefectures (*ti-ch'ü*) and nine prefecture-level municipalities (*shih*). Below this level the province is divided into counties (*hsien*) and county-level municipalities (*shih*). The province was badly affected by political conflict during the Cultural Revolution. During much of that time it was governed by a provincial Revolutionary Committee, which consisted of 11 members, nine of whom were selected from the People's Liberation Army, one from the "revolutionary cadres," and one from the "revolutionary masses." The Revolutionary Committee was replaced in 1980 by the People's Government, which is the administrative arm of the People's Congress. The People's Congress, acting largely through its Standing Committee, is an organ of the state, and its powers include enacting legislation, implementing state policies, and approving provincial economic plans and budgets. Its members are elected by the People's Congress at the next lower administrative level, and it in turn elects the members of the People's Government.

*Education.* Since its Imperial days K'ai-feng has remained the cultural and educational centre of Honan, although it has come to share that role with Cheng-chou. The first impact of Western learning came, as in the rest of China, through the primary and middle schools of Christian missions. Little real progress was made in the turbulent years between 1911 and 1949, and the vast mass of the people remained illiterate. Successful efforts were made by the government in the first years after 1949 to overcome illiteracy, and a real attempt at universal primary education was launched. Education is now based on six years of primary schooling and six years of secondary schooling. Honan has more than 30 institutes of higher learning.

*Health and welfare.* Modern Western medicine, like education, was introduced by Christian missions but made very little impact on the vast area of Honan. On attaining power in 1949, the People's government concentrated attention on public hygiene and preventive medicine. A doctor's training was cut from six to three years, and teams were dispersed throughout the province to teach hygiene, vaccinate, inoculate, and advise. Although a doctor's training has now reverted to six years, emphasis remains on the "barefoot doctor," midwife, and health worker. With the great development of coal mining in Honan, attention has been focused on silicosis prevention, which is being achieved mainly by improving working conditions. Kala-azar, the debilitating disease carried by sand flies, is also receiving special attention. As elsewhere in the country, traditional Chinese medicine has gained status.

Emphasis on the "barefoot doctor"

**Cultural life.** In an essentially agricultural province such as Honan, cultural life is centred in the rural community. While families, for the most part, retain and own their own homes, cultural life is focused in the community centre, with its reading room, library, and teahouses, in which the old tradition of storytelling has continued and is very popular. Loudspeaker radio is used to ensure communication. Traditional local music forms—such as the *chui-tzu* ballad and *yu-chü* opera—are popular, as are the performances of the province's many art troupes.

HISTORY

Honan abounds in prehistorical and early historical interest. Some of the most important evidences of the Neolithic beginnings of Chinese civilization are found in the northern part of the province. It was at Yang-shao in north Honan that a Swedish geologist and archaeologist, Johan Gunnar Andersson, in 1921 discovered an assemblage of Neolithic painted pottery that, together with many later finds, marked the presence of a well-established primitive farming culture, which has been named Yang-shao. The early farmers occupied the lands at the confluence of the Huang, Wei, and Fen rivers, the cradle of Chinese civilization. The other main Honan sites of the culture are at Miao-ti-kou, Lin-shan-chai, P'an Nan, and Hsi Yin. The early farmers, who were also part-time hunters and fishermen, lived in sunken circular or rectangular dwellings, sometimes of considerable dimensions. They grew foxtail millet, broomcorn millet, and kaoliang and had domesticated dogs and pigs. Cultivation with their primitive stone tools was comparatively easy in the easily worked loess (wind-borne) soil.

Cradle of Chinese civilization

Immediately to the east, at Lung-shan in Shantung Province, a different culture was discovered, known as the Black Pottery culture, as distinct from the slightly earlier Painted Pottery culture (Yang-shao culture). It was on these Yang-shao–Lung-shan foundations that the early civilization of the Shang (Yin) dynasty arose (18th–12th century BC) in north and west Honan, south Hopeh, and west Shantung. Excavations near An-yang and in Cheng-chou and Hsing-t'ai, Hopeh, revealed an advanced culture, having a hierarchical class structure, advanced buildings, and elaborate ritual in which beautiful bronze vessels were used. Based on the dating of oracle bone inscriptions, the Shang king P'an K'eng moved his capital to a site near An-yang in 1384 BC.

When the Shang kingdom fell to the Chou dynasty (1111–255 BC), An-yang lost its status as a capital. When the Chou capital, Hao (near modern Sian in Shensi Province), was destroyed in 771 BC by western tribes, Lo-yang (then known as Lo-i) took its place. During the period 771 BC to AD 938, the distinction of being the capital was shared alternately by Lo-yang and Ch'ang-an (modern Sian). Lo-yang was the capital during the following dynasties—the Tung (Eastern) Chou (771–256/255 BC), Tung Han (AD 25–220), Wei (220–265/266), Hsi (Western) Chin (265–311), Wei (386–534/535), and Hou (Later) T'ang (923–936/937). With the fall of the T'ang dynasty in 936/937, K'ai-feng, then called Pien, became the nation's capital and remained so until the Pei (Northern) Sung dynasty was overthrown by the Juchen invaders in 1126. After the sack of K'ai-feng in 1127, the Honan region continued to be the chief source of grain for Imperial storehouses. Both Lo-yang and K'ai-feng remained important because of their strategic locations in the gateway leading from the North China Plain into the Huai Basin, thence into the Yangtze Basin. Cheng-chou became important in the early 20th century as a railway junction and was made the provincial capital in 1954. (T.R.T./V.C.F.)

## Hopeh

Hopeh (Ho-pei in Wade–Giles romanization, Hebei in Pinyin) is a province in northern China on the Po Hai

(Gulf of Chihli) of the Yellow Sea. It is bounded on the northwest by China's Inner Mongolia Autonomous Region and by the provinces of Liaoning on the northeast, Shantung on the southeast, Honan on the south, and Shansi on the west. Hopeh means "North of the (Yellow) River." Hopeh has an area of 72,500 square miles (187,700 square kilometres). The provincial capital was at Paoting until 1958, when it was transferred first to Tientsin and then, in 1967, to Shih-chia-chuang, 160 miles (260 kilometres) southwest of Peking. The present capital is at the junction of three railways: the Peking–Canton line, China's north–south trunk line, and lines to Shansi and to Shantung. The large municipalities of Peking, the national capital, and of Tientsin lie within Hopeh Province but are independent of the provincial administration. Culturally and economically, Hopeh is the most advanced province in northern China.

## PHYSICAL AND HUMAN GEOGRAPHY

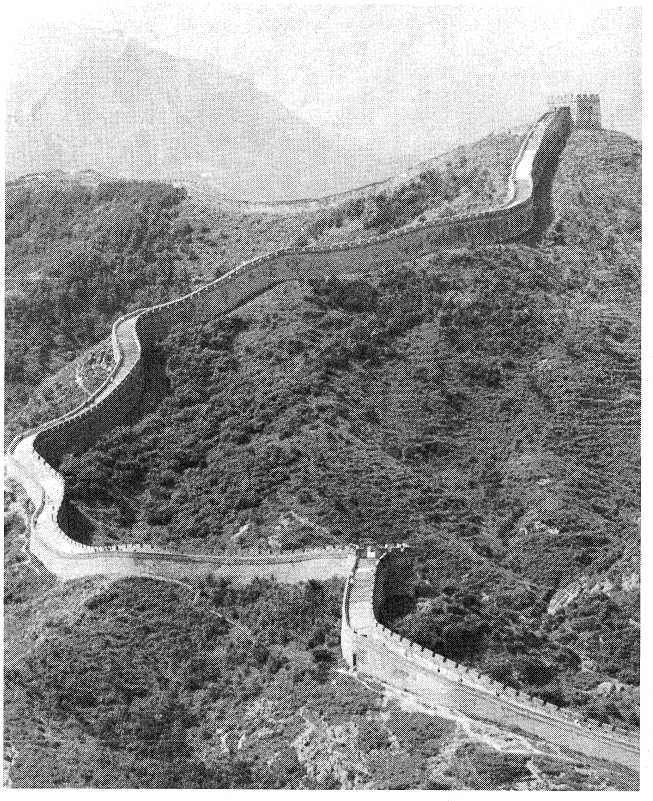
**The land.** *Relief.* Hopeh Province consists of two almost equal sections: the northern part of the North China Plain and the mountain ranges along the northern and western frontiers. The former is sometimes called the Hopeh Plain. It is formed largely by the alluvial deposits of the five principal tributaries of the Hai River, which flows past Tientsin to the sea. Two of them, the Yungting and the Pai, flow down from the northern highlands. The other three have their sources in the western part of Hopeh: the Ta-ch'ing and Tzu-ya rivers and the Southern Grand Canal (Nan-yün Ho).

Hopeh Plain

The Hopeh Plain slopes gently from west to east. It is bounded by the Yen Mountains on the north, the T'ai-hang Mountains to the west, and the Po Hai to the east. The mountains have at their base a string of alluvial fans. This inner belt of the Hopeh Plain is generally well drained. The groundwater level is usually less than 33 feet (10 metres) from the surface and is easily tapped for domestic water and irrigation.

The Yen Mountains form the northern rim of the North China Plain, displaying to the traveler an endless sea of rounded hills, with peaks averaging 4,900 feet above sea level. The Great Wall of China zigzags along its crests.

Ellen Warner—Black Star



The Great Wall of China as it traverses the crests of the Yen Mountains in Hopeh Province.

Beyond these mountains the Mongolian Plateau stretches from the northernmost part of Hopeh Province to the Mongolian People's Republic. This part of Hopeh was incorporated into the province in 1952, when Hopeh's boundaries were extended beyond the North China Plain for the first time. The rim of the plateau has an average elevation of 3,900 to 4,900 feet and is rugged and inhospitable to human settlement. Between the Yen Mountains are large basin plains, cultivated and well inhabited. Coal and iron are mined in the northern mountains.

To the west of the North China Plain sprawls the lofty north–south range of the T'ai-hang Mountains, separating the Hopeh Plain from the Shansi Plateau, its highest peak rising more than 9,000 feet. The range is pierced by a number of west–east streams whose narrow valleys (the famous "Eight Gorges" of T'ai-hang) are the routes of highways and railroads between the Hopeh Plain and the Shansi Plateau.

*Drainage and soils.* The major Hopeh rivers flow down from the loess-covered T'ai-hang Mountains and the Shansi Plateau. They carry a heavy load of silt after the summer downpours, depositing it in the shallow channels downstream on the plain, gradually silting them up and causing widespread floods in low-lying areas. Since 1949 vigorous measures for water control and soil conservation have been carried out together with reforestation in the upland areas. Numerous dams, generally small to medium-size, have been built upstream and in the tributaries to conserve the water for irrigation and other uses; flood-retention basins and storage reservoirs have been built downstream. The Tu-liu-chien River, connecting the Ta-ch'ing to the sea, helps to drain the extremely low-lying tract around the large Pai-yang Lake and the Wen-an Marsh. Water from the streams is used to wash away excess salt in the alkaline soil and to make it arable. Similar *chien-ho* ("reducing streams") have been completed for the Southern Grand Canal.

Rivers

The Hai River is only 35 miles long, from the city of Tientsin to the sea, but the drainage basin of its five tributaries covers two-thirds of the province. A number of flood-control and power-generation projects have been developed in the Hai Basin, including reservoirs to the northeast and northwest of Peking. Another major river is the Luan, which drains northeastern Hopeh. A major project of the 1980s was the construction of a diversion channel carrying water from the Luan to Tientsin. All the major Hopeh rivers empty into the Po Hai, a shallow sea with an average depth of only 100 feet. The water and nutrient matter brought down by the rivers nourish a rich marine fauna. In winter the surface water along the coast is frozen, but navigation is possible with the use of ice-breakers. There are three important ports: Tientsin, which is about 35 miles up the Hai, T'ang-ku, and the major coal-handling and oil-shipping port of Ch'in-huang-tao.

The most common soil in the Hopeh Plain is dark-brown earth developed on loessial alluvium, modified by cultivation over several millennia. It is extremely fertile—the famous "good earth"—yielding crops with little fertilization for thousands of years. New alluvium is distributed in the areas along the rivers by frequent flooding. In the mountains the soils vary: the upland hills have leached dark-brown soils, the more humid mountainous areas of the Yen and T'ai-hang ranges have brown forest soils suited to fruit trees, and the northernmost Chang-pei plateau has light-chestnut zonal soils.

*Climate.* The province has a continental climate. The January mean temperatures range from 25° F (−4° C) in the south to 14° F (−10° C) north of the Great Wall. The average July temperature is about 77° F (25° C) in the North China Plain, and 73° to 77° F (23° to 25° C) in the northern and western highlands. The annual precipitation (rain and snow) is more than 20 inches (500 millimetres) in most parts of the province. The summer months of June, July, and August are the rainy season.

*Plant and animal life.* The natural vegetation of the greater part of the province is broad-leaved deciduous forest, but, after many centuries of human settlement, cultivation, and deforestation, little of the original vegetation remains except in the high mountains and other inac-

cessible areas. Annual afforestation projects have seeded millions of acres in an effort to develop the forest upland economy.

The northernmost Chang-pei plateau has steppe grass of the Mongolian Plateau type. The higher mountains have coniferous forests. In the saline areas along the coast and in the low-lying depressions, plants that flourish in a salty environment dominate. There is a conspicuous absence of forests in the lowlands and lower hills. The flora is predominantly of a northern character. It includes the willow, elm, poplar, Chinese scholar tree (*Sophora japonica*), tree of heaven (*Ailanthus*), and drought-resistant shrubs.

Animal life

The present fauna includes elements of the temperate forest (such as the forest cat *Felis euptilus*) and of the cold-winter steppe (such as the camel), as well as some tropical elements from the Indo-Malay region (such as the tiger and monkey). The domestication of animals such as the dog, sheep, goat, cow, horse, donkey, mule, camel, and cat has led to the extinction or near-extinction of many wild species. The smaller mammals are better preserved, including moles, bats, rabbits and hares, rats, mice, and squirrels. Birds include the Mandarin duck (*Aix galericulata*), native to China. The Hopeh Plain was the home of Peking man, an extinct hominid of the species *Homo erectus,* who lived about 460,000 years ago and used tools and fire.

**The people.** The ethnic composition of the population is almost entirely Han (Chinese). Minority groups include the Hui (Chinese Muslims) and a tiny percentage of Mongols. Since nearly one-half of Hopeh Province is mountainous, the density of population is really much higher than the average of about 750 persons per square mile (300 per square kilometre) suggests. The highest population densities in Hopeh are found at the foot of the T'ai-hang Mountains, in the belt of alluvial fans. This is a district settled since antiquity, on the ancient highway from the Chung-yüan, or "Middle Plain," of the North China Plain to Peking and on to the regions north of the Great Wall. These piedmont plains have also been settled since ancient times. The rural settlement pattern is that of huge nucleated villages. Farther east and south of the alluvial-fan belt are the low-lying districts subject to flood, which have somewhat lower densities. The area north of the Great Wall and the remote mountainous areas have the lowest densities. Before 1949 there was substantial migration from northwestern Hopeh to Inner Mongolia. Peasants in southeastern Hopeh have also migrated in large numbers since the beginning of the 20th century to Inner Mongolia and China's northwest and the Northeast.

**The economy.** *Agriculture.* Hopeh is important for its production of cotton, wheat, corn (maize), peanuts (groundnuts), and fruit. The widespread introduction of tube-well irrigation in the late 1960s and early '70s made Hopeh and Kiangsu among the leading provinces in irrigated acreage.

*Industry.* Hopeh lies at the heart of one of two major industrial regions in China. The province developed a modest industrial base from the late 19th century onward, chiefly in coal, iron, textiles, and indigenous handicrafts. Tremendous industrial expansion took place during the 1950s: the spinning capacity of Hopeh's cotton belt was expanded considerably; a major coal belt, which stretches in a crescent through Hopeh and into northern Honan, provided the impetus for significant expansion of the coal-mining industry; and the incorporation into Hopeh (1952) of the Lung-yen iron-ore district of former Chahar Province speeded the development of the iron and steel industry. In the 1960s the emergence of the Hua-pei oil fields made Hopeh a major oil producer, and in 1983 China's first deep-horizon oil field, the Ma-hsi field, went into operation in the southern section of the Ta-kang oil field on the Po Hai coast, producing significant quantities of oil and natural gas.

Peking–Tientsin industrial region

These industries became the basis of the Peking–Tientsin industrial region, the largest and most important industrial centre in North China. Industrial production has diversified and expanded to include such key products as cement, agricultural equipment, and fertilizer. Light industries include textile and ceramics manufacture, food processing, and paper and flour milling. Tientsin, the region's second largest city, is the primary industrial and commercial centre of North China and the second most important trade centre in all China. Other major industrial cities of the region include T'ang-shan (largely rebuilt since an earthquake in 1976) and Ch'in-huang-tao, in eastern Hopeh; Pao-ting; Shih-chia-chuang, in western Hopeh; and Liu-li-ho, in Peking Municipality.

*Transportation.* Hopeh is well served by railroads. The province is at the centre of China's vast north–south railway network, and all of its major cities are connected by rail. Sea transport moves through Tientsin and Ch'in-huang-tao. The port of Ch'in-huang-tao was first opened to commercial activity in 1898; it is now one of China's "open" coastal cities, which play a key role in the country's foreign trade and investment. The port ranks third nationally after Lü-ta (formerly Dairen) and Shanghai in handling capacity.

**Administration and social conditions.** *Government.* Hopeh Province (*sheng*) is divided into nine prefectures (*ti-ch'ü*) and nine prefecture-level municipalities (*shih*). Below this level the province is divided into counties (*hsien*) and county-level municipalities (*shih*). The traditional subcounty administrative unit was the civil township, or rural district (*hsiang*), which was supplanted in 1958 by the commune. The communes were in turn replaced by the *hsiang* after the Cultural Revolution.

Administrative levels

*Education.* Public education has made major strides since 1949. The great majority of adults are now literate, and well over half of the population has received at least a primary school education. With more than 30 institutions of higher education, the province has sought to upgrade the technical level of its citizens as part of a drive toward modernization. The emphasis on broadening opportunities for education has led to the establishment of television and radio universities for part-time and continuing study, while vocational secondary schools serve the needs of Hopeh's industry.

**Cultural life.** Hopeh is linguistically and culturally part of the Northern Mandarin dialect area and shares many of the features of that regional culture. Living in the northernmost part of the Sinitic zone—historically subject to nomad incursions and political subjugation—Hopeh's people are traditionally depicted as orderly, submissive, and uncomplaining. Their cuisine features wheat cakes, mutton, and bean dishes. There are many local operatic and dramatic traditions, carried on by the province's numerous art and theatre troupes.

## HISTORY

Although the area of present Hopeh Province was settled very early, it lay for many centuries outside the sphere of most political and economic activity of the Chinese empire. Before incorporation into the Ch'in Empire in the 3rd century BC, the region was occupied by the states of Yen, Ch'i, and Chao.

Hopeh has long been an area of strategic significance. To the rulers of the Han dynasty (206 BC–AD 220), it was largely a frontier zone beyond which lay their main enemies, the Hsiung-nu people, and defense of the region with walls and permanent garrisons was therefore emphasized. To the expansionist emperors of the T'ang dynasty, Hopeh served as a starting point for large campaigns aimed at the conquest of Korea. In AD 755, military forces stationed in the area were used to temporarily overthrow T'ang rule in a revolt led by An Lu-shan. Hopeh grew in importance under the rule of a series of northern-based dynasties, including the Liao, or Khitan (907–1125); the Chin, or Juchen (1115–1234); and the Yüan, or Mongol (1206–1368). Peking first became the capital of all China under the Yüan rulers, who also completed work begun by the Chin on the Grand Canal linking Hopeh to the rice-growing regions of southern China.

Hopeh under T'ang rule

During the Ch'ing, or Manchu, dynasty (1644–1911/12) Hopeh was called Chihli ("Directly Ruled") Province and continued to be strategically important, especially as foreign imperialist pressure mounted during the 19th century. Li Huang-chang, the foremost military and political leader of his time, served for many years as governor general of

Chihli and was succeeded by Yüan Shih-k'ai, who became president of the Chinese republic in 1912. A period of domination by a succession of autonomous warlords in Hopeh followed Yüan's death in 1916. The warlord Yen Hsi-shan continued to govern independently in Hopeh until the Japanese invasion of 1937. After Japan's defeat the occupiers surrendered to the Chinese Nationalists in 1945. Chinese Communist forces took the province in January 1949, opening a new chapter in its long history.

(F.Hu./V.C.F./Ed.)

## Peking

For coverage of Peking, see the *Macropædia* article PEKING.

## Shansi

Shansi (Shan-hsi in Wade–Giles romanization, Shanxi in Pinyin), a province of northern China, has an area of about 60,200 square miles (156,000 square kilometres). Roughly rectangular in shape, Shansi is bounded by the provinces of Hopeh to the east, Honan to the south and southeast, and Shensi to the west and by the Inner Mongolia Autonomous Region to the north. The name Shansi ("Western Mountains"), testifies to the rugged terrain of the territory. The largest city and provincial capital, T'ai-yüan, is located in the centre of the province.

Shansi has always held a strategic position as a gateway to the fertile plains of Hopeh and Honan. Since ancient times it has also served as a buffer between China and the Mongolian and Central Asian steppes. A key route for military and trading expeditions, it was one of the major avenues for the entrance of Buddhism into China from India. Today it is important for its vast reserves of coal and iron, which form the basis of heavy industrial development, and for its production of cotton for export.

### PHYSICAL AND HUMAN GEOGRAPHY

**The land.** *Relief.* Two-thirds of the province is composed of a plateau, part of China's vast Loess Plateau, that lies at elevations between 1,000 and 3,000 feet (300 and 900 metres) above sea level. The plateau is bounded by the Wu-t'ai and Heng Mountains on the north, the T'ai-hang Mountains on the east, and the Lü-liang Mountains on the west. The eastern mountains average between 5,000 and 6,000 feet in height and reach their maximum elevation at Mount Shih-ku (8,300 feet), located in Hopeh Province. The highest peak in the west, Mount Kuan-ti, reaches an elevation of 9,288 feet, while the northern ranges are crowned by Mount Wu-t'ai at 10,033 feet.

Mountains and plateaus

Wang Wenxue—Xinhua News Agency



The white pagoda of T'a-yüan Temple in the Wu-t'ai Mountains, Shansi Province.

The Huang Ho (Yellow River) flows through a mountain gorge from north to south and forms the western border with Shensi Province. At Feng-ling-tu the river turns sharply eastward and forms part of the southern border with Honan Province. The southwest corner of the province is part of the highland region that extends from Kansu to Honan provinces and is covered with a layer of loess. The Fen River Valley comprises a chain of linked, loess-filled basins that crosses the plateau from northeast to southwest. The largest of the valley's basins is the 100-mile-long T'ai-yüan Basin. North of T'ai-yüan are three detached basins, which are areas of cultivation. Farther north the Ta-t'ung Basin forms a separate feature.

*Drainage and soils.* Several rivers drain eastward and southeastward, cutting valleys and ravines through the T'ai-hang and Wu-t'ai ranges, including the Hu-t'o and its tributaries. In the west several rivers cut across the Lü-liang Mountains and drain into the Huang Ho; principal among these is the Fen, which flows southward through two-thirds of the province. The northern mountains are drained chiefly by the Sang-kan, which flows eastward.

In the mountains, several types of light-brown and brown forest soils are common, with meadow-steppe varieties found at higher elevations. Alluvial soils in the central and southern portions of the province are formed mainly of calcareous (lime-bearing) brown soils deposited by the Fen River. There are also loess and lime deposits. Natural organic materials are not abundant, and salinity is excessive.

*Climate.* Shansi has a semiarid climate. The mean annual rainfall ranges from less than 10 inches (250 millimetres) a year in the northwest to a maximum of 20 inches in the southeast. Between 70 and 80 percent of the annual rainfall occurs between June and September. Temperatures range from a January mean of 19° F (−7° C) and a July mean of 75° F (24° C) at T'ai-yüan, to a January mean of 3° F (−16° C) and a July mean of 72° F (22° C) at Ta-t'ung. Winter droughts are common because the plateau is subject to the full force of the dry northwestern wind that blows in the winter from the Mongolian Plateau. In summer the southeastern monsoon (a rain-bearing wind) is blocked by the T'ai-hang Mountains. Hailstones are a common natural hazard, as are frequent floods, particularly along the course of the Fen.

*Plant and animal life.* Vegetation distribution primarily depends on the direction in which the mountain slopes face. The southern slopes are characteristically covered by species such as oak, pine, buckthorn, and honey locust, which are more tolerant of drier conditions than are the linden, hazel, maple, and ash that prevail on the more humid northern slopes. The province has long been cultivated, and such natural vegetation as remains consists mainly of shrubs and grasses; isolated forests occur on the north-facing slopes. Destruction of the original forest cover in ancient times eliminated most animal species.

**The people.** Most of the province's people are of Han (Chinese) origin and speak the Northern Mandarin dialect of Chinese. The small minority populations include the Hui (Chinese Muslims), in the T'ai-yüan–Yü-tz'u region, and some Mongols and Manchu around Ta-t'ung. Most of the populace lives in agricultural villages. The highest rural densities occur in the T'ai-yüan Basin, in the southeast around Ch'ang-chih, and in the Fen Valley.

Cities

The two principal urban areas are T'ai-yüan, the capital and leading industrial and mining complex, and Ta-t'ung, a mining and rail transport centre. Other manufacturing and transport centres include Yü-tz'u and Yang-ch'üan, both east of T'ai-yüan, and Ch'ang-chih in the southeast. Smaller cities are Ch'ü-wo (Hou-ma) and Lin-fen, both situated in the fertile Fen Valley; Fen-yang, immediately southwest of T'ai-yüan; and Yün-ch'eng, on the Hsieh Ch'ih salt lake in the southwest.

**The economy.** *Resources.* Shansi is China's major coal

region, producing one-quarter of the country's output. Proven reserves of anthracite and high-grade coking coal have supported the development of heavy industry and thermal generation of electricity. Iron ore is mined from vast deposits in the Ma-an Mountains district of central Shansi. The largest titanium and vanadium (metallic elements used in alloys such as steel) deposits in China are located near Fen-hsi. Other mined minerals include silver, zinc, copper, and edible salt.

*Agriculture.* Because of widespread erosion, only about one-third of the province is under cultivation. Extensive soil and water conservation efforts since 1949 have taken the form of terracing, afforestation, the digging of irrigation canals, diking of cultivated plots, soil desalinization, and land reclamation along rivers.

Agricultural production

In the extreme north the short growing season and long, cold winter limit cultivation to one annual crop of spiked millet, spring wheat, naked oats (oats with no covering on the kernels), potatoes, and sesame. In the rest of the province—except for the mountainous areas—the longer growing season permits three crops in two years or two crops in one year. Winter wheat, millet, soybeans, kaoliang (a variety of grain sorghum), corn (maize), and cotton are raised in adequately irrigated areas. Some tobacco and peanuts (groundnuts) as well as some fruits are produced in the central basins and on the Huang Ho floodplain.

Only a small part of Shansi's cultivated acreage is devoted to cash crops, such as cotton and sesame, the latter grown both for its oil seeds and for its fibre. Other cash crops include castor beans, rapeseed, and Indian hemp.

The relatively low ratio of population to land over much of Shansi's hilly terrain has traditionally fostered animal husbandry. Sheep are raised for their high wool yields. Domestic animals include pigs, horses, yellow oxen (for transport), donkeys, and chickens.

*Industry.* Most of the province's industries are concentrated in the T'ai-yüan–Yü-tz'u region. The iron and steel industry produces ingot steel, pig iron, and finished steel products. Heavy machinery, industrial chemicals, and chemical fertilizers are produced, as are cement, paper, textiles, milled flour, and wine. Other iron and steel centres include Yang-ch'üan, Ch'ang-chih, Ta-t'ung (which also produces cement and mining machinery), and Lin-fen.

*Transportation.* Shansi relies heavily on rail lines, both for intraprovince transport and for shipping raw materials, industrial commodities, and foodstuffs outside the province. The longest of these, the T'ung-p'u trunk line, runs from Ta-t'ung to Feng-ling-tu, in the southwest corner of the province. Additional branch lines connect the main line with newly opened industrial and mining sites. Major efforts have been made to relieve the pressure on Shansi's railways from ever-increasing freight volume and limited coal transport capability. Rail lines have been double-tracked and electrified, and trunk and spur lines have been constructed.

Long-distance, all-weather roads have been extended, especially near coal mines; many roads serve as feeder routes to the rail lines. The Fen River is navigable for small flat-bottomed boats as far north as Lin-fen. Freight traffic on the Fen, as well as on the north–south section of the Huang Ho, is insignificant, however.

**Administration and social conditions.** The chief provincial administrative body from 1967 to 1980 was the Shansi Provincial Revolutionary Committee. It was replaced in 1980 by the People's Government, which is the administrative arm of the People's Congress. The province (*sheng*) is divided into four prefecture-level municipalities (*shih*) and seven prefectures (*ti-ch'ü*). At the next lower level there are counties (*hsien*) and county-level municipalities (*shih*). An Office for Planning the Energy Resource Base of Shansi was established in 1982.

Education

The educational and medical institutions that were established in Shansi, mainly through foreign initiative, between 1898 and 1910 played a minor role in ameliorating the widespread poverty, illiteracy, and substandard health conditions that then prevailed. Shansi University, founded in T'ai-yüan by an English missionary in 1902, was one of the first in China to offer Western curricula in liberal arts, law, and medicine. Since 1949 technical schools for agriculture, mining, forestry, and machine technology have been established, as have universities, colleges, senior middle schools, and primary schools. The medical colleges and affiliated hospitals in T'ai-yüan offer treatment and full courses of study in both Western and traditional Chinese medicine.

Public works projects include a centralized water supply system based at Lan-ts'un that regulates the flow of the groundwater supply of the T'ai-yüan Basin, modernized sewerage and waste disposal facilities in the major cities, housing projects, and extensive "green belt" areas that are planted with thousands of trees.

**Cultural life.** Shansi's long-standing position as an avenue of communication between the North China Plain, the Mongolian steppes, and Central Asia gave rise to a rich and varied cultural and folkloric tradition. Several distinctive forms of Shansi opera became popular under the Ming and Ch'ing dynasties. Metalworking has been a specialty of Shansi craftsmen since the 2nd millennium BC. The province was also famous for the uniquely sculpted decorative tiles and glazed pottery figures used for temple decoration. The Chin-tz'u, near T'ai-yüan, is Shansi's best known temple complex; it was originally built in the 5th century AD. During subsequent periods it served as a monastery and as the centre for several religious cults.

Ancient handicrafts

## HISTORY

Pollen analyses from western and southern Shansi reveal that several cereal plants were grown there as early as the 5th to the 3rd millennium BC. During the Hsi (Western) Chou period (1111–771 BC) the fief of Chin (now a colloquial and literary name for Shansi) was established in the area of modern Ch'ü-wo (Hou-ma) along the Fen River in the southwest.

Under the Han dynasty (206 BC–AD 220) Shansi assumed what was to become its traditional role as a buffer state between the pastoral nomads to the north and west and the sedentary Chinese farmers to the south and east. A predilection for political autonomy was paralleled by a commercial aggressiveness that led to the rise in the 18th and 19th centuries of a class of Shansi bankers and merchants famous throughout China.

From the end of the Han dynasty until the reunification of the empire under the Sui dynasty in 581, Shansi came under the dominance of several short-lived dynasties, most prominent of which was the Wei dynasty (AD 386–534/535) of the Pei-ch'ao (Northern Dynasties). Buddhism prospered for the first time during the Wei period; it was from Shansi that the Chinese Buddhist monk Fa-hsien began his legendary journey to India. The Buddhist cave sculptures dating from this period and preserved at Yün-kang today constitute some of China's most precious art treasures.

From the 7th century until the end of the 14th century, control over the area shifted back and forth among local military leaders, invading Turkic and Mongol forces, and representatives of the Chinese dynasty in power. Some stability was restored during the Ming dynasty (1368–1644).

Antiforeign feeling ran high during the latter years of the Ch'ing (Manchu) dynasty (1644–1911/12), despite the fact that there was relatively little foreign influence in the province. A few manufacturing establishments were set up in T'ai-yüan in 1898, and a French- and Chinese-financed railway between T'ai-yüan and Shih-chia-chuang in western Hopeh was built from 1904 to 1907. In 1900 antiforeign feeling took a violent form when an English mission church in T'ai-yüan was burned by the I-ho ch'uan (a secret society that came to be known as the "Boxers"), and foreigners and Chinese Christian converts were killed. This led to the outbreak of what became known as the Boxer Rebellion, which eventually spread to Peking.

Boxer Rebellion

After the overthrow of the Ch'ing dynasty in 1911/12, the Shansi warlord Yen Hsi-shan (1883–1960) ruled as an absolute dictator until the end of World War II. Yen was instrumental in establishing the nucleus of a heavy industrial base and in opening the T'ung-p'u railway in 1934.

During the Sino-Japanese War of 1937 to 1945, the Japanese developed coal resources in the T'ai-yüan Basin and expanded heavy industry. They were, however, con-

tinually harassed by Communist guerrillas who operated from mountain bases. The agricultural and handicrafts cooperatives established at these bases were instrumental in facilitating economic and social recovery after Communist forces assumed control of Shansi in 1949. (B.Bo.)

## Shantung

Shantung (Shan-tung in Wade–Giles romanization, Shandong in Pinyin) is a north coastal province of China across the Yellow Sea from Korea. It has an area of 59,200 square miles (153,300 square kilometres). Shantung is China's third most populous province, its population exceeded only by that of Szechwan and Honan. The name Shantung means "Eastern Mountains" and was first officially used during the Chin dynasty in the 12th century.

The province consists of two distinct segments. The first is an inland zone bounded by the provinces of Hopeh to the north and west, Honan to the southwest, and Anhwei and Kiangsu to the south. The second is the Shantung Peninsula extending some 200 miles (320 kilometres) seaward from the Wei and Chiao-lai river plains, with the Po Hai (Gulf of Chihli) to the north and the Yellow Sea to the south, giving Shantung a coastline of 750 miles.

The inland zone, covering roughly two-thirds of the province's total area, includes a hilly central region, centred on the famous Mount T'ai complex, and a fertile and intensively farmed agricultural area on the north, west, and south, which forms part of the Huang Ho Basin and the North China Plain. The provincial capital, Chi-nan, is situated just west of Mount T'ai and three miles south of the Huang Ho, which flows from southwest to northeast through the province before emptying into the Po Hai.

Shantung Peninsula

The Shantung Peninsula, in contrast, is entirely an upland area and, with its seaward orientation and indented coastline, has traditionally depended on fishing, mining, and port-related activities. Long a focal area in the evolution of Chinese civilization and institutions, the province's natural inland–peninsular division is paralleled by a dual orientation in its past and present political and economic configurations. The eastern peninsula historically has coveted autonomy, whereas the inland portion has been closely tied to the inward-facing empire.

### PHYSICAL AND HUMAN GEOGRAPHY

The hill regions

**The land.** *Relief.* Shantung is dominated by two hill masses to the east-northeast of the Grand Canal and to the south-southwest of the present course of the Huang Ho. These hills are formed mainly of ancient crystalline shales and sedimentary rocks on their flanks and of hard, very ancient rocks with granitic intrusions in their core. Both masses are detached remnants of China's most ancient geologic core. The easternmost (peninsular) mass is connected to the Liaotung Peninsula by a submerged ridge, emerging periodically in the Po Hai as the Ch'ang-shan Archipelago. In fairly recent geologic times, the Shantung hill masses stood as islands in an inland sea that separated them from the T'ai-hang Mountains of Shansi to the west.

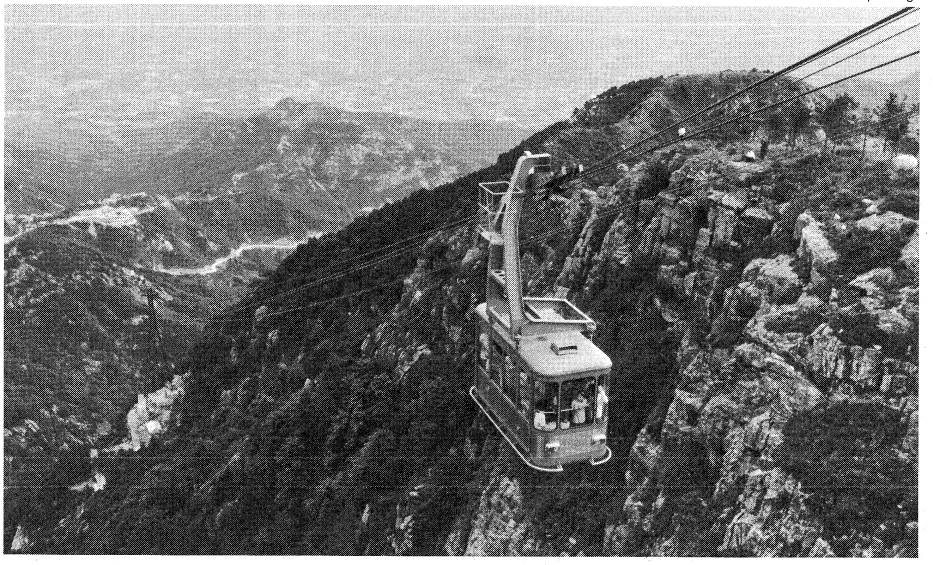
A broad, marshy depression, the Chiao-lai Plain (sometimes known as Wei-hsien Valley), extends for about 100 miles from Lai-chou Bay in the Po Hai, south to Chiao-chou Bay in the Yellow Sea, near Tsingtao, and westward into the North China Plain. The generally flat surface of the plain is interrupted occasionally by bedrock-derived monadnocks, or residual rocks or hills, that have resisted erosion. Another depression, part of the inland zone of western Shantung, forms the central segment of the North China Plain. It slopes eastward into a northwest–southeast trough skirting the western perimeter of the central Shantung hill mass and is filled with a mixture of loess and alluvial materials (sand, clay, and gravel), along with more recently deposited alluvium, resulting from the building up of the Huang Ho floodplain. Five narrow lakes forming part of the Grand Canal system stretch out along this depression and are also linked to a series of saline marshes (indicative of earlier swamp conditions) that separate the fertile margin at the western edge of the central hills from the main sections of the North China Plain to the south and west.

Of the two main hill masses, the westernmost (inland) complex is the most extensive. It consists of a northern series of three parallel faulted ranges—the Hsing, Lu, and T'ai, which stretch northeastward for more than 200 miles—and a more diversified, lower, and more exposed southern portion. The granitic T'ai range, dominated by Mount T'ai, the most famous of China's five sacred mountains, attains a maximum elevation of 5,000 feet. The mountains of the peninsular mass to the east seldom rise above 700 feet. There, surface erosion has etched irregular and deeply cut valleys, and rounded hills contrast sharply with small intermontane basins. Both the north and south coasts of the peninsula are rocky, with hills dropping precipitously to the sea and separating a series of intensively cultivated crescent-shaped plains.

*Drainage.* Drainage is predominantly radial and subject to the prevailing configuration of the mountains. The only navigable river (other than portions of the Huang Ho) is the Hsiao-ch'ing, which emerges from a small spring-fed lake in a limestone outcrop zone near Chi-nan and flows parallel to the Huang Ho, before emptying into Lai-chou Bay. The southern hills, in contrast, are drained by several rivers in arable valleys, typified by those of the Tung-wen River system, which eventually terminate in the marshy plain east of the Grand Canal in Kiangsu Province.

*Soils.* The soils of Shantung fall into two broad categories associated with upland or lowland distributions.

© D.E. Cox—CLICK/Chicago



Cable car carrying passengers to the top of Mount T'ai in Shantung Province.

Shantung brown soils

The so-called Shantung brown soils are found over most of the two major hill masses and include a variety of brown forest and cinnamon-coloured soils formed through clay accumulations and sod processes.

A distinctive variant of the typical Shantung brown soil is the recalcified soil (soil that has been made hard or stony by the deposit of calcium salts); it is found on the northern perimeter of the central hill mass. Calcareous alluvial soils predominate in both lowlands and plains. They are usually quite fertile, depending on both the length of time they have been cultivated and their proximity to urban centres (where heavier fertilization with human and animal wastes results in rich, dark-coloured soils). Silty alluvium covers most portions of the North China Plain area of the province.

The sandy ginger soil

Another distinctive soil type found in central and western Shantung on the North China Plain is the subsurface *sha-chiang t'u,* or "sandy ginger soil." This appears at the lowest elevations of alluvial plains where surface water remains unevaporated for several months until the dry season and also in sections of the plains subject to annual alluvial inundation. Such soils are always covered with alluvium or redeposited loess. Their name derives from the appearance of lime concretions that resemble the shape of ginger roots. Other *sha-chiang t'u* soils develop impervious layers of limestone hardpan.

*Climate.* Shantung falls within the North China climatic region, which extends from the Huai River in the south to the Hopeh–Liaoning border in the north. It is characterized by a continental climate with cold winters and hot, dry summers. Climatic variation prevails, however, between the peninsular and inland zones of the province.

The inland zone, especially in its northern sections, is subject to the full effect of the winter monsoon, when cold, northwesterly winds continue through December. By March wind direction gradually reverses, and warmer, southeasterly winds prevail throughout the summer. In the inland zone, annual precipitation ranges from 10 inches in northwest Shantung to 20 to 24 inches as one approaches the mouth of the Huang Ho. Of the total annual precipitation, 70 to 80 percent falls in summer. The interior areas of Shantung are also subject to severe winter and spring dust storms, sometimes followed by droughts and frequent summer floods. Temperatures in the inland zone range from a mean January reading of 25° F (−4° C) in the northern interior to a mean of 82° F (28° C) in July. This area is subject to freezing temperatures during one to three months, with frosts common from late October to April. Rivers often freeze over for extended periods during the winter months. In the interior zone the growing season extends 200 to 250 days.

Rainfall and temperature

The maritime orientation of the Shantung Peninsula tends to modify the climatic extremes of the inland zone. The northern half of the peninsula is subject to winter snow and rainstorms and to extensive coastal ice from the mouth of the Huang Ho to Wei-hai and Chefoo (Yen-t'ai); the southern half is somewhat warmer. Mean January temperatures range from 25° F (−4° C) on the northern coast of the peninsula to 32° F (0° C) in the south. There is less temperature difference during the hot summer months when the mean July temperature is 79° F (26° C), but the ports of Chefoo and Tsingtao are cooler than interior stations. Maximum summer temperature in these ports rarely exceeds 77° F (25° C). Sea fog is common along the north and south coasts of the peninsula. Because of the high relative humidity, annual mean precipitation over the peninsula reaches 31 inches, with less seasonal contrast than in the interior of the province. Heaviest precipitation occurs on the south-facing slopes of the central and peninsular hill masses.

*Plant life.* The limited natural vegetation that remains in the intensively cultivated inland zone of Shantung is found in minor depressions in the flat, alluvial landscape. Species there included reeds, grassy legumes, and several varieties of shrubs, notably tamarisk. Halophytic (salt-tolerant) vegetation is common in alkaline and saline soil areas along the coasts of the Po Hai and southern Shantung near the Kiangsu border. Many of the halophytic shrubs are harvested for fuel and are used for salt manufacture.

Halophytic plants

Lienliu, a shrub with long willowy branches, is used for basket weaving, while other plants are woven into thatch mattings and sunshades. Poplar, pine, and arborvitae (an aromatic evergreen tree of the cypress family) are planted around settlements, along roads, and on the coasts.

The mountainous zones of Shantung are almost completely deforested, with only a small part of the area covered by scattered deciduous and coniferous forests interspersed among barren, eroded hills. Several types of pine grow at higher elevations on rocky, shallow soils in association with alpine meadow species. On the lower slopes and in the valleys, mixed oak, elm, cedar, linden, ash, maple, and chestnut forests appear along with such economically important fruit trees as apple, pear, apricot, and peach. Other deciduous species found at the lower elevations include the pagoda (or Chinese scholar) tree, the white mulberry, Persian walnut, silk tree, and acacia. For centuries Shantung forests were overharvested for fuel and timber, and natural regeneration became extremely difficult. Since 1949, reforestation and closer regulation of timber harvesting has resulted in more extensive growth.

Despite the obliteration of much of Shantung's natural vegetation cover, the peninsular zone still exhibits an interesting mixture of northern and southern vegetation. Along with common northern plants, uniquely southern varieties, such as wingnut, magnolia, and styrax, are common.

*Animal life.* Through long periods of human settlement, intensive cultivation, and destruction of forests, animal life has suffered drastic decline. Animals include roe deer and field and harvest mice; birds include mandarin ducks, dollar birds (belonging to the roller group), and large owls. Even with recent attempts at reforestation, formerly extensive populations of native birds and mammals have almost vanished. Species of insects, beetles, and moths, however, are still unusually diverse and varied.

*Settlement patterns.* The two largest cities are Tsingtao and Chi-nan, followed by the Tzu-po conurbation, a leading mining and industrial zone at the northern edge of the central hill mass, about 20 miles east of Chi-nan. Other cities include Chefoo and Wei-hai, ports and fishing centres on the northeast coast of the peninsula; Wei-fang, an industrial and commercial town on the central Chiaolai Plain; and Hsin-t'ai, a mining town south of Tzu-pa.

Principal cities and towns

The greatest rural population densities are found in three areas. The first is one of the earliest settled places in the province, where irrigation works were constructed as long ago as the Han dynasty; it lies along the foothills of the central hill mass. The second, the southwestern Ho-tse–Ting-t'ao–Chi-ning area, is bounded on the northwest by the Huang Ho and on the southwest by the former course of the Huang Ho. This area was frequently subject to flooding, but because of its fertility and level terrain gradually became densely settled. The third area constitutes a fertile, irrigated strip along the north coast of the Shantung Peninsula between I-hsien and Lung-k'ou.

**The people.** Shantung's population is predominantly Northern Mandarin-speaking and of Han (Chinese) origin, but there are small concentrations of Hui (Chinese Muslims) in Chi-nan, the capital, in Chou-ts'un (near Tzu-po), and in Chi-ning and Lin-ch'ing (trading centres on the Grand Canal in western Shantung). The population, more than 90 percent rural, is fairly evenly distributed over the level, cultivated areas of the province.

**The economy.** Shantung has a diversified agricultural and industrial economy. A broad range of food and cash crops is grown for internal consumption and export to other provinces and overseas. The province's industrial base has been expanded since 1949. Before World War II, light industrial enterprises produced limited quantities of light products. Although the province often suffered a food deficit, agricultural products were continuously exported along with salt, coal, iron ore, and bauxite. Since 1949 relatively greater emphasis has been given to the development of industry, mining, and electric-power generation, although the absolute overall level of agricultural output has continued to rise. Shantung attained food self-sufficiency in 1970, while still increasing cash-crop production.

*Resources.* Shantung's industrial base is supported by extensive mining activities, principally coal mining, which

was originally developed by German concessionaires in the early 20th century. Considerable mechanization of coal-mining operations has taken place since 1949. There are also major iron-ore deposits located north of Tzu-po at Chin-ling-chen, some bauxite is mined near Nan-ting (Tzu-po), and gold is scattered throughout the peninsular hills. Edible salt is produced on both the north and south coasts of the Shantung Peninsula.

Oil and oil products have exerted an increasing influence on the economy of the province. The Sheng-li oil field, China's second largest continental oil production area, is located on the mouth of the Huang Ho in the Po Hai. The field yields a type of oil especially suitable for fuel. A lighter oil is produced at the Tung-p'u field, on the Shantung–Honan border. A pipeline completed in 1978 connects the Sheng-li oil field with those of the North China Plain in Hopeh and the ports and refineries of the lower Yangtze River area.

Major emphasis since the late 1970s has been given to increasing electric-power generation. High-voltage transmission lines and feeder lines to rural areas extend throughout the province and have substantially increased the supply of rural electric power, as well as the amount of electrically irrigated and drained acreage.

*Agriculture.* The success of agriculture in Shantung since 1949 is attributable to extensive investment in irrigation, flood-control, and soil-conservation measures; drainage of alkalinized and salinized land; and increased mechanization. More than 60 percent of the province's wasteland has been reclaimed and cultivated, and in most irrigated areas the productivity ratio has improved from three crops in two years to two crops in one year. The leading food crops—wheat, corn (maize), soybeans, kaoliang (a variety of grain sorghum), spiked millet, and sweet potatoes—account for most of the total cultivated acreage of the province. The remaining arable land is given over to cash crops, which contribute substantially to agricultural earnings.

rincipal ash crops

Peanuts (groundnuts), the leading cash crop, are grown primarily in the peninsular uplands and in the south central sector. The large variety of peanuts grown in Shantung is especially well suited for oil pressing, and Shantung is a leading manufacturer of peanut oil for cooking. Shantung's other major cash crop, cotton, is grown throughout the province but is concentrated in the western and northern sections on the intensively irrigated lands near the mouth of the Huang Ho. Other cash crops include tobacco, grown chiefly on irrigated land in the vicinity of I-tu and Wei-fang; hemp, produced on low ground in the southwest; and fruit, grown on lower slopes of the central and peninsular hill masses.

Animal husbandry plays an important role. The most common animals are pigs, yellow oxen, and donkeys. Sheep are also raised in the uplands. Sericulture (silkworm raising), another important subsidiary activity, has been carried out in Shantung for hundreds of years. The popular fabric known as shantung was originally a rough-textured tussah, or wild silk cloth, made in the province. Silkworm raising is most common in the central hills near I-tu, Lin-ch'ü, Tzu-ch'uan, and Lai-wu, and most of the raw silk is sent to other provinces for processing and spinning.

Shantung's seaward orientation and its excellent harbours, as well as the convergence of cold and warm currents in offshore waters, have fostered a thriving ocean fishing industry, complemented by the intensive development of pisciculture in the province's western lake region. Trawlers and smaller fishing craft operate from ports around the peninsula and off the Huang Ho Delta. The ocean catch consists mainly of eels, herring, gizzard shad, fish roe, and several varieties of shrimp and crab. Freshwater varieties raised artificially are chiefly carp and crucian carp.

*Industry.* The province is still especially well known for its light industrial products, despite post-1949 gains in heavy industry. Tsingtao, the major manufacturing centre, has a large textile industry, a locomotive works, and chemical, tire, and machine-tool factories. Pre-World War II oil pressing (peanut oil), cigarette making, flour milling, brewing, and beverage distilling installations are still important. Chi-nan—long famous for its silks, precious stones, and handicrafts—now also manufactures trucks, agricultural machinery, machine tools, precision instruments, chemicals, fertilizers, and paper. Tzu-po produces glass, porcelain and ceramics, and textiles. Wei-fang is an important food-processing centre, and it also has metal-processing and textile factories. Some of Shantung's better known handicraft goods are embroidered tablecloths from Chefoo and Lin-tzu, straw braids for hat weaving from P'ing-tu (east of Wei-fang), poplins, pottery, and ceramics.

ight ıdustry nd hand- rafts

*Transportation.* Shantung's earliest railways were built in the first decade of the 20th century during the time of the German concession. One of the lines traverses the province from north to south and another east to west, connecting Tsingtao and Chi-nan. Since 1949 new lines have been built, including a major trunk line from Tsingtao north to Chefoo.

Shantung's highways connect every district in the province, but many of them have earthen surfaces and are used either for short-haul transport or as feeder routes for the major railways. Truck traffic accounts for a majority of the total annual vehicular movement over Shantung's highways, as compared with only a small proportion in other North China provinces.

Limited use of inland waterways

Except for portions of the Huang Ho and of the Hsiao-ch'ing River in northern Shantung, part of the Grand Canal in the west, and the I River in the southeast, inland-waterway transport is limited. The chief route—for shallow-draft craft only—extends upstream from Li-chin, about 50 miles inland from the mouth of the Huang Ho, to Ch'i-ho, the main Huang Ho river port in Shantung and just northwest of Chi-nan. The Grand Canal is navigable only to a limited extent south of the Huang Ho.

Shantung has a number of excellent seaports. Tsingtao is the largest in tonnage handled, although Chefoo, Wei-hai, and Lung-k'ow on the north coast of the peninsula also handle a considerable amount of shipping. Coastal shipping also plays an important role in Shantung's economy. Tsingtao alone handles more than one-third of the province's intraprovince trade. Trade between Tsingtao and Shanghai and Tsingtao and Lü-ta is particularly heavy.

**Administration and social conditions.** *Government.* Shantung is divided into six prefectures (*ti-ch'ü*) and eight prefecture-level municipalities (*shih*). At the next lower administrative level there are counties (*hsien*) and county-level municipalities (*shih*). The Shantung Provincial Revolutionary Committee, the chief provincial administrative body from 1967, was replaced in 1980 by the People's Government, which is the administrative arm of the People's Congress. Until the early 1980s the rural people's communes, made up of production teams and brigades, served as the lowest administrative units. With the institution of family farms as primary production units, commune labour allocation, production, and marketing became less important. In many areas, county seats operate as coordinating centres for the production and distribution of commodities produced in the areas under their administrative jurisdiction.

Higher education complex in Chi-nan

*Education.* Most of Shantung's institutions of higher education are located in the provincial capital, Chi-nan, with smaller or special-purpose schools scattered widely throughout the province. Among those in Chi-nan are the Shantung Medical College and the Shantung Institute of Technology. Shantung University is in Tsingtao. Tsingtao is China's major centre for research training in marine science and technology. Institutions include the Institute of Oceanology of the Academia Sinica (Chinese Academy of Sciences) and the Shantung Oceanography College, which is under the jurisdiction of the national-level Ministry of Education.

*Health and welfare.* Before 1949 Shantung was particularly hard-pressed by the pressure of population on the land, by the common occurrence, especially since the latter half of the 19th century, of floods, droughts, dust storms, excessive soil salinization and alkalinization, and insect infestations, and by frequent military and civil disturbances. Few serious attempts were made by officials of either the Ch'ing (Manchu) dynasty (1644–1911/12) or, later, by the Republic of China to ameliorate the difficult social conditions of the peasant population. With the ex-

ception of missionary-financed and missionary-controlled undertakings in areas under foreign influence or administration, such as Tsingtao, Chefoo, and Chi-nan, modern intensive health-care facilities were virtually nonexistent, and there was only token support for public higher education. Water supplies, environmental sanitation facilities, and public housing were similarly inadequate to the needs of the populace, and public health services were neglected and understaffed.

Since 1949 the public health services in both rural and urban areas have been improved, and formerly common ailments such as kala-azar (a severe infectious disease transmitted by the sand fly), leprosy, and a variety of nutritional-deficiency diseases have been eliminated. Most large and medium cities now have adequate water-supply systems, often built in conjunction with multipurpose water-conservancy schemes to improve and stabilize the watersheds of nearby rivers. Along with water supply, the construction of sewage treatment facilities in many cities has also helped raise public-health standards.

Multipurpose systems of water supply

Not only has extensive tree planting enhanced the beauty of most Shantung cities, but "greening" has been officially designated as a primary task of urban reconstruction in order to ameliorate the effects of harsh climates and to improve health conditions. In Tsingtao alone, some 4,000,000 trees were planted from 1949 to 1959, while in Chi-nan, a green belt has been built on the site of some dilapidated sections of the ancient city wall. Along with urban reforestation, recreational facilities have been expanded, improved, and made readily available for public use. Many famous temples, hot springs, shrines, parks, lakes, and museums are frequented by the populace. In Chi-nan, a city famous for its hot springs, where for centuries poets, scholars, and officials enjoyed diverse pleasures, several new parks have been built and old buildings restored. Tsingtao, known as the most pleasant beach resort in North China, is also famous for its parks.

**Cultural life.** Shantung's rich cultural and folklore tradition is most clearly evidenced in the temples, shrines, legends, and cults associated with Mount T'ai and with the temple and tomb of Confucius at Ch'ü-fu, north of Chi-ning. Despite official disavowal from 1949 until the mid-1980s of their religious, parareligious, animistic, and superstitious connotations, the temples, shrines, and their surrounding areas have been restored, renovated, and converted to public parks so as to assure their preservation as important symbols of the national cultural heritage.

Mount T'ai

Mount T'ai—known also as Tung-yüeh, or "Eastern Peak," to distinguish it from the "Southern Peak" (in Hunan), the "Central Peak" (Honan), the "Western Peak" (Shensi), and the "Northern Peak" (Shansi)—is the most prominent of these five sacred mountains where the emperors once offered sacrifices to Heaven and Earth. It was also the place where for centuries Buddhists, Taoists, and Confucianists built more than 250 temples and monuments to honour deified historical personages and to immortalize the sacred presence and supernatural powers of the supreme mountain deity of Mount T'ai. The mountain was deified at least as early as Han times, and in the Sung dynasty it was elevated by the emperor Chen-tsung to the position of "Equal with Heaven." Incantations and prayers offered to the deity of Mount T'ai by countless emperors are inscribed in stelae along the ascent to the summit, and temples are distributed in T'ai-an and on the mountain itself.

The Temple of Confucius, Confucius' tomb, and the residence of the Kungs (Confucius' lineal descendants) at Ch'ü-fu are also maintained as national historic monuments. Both the temple and the Kung residence are laid out with elaborate temples, monuments, pavilions, and gates, and have collections of stelae dating, in some cases, from the Han dynasty.

### HISTORY

A Neolithic culture—known as the Lung-shan because of archaeological remains discovered near the township of that name—existed on the Shantung Peninsula in the 3rd millennium BC; it played a key role in the establishment of a common rice-based cultural grouping that apparently spread along the Pacific seaboard from the peninsula to Taiwan and eastern Kwangtung.

Western Shantung formed part of the Shang kingdom (18th–12th century BC). By the Ch'un-ch'iu (Spring and Autumn) period (770–476 BC) it had become the centre of political and military activity that resulted from the eastward expansion of the Chou, following their conquest of the Shang. One of the small southern Shantung states was Lu, the birthplace of Confucius and Mencius. Also in the "Eastern Territory"—an early name for Shantung—was Ch'i, extending over the major part of the peninsula; it became an important economic centre, exporting hemp clothing, silk, fish, salt, and a unique variety of purple cloth to all parts of China. Beginning in the Six Dynasties period (AD 220–589), Shantung became North China's leading maritime centre, receiving commodities from the South China coastal area (now Fukien and Kwangtung) for transshipment to destinations north and south of the Huang Ho. Thus, Shantung has been a part of China from its very beginning as an organized state.

Grand Canal

In 1293 the Grand Canal, running generally north to south, was completed, making western Shantung a major inland trading route. Yet even after the completion of the canal, maritime trade still remained important to Shantung, and the peninsula retained its dominant economic position. In the great agricultural areas of the province, however, early deforestation and the long-established practice of clearing land for cultivation without providing for flood prevention and control measures led to serious and ultimately disastrous erosion and wastage of valuable agricultural land.

In the 19th century these problems were worsened by shifts in the course of the Huang Ho. From 1194 until the early 1850s the Huang followed the original bed of the Huai along the Shantung–Kiangsu border before emptying into the Yellow Sea. After 1855, when a series of devastating floods was followed by extensive dike construction, the river changed to its present course some 250 miles to the north. Hardships and food shortages from floods and other natural calamities increased in intensity throughout the 19th and 20th centuries. This resulted in a substantial emigration of Shantung peasants to the Northeast (Manchuria) and to Inner Mongolia, with more than 4,000,000 people emigrating between 1923 and 1930.

In the closing decade of the 19th century Shantung came under the influence of German, British, and Japanese interests. It was occupied briefly by Japanese troops after the Sino-Japanese War of 1894–95. In 1897 Germany landed troops, and in 1898 a treaty was signed by which China ceded to Germany, for 99 years, two entries to Chiaochou Bay and the islands in the bay and granted the right to construct a naval base and port, Tsingtao. Germany used Tsingtao as a base to extend its commercial influence throughout the peninsula; it developed coal mines and constructed a railway (1905) from Tsingtao to Chinan. Similarly, in 1898 Great Britain obtained a lease for Wei-hai-wei (modern Wei-hai), another strategic port near the northern tip of the peninsula. This was in response to the Russian occupation of Port Arthur (now Lü-shun). With the advent of World War I, Japan took over German interests in the peninsula and in 1915, as one of its infamous 21 Demands, compelled the Chinese to give official recognition to the renewed occupation. Taking up the Shantung question, the imperialist powers decided in 1919 to grant Japanese occupation, which Japan maintained until 1922.

In the Sino-Japanese War of 1937–45, even though the Japanese had gained control of most of Shantung by the end of 1937, they miscalculated Chinese strength and suffered a serious defeat—their first of the war—at T'ai-erh-chuang, in southern Shantung, in 1938. In the postwar struggle between the Chinese Communists and Nationalists, Shantung came under Communist control by the end of 1948. (B.Bo./Ed.)

## Tientsin

For coverage of Tientsin, see the *Macropædia* article TIENTSIN.

# LOWER YANGTZE VALLEY

## Anhwei

Anhwei (An-hui in Wade–Giles romanization, Anhui in Pinyin), one of the smallest provinces of China, covers an area of 54,000 square miles (139,900 square kilometres) and stretches for 400 miles (640 kilometres) from north to south. Anhwei, which is landlocked, is bounded by the provinces of Kiangsu to the northeast, Chekiang to the southeast, Kiangsi to the south, and Hupeh and Honan to the west. Its northern extremity barely touches the southern extremity of Shantung. Its name means "beautiful peace" and is derived from the names of two cities—An-ch'ing and Hui-chou (now She-hsien). The capital, Hofei, is located in the heart of the province. Anhwei was long one of China's poorest and most undeveloped areas. Since 1949, however, successful attempts have been made to utilize the province's economic and human resources. Vast irrigation schemes on the major rivers have alleviated severe periodic flooding and have also provided increased agricultural land and electric power.

### PHYSICAL AND HUMAN GEOGRAPHY

**The land.** Anhwei lies in the path of the Neo-Cathaysian Geosyncline (a great downward flexure of the Earth's crust), which stretches across the entire length of eastern China from Heilungkiang on the Soviet border to Kwangsi on the Gulf of Tonkin. The floor of the geosyncline is steadily sinking under the weight of the silt carried by the Huang Ho (Yellow River) and the Huai River. The sediment is estimated to be more than 2,000 feet (600 metres) deep.

*Relief.* The northern portion of Anhwei Province is occupied by the North China Plain—an immense level surface that has periodically been flooded by its dominant rivers. The southern section of the province, the Yangtze Valley, is separated from the northern plain by a series of mountains that stretch roughly from west to east. The Ta-pieh Mountains—an extension of the Tsinling–Fu-niu ranges lying to the north of the Yangtze—form a convex curve of steep slopes facing east and northeast on the southwestern Hupeh–Anhwei border. The Paichi Mountains lie south and east of the Yangtze and form the southeastern border between Anhwei and Chekiang. Composed mainly of granite, metamorphic rock (rock formed in the solid state by heat and pressure), and sandstone, they include the Huang Mountains, a range that rises to a height of 6,000 feet and is beloved by poets and artists for its massive shape and lush vegetation.

The Huai and Yangtze river basins

*Drainage.* The northern plain is drained by the Huai River. Its many left-bank tributaries rise in the western mountains and flow eastward across Honan Province into Anhwei. The shorter right-bank tributaries rise in the Ta-pieh Mountains. The Huai flows across the level plain and drains into Hung-tse Lake, which lies just across the eastern border with Kiangsu Province. The river basin is subject to widespread and disastrous floods.

The watershed between the Huai and the Huang to the north is barely perceptible. At least three times the Huang Ho has changed its course to flow south of the Shantung Peninsula, joining its waters with the Huai to flow into the Yellow Sea. Because of its susceptibility to disastrous floods, the Huai Basin was chosen in 1949 to be the site of the first large-scale water-conservation project to be undertaken by the People's Republic. The scheme entailed building several dams on the upper reaches of the Huai and its tributaries in order to control the flow of floodwaters, constructing or rebuilding hundreds of miles of dikes and irrigation and drainage canals along the main rivers, clearing the entry to and the exit from Hung-tse Lake, and digging the Su-pei Canal from the eastern edge of the lake to the Yellow Sea. Since 1956 there have been no serious inundations, local flooding has been controlled, and irrigation has successfully compensated for threatened drought.

The low line of hills (Hua-yang Mountains) that extends northeast from the Ta-pieh range to Hung-tse Lake marks the divide between the Huai and Yangtze river basins. The Yangtze plain is studded with lakes that, in time of flood, join the river and increase its width in places to five miles. The change of river level between summer and winter is not as great as it is in Hupeh; nevertheless, winter navigation is difficult. The Yangtze plain is crisscrossed by canals that are used for irrigation, drainage, and transport.

*Soils.* The province's complex soil structure can be divided broadly into three categories. The uplands to the north and south of the Yangtze are mainly podzolized (leached) old and young red earths that are susceptible to erosion but are valuable for the cultivation of tea. The Yangtze floodplains are composed chiefly of alluvium and rice-paddy soils that have a slightly acidic character. The alluvial soil of the Huai Basin—particularly the area drained by the left-bank tributaries and extending into the North China Plain—is, however, calcareous (chalky). It includes curious mineral masses that are known as *shachiang t'u* ("sandy ginger soils") because they resemble ginger roots. They form in low-lying places where the ground is waterlogged, rarely occur on the surface, and sometimes form a hardpan, or basin, some feet below ground level.

*Climate.* Anhwei shares with much of the rest of China the seasonal monsoon climate characterized by hot, wet summers and cooler, dry winters. As it extends for about 400 miles from north to south, however, Anhwei experiences appreciable variations in climate. Mean January temperatures in the north of the Huai Basin range between 32° F (0° C) and 36° F (2° C). Mean July temperatures in both regions are 82° F (28° C). The north has between 200 and 230 frost-free days, while the south experiences between 230 and 250. The relative humidity is considerably higher in summer in the Yangtze Valley than in the Huai Basin. Northern Anhwei experiences some discomfort when winter dust storms, brought by bleak winter winds, sweep down from the Shansi Plateau.

Rainfall distribution

There is a marked difference between the rainfall of the Yangtze Basin and that of the Huai plain. In the south the precipitation amounts to between 47 and 71 inches (1,200 and 1,800 millimetres) and is relatively evenly distributed. In the north the mean annual rainfall is between 24 and 31 inches, more than half of which falls between June and September. Rainfall becomes progressively unreliable as one progresses northward until the famine lands of the northwest are reached. Summer evaporation in northern Anhwei is intense.

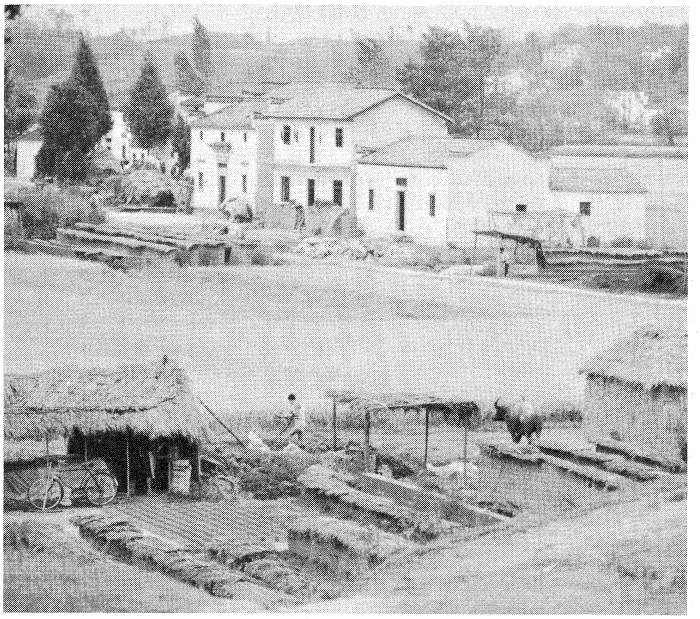
*Plant and animal life.* The original vegetation was forest and woodland, but centuries of intensive settlement have led to deforestation everywhere except in the western and southern uplands. This process, accompanied by soil erosion, has been very rapid since the mid-18th century. Cultivated crops and grass have replaced the trees, but fuel gatherers annually strip the hillsides of grass. The only remains of the forests on the plains are the bamboo groves and woods surrounding temples and villages. Since the 1950s, resolute efforts at afforestation have been made, and fuel gathering has been restricted.

One of the world's rare animals, the Yangtze alligator, lives in the environs of Wu-hu. It is less than six feet long and feeds mainly on fish and small animals. Apart from rodents and reptiles, very few wild mammals remain in the densely settled and cultivated plains.

Population distribution

**The people.** The regions of densest population are the tributaries and banks of the Huai above Pang-pu and the diked areas along the right bank of the Yangtze. Generally, the villages in the Huai Basin are somewhat larger than those in the south. In the Yangtze Valley, villages are farther apart, and there are more scattered homesteads. Population density is considerably less in the hilly region that separates the two river basins, with the lowest densities in the southern uplands.

As in the rest of China, there is no precise demarcation between rural and urban population. Places of 2,000 people and fewer that are essentially engaged in agriculture



Rural homestead in a village near T'un-hsi in southern Anhwei Province.
© D.E. Cox—CLICK/Chicago

are classified as rural. There are four large towns—Ho-fei, the capital; Huai-nan; Pang-pu; and Wu-hu. Towns of lesser importance are An-ch'ing, the former provincial capital; Ma-an-shan; and Ta-t'ung. Since 1954 many of these towns have grown rapidly because of industrialization, and new towns have also appeared.

The population is almost totally Han (Chinese), and there are no pockets of aboriginal peoples as there are in the provinces south of the Yangtze. Until the 7th century AD there were large numbers of Hakka, but they were later driven south into Fukien and Kwangtung. Those of their descendants who remain have been integrated with the main body of the Chinese population.

**The economy.** Until 1949 Anhwei was regarded as the most backward province of eastern China. Most of its population was rural, and the standard of its agriculture was low because of poor use of water resources. Mineral resources were little developed. Since the 1950s there have been great advances in agriculture and industry. Improvements in irrigation now allows more land in the south to be double cropped in rice; and other water conservancy and land improvement measures and a return to household farming have improved yields and increased per capita net farm income. The Shanghai special economic zone, established to promote industrial growth, includes Anhwei Province, and long-range planning envisage the development of Anhwei's mineral resources on the model of the Ruhr Valley in West Germany.

*Agriculture.* Wheat is the predominant crop in the Huai Basin to the north, and—more importantly—rice is grown in the Yangtze Basin to the south. In the relatively wetter Yangtze Valley well over half of the cultivable land is devoted to rice, while in the drier Huai Basin about one-third of the land is under wheat. Most of the land produces two crops a year. Anhwei is also one of China's most important soybean producers; the beans are grown mainly in the north in rotation with wheat or barley. Increasing amounts of kaoliang (a variety of grain sorghum) and millet are also grown in the north. The main industrial crops are vegetable oilseeds, cotton, tea, fibres, and tobacco. Among the vegetable oils, the most important are rapeseed, peanut (groundnut), and sesame. Cotton is grown mainly on the northern Huai plain. Hemp, jute, and ramie (an Asian nettle that yields a fibre used for making textiles) are also grown.

Tea and silk cultivation

Anhwei has been famous for its tea since the 7th century, when teas were exported to the rest of China, as well as abroad. This trade became depressed in the late 19th and early 20th centuries but has now revived; Keemun black tea is especially famous. The main areas of cultivation are on the slopes of the Ta-pieh Mountains, north of the Yangtze, and on the Paichi Mountains along the Anhwei–Chekiang border. Sericulture, the production of raw silk by the raising of silkworms, has also been revived since the 1950s. During the years of warfare from 1937 to 1949, many of the mulberry trees were felled to deprive guerrilla forces of cover. These have been replaced, and both the mulberry-feeding moth (*Bombyx mori*) and the tussah silkworm are reared, providing the raw material for Nanking brocades and Wu-han silk fabrics.

Large domesticated animals are used almost exclusively for draft purposes. Pigs are the main source of meat, and sheep are raised in increasing numbers in northern Anhwei. Numerous rivers and lakes abound in fish, mainly carp and white bream. Fish are bred all along the Yangtze.

*Industry.* Anhwei's industrial development was quite limited before the 1950s, based mainly on the partial development of its rich copper, iron, and coal deposits. In the 1950s there was substantial development of the Huai-nan coal basin—first worked in the 1920s—followed in the 1960s by the development of the Huai-pei basin. This coal is the primary source of power for Anhwei's mainly thermally generated electricity supply, although hydroelectric power units on tributaries of the Huai supply the province's major industrial centres. Anhwei is also rich in high-quality iron ore, located near the Kiangsu border. These reserves, first developed by the Japanese during World War II, were further developed in the 1950s. A large copper shaft mine and smelter were built at T'ung-ling in the 1950s on the site of a T'ang dynasty copper lode, and T'ung-ling developed into a major supplier of blister copper.

The capital, Ho-fei, is the province's major industrial centre; its development began in the 1950s with the transfer of a number of textile and light industry plants from Shanghai. In the 1960s steelmaking and machine-tool plants were established to serve provincial mining, electrical, and chemical industries, and the city also developed into a producer of chemical fertilizer. Pang-pu, in northern Anhwei, has developed into an important supplier of agricultural machinery and a major food-processing centre. Wu-hu, on the southern bank of the Yangtze, has been a commercial centre since the 1960s, with a port that plays an increasingly important role in both domestic and international trade.

*Transportation.* Throughout the centuries waterways have been the main means of communication. Water-conservation schemes have increased the navigability of rivers and canals, and traffic on them is heavy. The ports of Ma-an-shan, Wu-hu, T'ung-ling, Ta-t'ung, and An-ch'ing on the Yangtze can be reached by oceangoing vessels of 15,000 tons during the high-water summer months. The principal railway in Anhwei enters the province from Suchow across the northern Kiangsu border and runs south to Pang-pu, where it divides—the main line running southeast to Nanking in Kiangsu Province, and a branch line running south to Ho-fei, from where it runs southeastward to the Yangtze opposite Wu-hu. Ho-fei stands at the centre of the province's highway system, with main roads running to Nanking, Pang-pu, and Wu-han.

**Administration and social conditions.** *Government.* From 1950 to 1954 Anhwei was included in the East China greater administrative region, which embraced all the east coast provinces from Shantung to Fukien. In 1954 provincial (*sheng*) government was made directly subordinate to the national government. Anhwei is subdivided into eight prefectures (*ti-ch'ü*) and eight prefecture-level municipalities (*shih*). Below this level it is divided into counties (*hsien*) and county-level municipalities (*shih*). The provincial Revolutionary Committee appointed by the central government during the Cultural Revolution was replaced in 1980 by the People's Government, which is the administrative arm of the People's Congress.

*Education.* Western learning was introduced in the early decades of the 20th century through the teaching of Christian missions. The National Government after 1928 attempted to expand education throughout the province, but most people remained illiterate. After 1949 the prob-

Role of the communes in education

lem of illiteracy was attacked with vigour. The communes established in the late 1950s were then made responsible for primary and middle-school education, and a program combining work and study was introduced. On the slopes of the Ta-pieh Mountains, several schools of sericulture have been established; and the tea-growing regions now have special schools for farming and study. Despite considerable advances in education, however—Anhwei has some 50,000 primary and secondary schools and more than 30 colleges and universities—only about two-thirds of the adult population is literate.

*Health and welfare.* Medical and health services that were developed after 1949 were at first concerned primarily with public hygiene and preventive medicine. Medical teams were sent into the countryside to inoculate the population and to teach and advise on matters of public health. From 1960 hospitals were built in the communes and all the major towns. The emphasis on public hygiene has, however, been maintained. Both Western and traditional medical practices are employed.

**Cultural life.** The Anhwei region was one of the earliest areas in the Yangtze Basin to be settled by Sinitic peoples. Many of the current linguistic and cultural traits of the region were shaped during the Nan (Southern) Sung period (1127–1279), particularly the characteristic conservatism of language and art forms. A number of separate regional subcultures continue to exist, including the Hui-chou culture from the region around She-hsien, renowned for its commercial and clan traditions.

HISTORY

During the Chan-kuo (Warring States) period of the Chou dynasty (475–221 BC) Anhwei formed part of the large southern state of Ch'u. Between 221 and 206 BC the Ch'in dynasty unified the states, and a great southward migration along the natural highway of the North China Plain and the Huai River basin began. Anhwei became the first part of southern China to be settled by the Han. Unrest following the fall of the Han dynasty in AD 220 led to further immigration into the area.

Growth under the early canal system

The Yangtze River basin subsequently became the granary of the empire, and an improved transport and canal system was developed across northern Anhwei to carry tribute grain to the capital from Su-chou to Pien (modern K'ai-feng) in Honan and from there to Lo-yang in Honan. It was later superseded by the New Pien Canal, built during the Sui dynasty (581–618). The new canal ran along the Kuei River and then cut across the region to Pien, forming the main line of communication to the capital. During the 12th century Anhwei was the scene of bitter battles between the Nan Sung emperors and the invading Juchen. After the establishment of the capital at Ta-tu (modern Peking), the Yüan, or Mongol, dynasty (1206–1368) constructed the Grand Canal to the east, connecting Hang-chou in Chekiang Province with Ta-tu, and the previously built waterways fell into disuse.

In the early 1850s the Huang Ho made one of its great changes in course, flowing into the Po Hai, north of the Shantung Peninsula instead of south into the Yellow Sea. The loss of water for the Huai Basin resulted in great distress for the farmers of northern Anhwei. Subsequent peasant risings—together with the Taiping Rebellion of 1850–64—resulted in widespread devastation.

The Yangtze was opened to foreign shipping in 1860, but it was not until 1877 that the walled city of Wu-hu in Anhwei was opened to international trade. Although it was the province's only treaty port, the city never figured prominently in overseas commerce.

In 1938 the Huang Ho was temporarily diverted south of Shantung by the Nationalist government, which blew up the river's dikes in Honan in an attempt to stem the advance of Japanese invaders. The river waters then surged south to Hung-tse Lake on the Anhwei border, flooding an enormous area at the cost of about 900,000 lives. During World War II most of Anhwei was occupied by Japanese forces, but, because of the resistance of the Chinese inhabitants, Japanese control could be effectively enforced only during daylight hours. Between 1946 and 1949 the province was controlled by the Nationalist forces.

From 1949 to 1952 Anhwei was administered in two separate parts. The section north of the Yangtze, which came under Communist military control in 1949, was constituted as the North Anhwei Administrative District. The South Anhwei Administrative District was established several months later, after the People's Liberation Army (PLA) crossed the Yangtze and based its administration in Wu-hu. In August 1952 the province was reunified under the leadership of Zeng Xisheng (Tseng Hsi-sheng), a long-time veteran of the PLA. Anhwei's provincial administration experienced relatively greater leadership turnover than other provinces in the 1950s and '60s and a major shift in leadership during the Cultural Revolution. During the 1970s Anhwei supported many of the radical programs of the Cultural Revolution, but in the late 1970s the province became a base for many of the moderate reforms advocated by the post-Mao leadership.

(T.R.T./V.C.F./Ed.)

## Hupeh

Hupeh (Hu-pei in Wade–Giles romanization, Hubei in Pinyin) Province lies in the heart of China and forms a part of the middle basin of the Yangtze River. Until the reign of the great K'ang-hsi emperor (1661–1722) of the Ch'ing dynasty, Hupeh and its neighbour Hunan formed a single province, Hukuang. They were then divided and given their present names: Hupeh, meaning, "North of the Lakes" (of the Yangtze River); and Hunan, "South of the Lakes." Hupeh has an area of 71,800 square miles (185,900 square kilometres). Its capital is Wu-han, the composite name of the three cities of Han-k'ou, Han-yang, and Wu-ch'ang, which lie at the confluence of the Han River and the Yangtze at a point approximately 600 miles (1,000 kilometres) from the sea and halfway between Shanghai and Chungking.

PHYSICAL AND HUMAN GEOGRAPHY

**The land.** *Relief.* Almost all of Hupeh Province lies immediately north of the Yangtze River. Hupeh is bounded on the north by the eastern extension of the axis of the Tsinling, T'ung-pai, and Ta-pieh mountains. In the southeast the Mu-fu Mountains divide the province from Kiangsi. Along its central southern border there is no clear physical divide apart from the Yangtze itself: a lake-studded alluvial plain continues uninterruptedly southward to Tung-t'ing Lake in Hunan. The Wu-ling Mountains form the boundary between southwest Hupeh and northwest Hunan. Western Hupeh has highlands that lie at an altitude of more than 6,000 feet (1,800 metres) and consist of the eastern extension of two ranges, the Ta-pa and Fang-tou mountains, marking the boundary between Hupeh and Szechwan.

The level of the land falls rapidly, from west to east, to the lake plain, much of which is no more than 200 feet above sea level. This flat or gently undulating country is often suddenly interrupted by steeply rising isolated hills or ranges. The plain is the remnant of a former depression or old lake basin formed in the Pliocene Epoch (from 7,000,000 to 2,500,000 years ago), which has largely been filled with eroded red sandstone from Szechwan. The process of filling in is not yet complete; in consequence, large areas adjoining the Yangtze and Han rivers are covered by innumerable shallow lakes.

Rivers

*Drainage and soils.* The Yangtze cuts its way from the Szechwan Basin through the Ta-pa Mountains in a series of magnificent gorges and descends rapidly to the Hupeh Plain at I-ch'ang. The bed of the river at I-ch'ang is only 130 feet above sea level and is 960 miles from the sea. From this point onward its velocity decreases and its bed widens as it winds its way across the province from west to east. Finally, it forces a passage between the Mu-fu and Ta-pieh ranges into Anhwei and Kiangsi provinces. There the river again narrows to less than half a mile in width. In its course through Hupeh the Yangtze receives the waters of two tributaries, the Han and the Ch'ing rivers. It also receives, through Tung-t'ing Lake, the entire drainage of Hunan. The Han, itself a considerable river even by Chinese standards, rises in the Tsinling Mountains and

flows eastward in Shensi Province for about 200 miles. On entering Hupeh, it turns south in a much broader valley, or floodplain, and widens its bed, which varies from half a mile to a mile in width over much of this stretch. About 100 miles from its confluence with the Yangtze at Han-k'ou (*k'ou*, "mouth"), it turns east, threads its way through a maze of lakes, and, in the last few miles, narrows its bed to a mere 250 yards—a factor that is responsible for much flooding in summer.

The variation in the regime of the Yangtze between summer and winter is striking. At Han-k'ou, where the river is nearly a mile wide, the average difference between summer and winter levels is 45 feet. In winter the river is sluggish, with many shallows, and is navigable up to Han-k'ou only by specially built flat-bottomed river steamers. With the coming of the spring and summer rains the change is dramatic, and the river comes down as a mighty flood. In times of exceptional flood, as in 1931 and 1954, the flow reaches astronomical figures. Flooding of large areas of the surrounding low-lying land is normal each summer, when river and lakes combine. Marco Polo, who visited the area in the 13th century, reported that in places the river was more than 10 miles wide; his report was discredited as a gross exaggeration, but he had seen the Yangtze in the summer flood. In summer Han-k'ou is an ocean port, capable of receiving vessels of 10,000 to 15,000 tons. There are, however, navigational hazards at this time. Great care has to be taken that a vessel does not stray from the true river course and become grounded in more shallow water. If this happens, refloating is an urgent matter as the river is prone to quite rapid falls in level, and the vessel may be left stranded for a year.

Hupeh lies in a neutral soil zone between the pedocals (soils of arid or semiarid regions, enriched in lime) of North China and the pedalfers (soils of humid regions, enriched in alumina and iron) of the South. The uplands are mainly brown mountain earth, the lower hilly lands yellow-brown soil, and the lowlands alluvium and red earth.

*Climate.* Hupeh's rainfall follows the general Chinese seasonal pattern, governed by the rain-bearing monsoon winds. Han-k'ou has an average minimum rainfall of about an inch (25 millimetres) in December and a maximum of almost 10 inches in June, with a total annual fall of about 50 inches. Rainfall throughout the province decreases from southeast to northwest. Much of this rainfall is caused by cyclones, which pass down the Yangtze Valley from west to east. Occasionally, such as in the summer of 1931, a series of cyclones passes down the valley in rapid succession, bringing phenomenal rainfall and disastrous flooding. Since 1949 considerable effort has been directed toward flood control measures in the valley.

Cyclones

Hupeh winters, although usually short, are often rigorous, with heavy and glazed frost and some snow brought by bitter north winds in January and February, when the average temperatures are 40° F (4° C) and 43° F (6° C), respectively. Summers are hot, with July temperatures averaging 85° F (29° C), and are long and oppressive because of the high relative humidity. Any light breeze by day tends to die out in the evening, leading to intolerable nights when mothers bring their bamboo beds into the streets and sit fanning their children through the weary hours. There are about 270 frostless days in the south and 250 in the north.

*Plant and animal life.* The natural vegetation of Hupeh is dense forest, but this was cleared from the lowlands and hills many centuries ago, leaving only the western highlands densely wooded. The forests and woodlands consist mainly of *ma wei* (pine), *shu mu* (cedar), camphor, yellow sandalwood, maple, and poplar. As a result of deforestation, soil erosion has been serious. Despite sporadic efforts to plant the hillsides with trees during the early decades of the 20th century, the poverty of the people and the demand for fuel led to the continued stripping of trees. Since 1949 there has been a determined effort at afforestation and its maintenance.

There is a sparsity of large wild animals. Some small barking deer are found in the scant cover on the hills rising from the plain. Deer and wild pig are plentiful in the wooded mountains in the west. There is abundant birdlife, including wild duck and pheasants.

*Settlement patterns.* Population distribution, as in the rest of China, is predominantly rural. The main concentrations of rural population are found in the lake plain around the Yangtze and lower Han, notably from Wu-han downriver to Huang-shih; the lower Han Basin below Chung-hsiang (former An-lu) to Wu-han; and between I-ch'ang and Sha-shih. There is a smaller concentration at Hsiang-fan at the confluence of the Han and Tang rivers. Here the villages are often strung along high, mud riverbanks, which give safety in time of flood. Villages are small, usually 10 families or less, and are usually about one mile or so apart. Urban population is concentrated in a few large towns and a large number of small ones. The conurbation of Wu-han is the second largest industrial and commercial centre in the Yangtze Basin. Other large towns are Huang-shih, Sha-shih, I-ch'ang, and Hsiang-fan. In the past, many of the larger towns were walled; many of these walls have been demolished and the stone used for building and road construction.

Population centres

**The people.** Hupeh's ethnic composition is homogeneous, being overwhelmingly Han (Chinese). Their dialect is closely akin to pure Mandarin. Most of the minority peoples are Hui (Chinese Muslims), widely scattered throughout north Hupeh and the Han plain. There are some T'u-chia and Miao people in the highlands of the southwest.

**The economy.** *Agriculture.* Hupeh is located in the agricultural transition zone between the wheat-growing North and the rice-growing South; it is one of China's leading rice-producing provinces. In south and southeast Hupeh, where rainfall is greater and irrigation more easily practiced, most of the cultivated land is devoted to rice growing. In northern parts, where rainfall is less and variability greater, rice occupies less of the cultivated area and wheat much more. Most of the paddy area is planted with a single crop—middle-season rice (rice planted in the middle of the season after winter wheat or barley has been harvested)—newer strains of which have a growing period of only 90 days. Winter crops grown on paddy fields are usually wheat, barley, and broad beans. Irrigation in the hilly lands is predominantly by means of gravity from ponds dammed higher up in the valleys. On the plains, where water has to be raised, wooden paddle pumps operated by hand are still used, but electrical pumping stations are rapidly replacing human labour. Food production decreases rapidly westward, where cultivation is confined mainly to deep valleys in the highlands.

© Anne Sager 1979—Photo Researchers, Inc.



Steel-rolling mill at the Wu-han Iron and Steel Corporation in Hupeh Province.

Cash crops

Hupeh ranks high among the Chinese provinces as a producer of cash crops, of which cotton is the most important. The main growing area lies north of the Yangtze in a belt stretching from Sha-shih eastward along the lower Han to Wu-han. Other important economic crops are vegetable oils (sesame, peanut [groundnut], and rapeseed) and fibres (ramie and hemp). Ramie is the fibre from which grass cloth or China linen is made. Some tea is grown on the hills in the southeast. Tung oil, a valuable forest product used in paints and varnishes, comes mainly from the western regions and the upper reaches of the Han and Yüan rivers.

*Industry.* Hupeh's mineral wealth consists chiefly of iron, copper, and phosphorus ores; coal; and gypsum. Some of China's richest and best iron ore is found at Ta-yeh in southeastern Hupeh. The exploitation of this ore and of coking coal from P'ing-hsiang in Kiangsi was the basis for the founding of an ironworks at Han-yang at the end of the 19th century. Ore from Ta-yeh and other mines was also the basis for the establishment of the Wu-han Iron and Steel Corporation, one of China's largest integrated ironworks. Completed in 1961, these works cover about four square miles and contain the largest blast furnaces in China. The Hunan No. 2 Auto Plant in Wu-han is a major national supplier of jeeps and small trucks. Huang-shih has also developed as a large iron and steel centre. Copper is found at Yang-hsin in the east and also at Ta-yeh. Reserves are large compared to those of other provinces, and production has increased considerably. Bituminous coal is found in the west and anthracite (hard coal) in the south and east. There are large reserves of gypsum and salt in the northeast, and a number of foreign-built plants produce chemical fertilizers.

With the restoration of its traditional role as a national centre of trade and transportation, the tri-city area of Wu-han has come to play an important role in the province's economic development. Wu-han is the second largest industrial and commercial city in the Yangtze Basin. Of its three constituent cities, Han-k'ou is the commercial and industrial centre; Han-yang, formerly residential, is now largely industrialized; and Wu-ch'ang is the administrative, educational, and cultural hub of the province. In 1983 the conurbation was given economic power on a level with the provincial government.

*Transportation.* For more than 2,000 years waterways have been the main means of communication in Hupeh. Wu-han, known historically as "the thoroughfare of nine provinces," is the largest inland port in the country. The Yangtze and Han rivers, with their tributaries, are used by all manner of craft. Large ocean freighters reach Han-k'ou, and small steam launches penetrate much farther inland. Steam and oil-fired craft, however, carry only a small portion of the total waterborne freight: huge coaster junks from Chekiang and Fukien provinces sail to Sha-shih and I-ch'ang, and the small stern-oared *hua tzu*—each rowed from the stern by one man—ply the smaller streams. In addition to the rivers, the lake plain is a network of drainage channels that are used for communication by the local people.

Railways

Until 1958 Hupeh's railways consisted entirely of the Peking–Han-k'ou and Wu-ch'ang–Canton line, which ran from north to south across the province. Because of political unrest, corruption, and lack of funds, by 1949 the Peking–Han-k'ou line was in a parlous state; rapid repair work was carried out by the Communist government. In 1959 the completion of the bridge over the Yangtze between Han-yang and Wu-ch'ang—the first bridging of the river over a length of 3,400 miles—wrought a revolution in the system by greatly increasing the value and efficiency of the whole north–south line from Peking to Canton.

More than half of the pre-1949 road network was rendered unusable by the Sino-Japanese War and subsequent civil war. Since 1949, much reconstruction and repair work has been done, and new roads have been built. Wu-ch'ang has become an important centre for air traffic, second only to Peking. Air services, formerly entirely under central government control, have been supplemented by a regional carrier.

**Administration and social conditions.** *Government.* From 1950 to 1954 Hupeh was part of the Central South greater administrative region. In 1954 provincial (*sheng*) government was established directly under the central government. In 1958 local government was greatly modified by the formation of communes, which took over the duties of the rural districts and market towns (*hsiang* and *chen*) and were made responsible for the functioning of all local life at this level. During the early years of the Cultural Revolution (1966–69) Hupeh was governed by a Revolutionary Committee composed of party cadres, the army, and the revolutionary mass organizations. The Revolutionary Committee was replaced in 1980 by the People's Government, which is the administrative arm of the People's Congress. The commune system was abolished in the 1980s, and the township government pattern of the 1950s was reestablished. Hupeh has six prefectures (*ti-ch'ü*), eight prefecture-level municipalities (*shih*), and one autonomous prefecture (*tzu-chih-chou*). Below that level are counties (*hsien*) and county-level municipalities (*shih*).

Provincial subdivisions

*Education.* The educational pattern in Hupeh is similar to that in the rest of the country. From 1949 onward determined efforts have been made to overcome illiteracy. By 1970 it was estimated that nearly two-thirds of the people were literate, and this percentage has steadily grown. Wu-ch'ang, which was the early capital of the ancient province of Hukuang, has remained the educational and cultural heart of Hupeh. Under the Nationalist (Kuomintang) government (1928–49), a national university was built on one of the three large lakes outside the old walled city, and a Christian university was established inside the city itself. After 1949 both these institutions were incorporated into the new educational system, which now includes more than 50 institutions of higher learning.

Medical service

*Health and welfare.* Before 1949 there were large, efficient modern hospitals, run by Christian missions and secular bodies in both Han-k'ou and Wu-ch'ang; good though many were, they were inadequate to meet the needs of the rural people. Insofar as rural needs were met at all, they were served by medical missionaries and nurses, scattered sparsely throughout the province, as well as by Chinese doctors, herbalists, and acupuncturists. From the 1950s the city hospital services were greatly enlarged, offering a choice of Western or Chinese medicine; most attention has, however, been paid to public health and to preventive medicine. Debilitating diseases such as schistosomiasis (a parasitic disease) and malaria were attacked; drinking water and the proper disposal of sewage were supervised; standards of personal hygiene and of the cleanliness of streets and public places were raised. These measures, and the equitable distribution of food, have served to improve health and increase production.

**Cultural life.** In common with all other provinces, Hupeh has experienced considerable change in its cultural life since 1949. The great extension of education and the increase in literacy have had a far-reaching effect. In the cities, museums and libraries have been opened and are much patronized. Large stadia, sports halls, and swimming pools have also been built. The theatre still retains great popularity, particularly the regional operatic form known as Chu opera.

The rural areas, no less than the towns, have undergone great cultural change. Electricity has been extended to villages and hamlets. Every village of any size now has its own stores, its library, and its hall, in which meetings are convened, health clinics are held, and table tennis is played. Being probably better lighted than individual homesteads, the hall has become the place where villagers assemble to chat or listen to the radio. Storytelling—an agelong profession, which is still very popular—serves to preserve folklore. Country life is enlivened by occasional visits of professional players, entertainers, and acrobats.

## HISTORY

When China was slowly evolving in the Honan–Shansi region during the Shang and Chou dynasties (18th to 3rd century BC), Hupeh formed part of the kingdom of Ch'u. It was subjugated by Shih huang-ti (reigned 221–210/209 BC), who created the first united empire of China; it was finally assimilated into the Chinese state under the Han

dynasty (206 BC–AD 220). Hupeh at that time was described by the ancient Chinese historian Ssu-ma Ch'ien as:

> a large territory, sparsely populated, where people eat rice and drink fish soup; where land is tilled with fire and hoed with water; where people collect fruits and shellfish for food; where people enjoy self-sufficiency without commerce. The place is fertile and suffers no famine and hunger. Hence the people are lazy and poor and do not bother to accumulate wealth.

From this time on, the facility of communications afforded by its river system has caused Hupeh to figure prominently in Chinese history.

Since the mid-19th century it has been the centre of many momentous events, sometimes to its sorrow. The Taiping Rebellion, led by a Hakka, Hung Hsiu-ch'üan, broke out in Kwangsi in 1850, after which the rebel armies moved north, taking Wu-ch'ang in 1853. During the succeeding 10 years the central plains of Hupeh and Hunan were devastated by fighting and banditry. After China's defeat in the second Opium, or Arrow, War of 1856–60, the Hupeh cities of Han-k'ou, I-ch'ang, and Sha-shih were opened to Western nations as commercial ports. From this time on, European influence in central China steadily increased. Han-k'ou became the head of international oceangoing traffic. In the first 20 years (*i.e.*, until 1880) trade was based almost exclusively on tea, but, with increasing Indian and Ceylonese competition, Han-k'ou became the centre for the collection and processing of other central Chinese raw materials, notably vegetable oils, egg products, and tobacco.

Early role as a trade and industrial centre

Hupeh's industrialization began with the establishment of the Han-yeh-p'ing ironworks in Han-yang by Chang Chih-tung, the governor of the province, who also established a cotton mill in Wu-ch'ang opposite Han-k'ou. The ironworks had a checkered career. At first it enjoyed some government protection and tax exemption but later suffered from internal political unrest and instability, lack of capital, and poor management. Subsequently a Japanese concern gained financial control with a view to securing ore from Hupeh for its ironworks in Japan. The Han-yang works were allowed, even induced, to fall into decay. They were destroyed by bombing during the Sino-Japanese War of 1937–45 and were restored only after the advent of the Communist government in 1949.

The Chinese Revolution of 1911–12 began in Hupeh. The army in Han-k'ou mutinied, and the soldiers, led by their commander, Li Yüan-hung, took the cities of Han-k'ou, Han-yang, and Wu-ch'ang. Yüan Shih-k'ai led his northern troops, on behalf of the Emperor, against them and retook Han-k'ou but was unable to cross the Yangtze and eventually retired. This was the only significant fighting during the revolution. The province was ruled by a warlord (*tu-chün*) from 1916 to 1927, but from 1928 to 1938 there was some attempt at local government of a democratic Western pattern. When Nanking was taken by the Japanese in 1937, Han-k'ou became a temporary headquarters for the Nationalists; after the Nationalist retreat to Chungking in 1938, much of Hupeh came under Japanese control. A period of near chaos after the Japanese defeat in 1945 ended with the establishment of the People's Republic in 1949. (T.R.T./V.C.F./Ed.)

## Kiangsu

A province on the east coast of China, Kiangsu (Chiang-su in Wade–Giles romanization, Jiangsu in Pinyin) is bounded by the Yellow Sea and by the provinces of Chekiang to the south, Anhwei to the west, and Shantung to the north. It occupies an area of 39,600 square miles (102,600 square kilometres). The provincial capital is Nanking, which was the southern capital of China during the Ming dynasty (1368–1644) and the capital under the National Government (1928–49). Kiangsu became a separate province in 1667 (the sixth year of the reign of the K'ang-hsi emperor). The name is derived from the prefixes of Chiang-ning and Su-chou, the names of the two most important prefectures within the province at that time.

The province consists almost entirely of alluvial plains divided by the estuary of the Yangtze River into two sections, Chiang-nan (literally, "South of the River") and Su-pei (northern Kiangsu). Chiang-nan is fertile and well watered, famed for its silk and handicrafts, and very densely populated and industrialized. The cities of Su-chou, Nanking, Wu-hsi, and Shanghai are all located in this region. Shanghai is situated at the mouth of the Yangtze River, although administratively the Shanghai Municipality is not a part of Kiangsu Province but is controlled directly by the State Council of the central government. Su-pei is relatively poor in comparison with Chiang-nan. The northernmost section of Su-pei, from Suchow to the sea, is actually part of the great North China Plain in its physical geography, as well as in its agriculture and general way of living; it is the poorest section of Kiangsu and is densely populated.

**The land.** *Relief.* The most important physical characteristic of the province is its wide alluvial plain, stretching from north to south, at a low elevation above sea level. Most of the soils are thus alluvial, both calcareous and noncalcareous, and including some saline soils. There is an intricate network of rivers and canals, lakes and ponds, all protected from floods by dikes. Most of the province is less than 150 feet (45 metres) above sea level. Hills of moderate elevation are found only in the southwestern corner of the province and in the extreme north along the Shantung border. Mount Yün-t'ai, in northern Su-pei near the Yellow Sea, is the highest point in the province, at 2,050 feet (625 metres). Nearly 10 percent of the total area is occupied by shallow lakes and reedy marshes. The silt of the great rivers encroaches constantly on the sea, leaving seaports of former ages dry. In coastal areas below the high-water level, cultivation is carried on in polders (areas protected from the sea, mainly by dikes). Extensive canalization and a vast development of polders have been systematically carried out during the 20th century. This section of the surface of the Earth has been completely altered by human hands.

*Drainage.* Chiang-nan is drained primarily by the Yangtze River, which enters the province to the southwest of Nanking on the Kiangsu–Anhwei border and flows generally east and southeast before reaching the East China Sea. The waters from upstream meet tidewaters at Nanking. The river becomes broader at Chen-chiang, widening to more than 11 miles at Nan-t'ung and more than 56 miles at its mouth. It carries an enormous load of silt to the sea annually, depositing it to form the Yangtze Delta. Tides and currents carry some of the sediment to form sandbars in the estuary and along the coast. The delta itself grows at an average rate of about 82 feet a year.

Su-pei's major drainage systems are Hung-tse Lake and the Huai River, which flows into the lake; Kao-yu Lake, through which waters from Hung-tse Lake reach the Yangtze; the Su-pei Canal, which drains Hung-tse Lake; and the Grand Canal, which runs through the entire province from north to south and connects Su-pei with the Yangtze Delta. During several periods in Chinese history, northern Kiangsu was also drained by the Huang Ho (Yellow River), which occasionally left its course and flowed into the Huai. Formerly, the Huai flowed into the sea, but when its channel was gradually usurped by the Huang, beginning more than a thousand years ago, it was unable to reach the sea and instead emptied itself into Hung-tse Lake.

The Kiangsu lowlands are floodplains formed by the alluvial deposits of the Yangtze, Huai, and Huang rivers and their tributaries. Using the Yangtze and the old channel of the Huai as convenient landmarks, the area of these plains may be divided into three sections.

The three sections of the plains

The Chiang-nan plain south of the Yangtze forms the principal part of the Yangtze Delta, characterized by flatness and lying only 10 to 16 feet above sea level. It is crisscrossed by streams and canals and dotted with ponds and lakes, forming an elaborate network of flowing water, meticulously maintained by farmers. This area actually has the highest stream density in China: within it, no place is more than 300 feet from the drainage system of T'ai Lake (the southern shore of which forms much of the Kiangsu–Chekiang border). The canals were all dug by farmers of the area. Isolated hillocks dot the edge of the T'ai Lake area, which adds to its enchanting beauty.

The lakes were parts of former shallow bays and inlets of the sea, obstructed and enclosed by the steady advance of the Yangtze Delta. After being cut off from the sea, the water gradually decreased in salinity and formed freshwater lakes. T'ai Lake is connected with the Yangtze and its estuary by many distributaries. The Chiang-nan Canal (the name for the section of the Grand Canal south of the Yangtze), which runs through the full length of the T'ai Lake plain from northwest to southeast, cuts across all the distributaries connecting the T'ai Lake basin with the Yangtze, thus forming a vital link of the T'ai Lake system.

Between the Yangtze and the ancient channel of the Huai is what Chinese geographers call the Yangtze–Huai plain, built by the alluvium of the two rivers. The centre of this plain is only 6½ to 13 feet above sea level, while its periphery stands at about 17 to 33 feet. It is considered to be a section of the Yangtze Delta, as it has the same topographical elements, including alluvial deposits and drainage. As a sluggish tributary of the Yangtze, the Huai formerly caused widespread floods during the high-water season, but a water-control project has permanently restricted the high waters of the Huai.

North of the old channel of the Huai is the Suchow–Huai plain, built of the alluvium of the Huai and Huang rivers and standing about 30 to 150 feet above sea level. In the northern part of the plain are low hills with heights of about 650 feet.

Yangtze Valley and North China climates

*Climate.* Within the province, two subtypes of climate may be distinguished: the Yangtze Valley climate, in central and southern Kiangsu, and the North China climate, to the north of the old Huai River. The former is humid subtropical, while the latter is cool, temperate continental, with greater extremes of temperature. Nanking in the south has a mean temperature of 36° F (2.2° C) in January and 82.4° F (28° C) in July. For northern Kiangsu, the mean January temperature is below 32° F (0° C), but summer is as hot as in the south. Annual precipitation generally increases from north to south, ranging from 24 to 47 inches (600 to 1,200 millimetres), that of Nanking being 41 inches. Seasons are distinct in both north and south. Between spring and summer, the south receives prolonged rains of cyclonic origin, typical of the Yangtze Valley and extremely useful for rice growing. The coast is often visited by destructive typhoons between late summer and early autumn.

*Plant and animal life.* In Su-pei grow temperate broad-leaved deciduous trees, typical of the North China Plain, while in southern Kiangsu are found subtropical mixed broad-leaved deciduous and broad-leaved evergreen trees, typical of the Yangtze Valley. As the whole of the province has been cleared for cultivation since ancient times, no primary forest remains. In natural flora, it is a markedly depleted territory, because of the dense population and intensive cultivation. There is a warmth-loving and moisture-loving fauna characteristic of the monsoon climate of East Asia. The fauna has considerable economic significance, fish, ducks, crabs, and shrimps being important sources of food. Fish raising is highly developed—the numerous ponds, reservoirs, lakes, canals, and streams are stocked with hundreds of millions of fry that are shipped to other provinces and are also exported to other countries.

There are also numerous agricultural pests, such as rodents and insects, which harm cultivated plants and trees. Great strides have been made in the control of the more common pests, but the insects that damage trees have not yet been brought under complete control.

*Settlement patterns.* Although almost 90 percent of Kiangsu's population is rural, the province contains many of the largest cities of the Yangtze Delta. The population distribution patterns of Kiangsu Province and the municipality of Shanghai are inseparable geographically and economically. Population density is higher in the north of the province, a fact explained by its earlier development, which dates from ancient times, and its importance as a communication link between North China and the Lower Yangtze Valley. Even the hilly district in southwestern Kiangsu has very high population densities in comparison with Europe and the United States. Shanghai is the largest municipality in China and one of the 10 largest in the world; it is not administratively a part of Kiangsu, being controlled directly by Peking. Other large cities in the region are Nanking, which is the largest city of Kiangsu proper and is its administrative and cultural centre; Su-chow (Hsü-chou), in northern Kiangsu; Su-chou, east of T'ai Lake; and Wu-hsi, in Chiang-nan.

The villages are distributed very close to one another on the Yangtze Delta, generally less than one-third of a mile apart. They are located mostly on the banks of rivers and canals. Villages with several scores of households are the most common. Communication between villages is usually very easy, thanks to canals and barges, rural roads, and the ubiquitous bicycle, somewhat as in the Low Countries of Europe. The houses are usually well built of brick baked in local kilns. Dwelling conditions are fair to good by Chinese standards.

Dispersed rural settlement is the rule along the coast and the rivers of northern Kiangsu. Dwellings are found singly along the riverbanks and quite close to each other in groups of two, three, or four among the rice fields.

**The people.** The population of Kiangsu is entirely Han (Chinese), with the exception of a few Hui (Chinese Muslims). The inhabitants of Chiang-nan speak the Wu (Shanghai) dialect, while those of northern Kiangsu and the Nanking area speak the eastern Mandarin dialect.

**The economy.** *Agriculture.* Kiangsu is one of the richest provinces in China, with a significant agricultural sector. Output is enhanced by multiple cropping, powered irrigation, tractors, and chemical fertilizers. The T'ai Lake plain produces rice, wheat, cotton, fruit, silk, tea, and fish. Tea is grown in the southwestern uplands around I-hsing, which produces the famous I-hsing china tea sets. Cattle, pig, and poultry raising are an important source of food and income, especially since the number and size of private plots allowed to each household increased. Fishing and pisciculture are other sources of food.

Crops

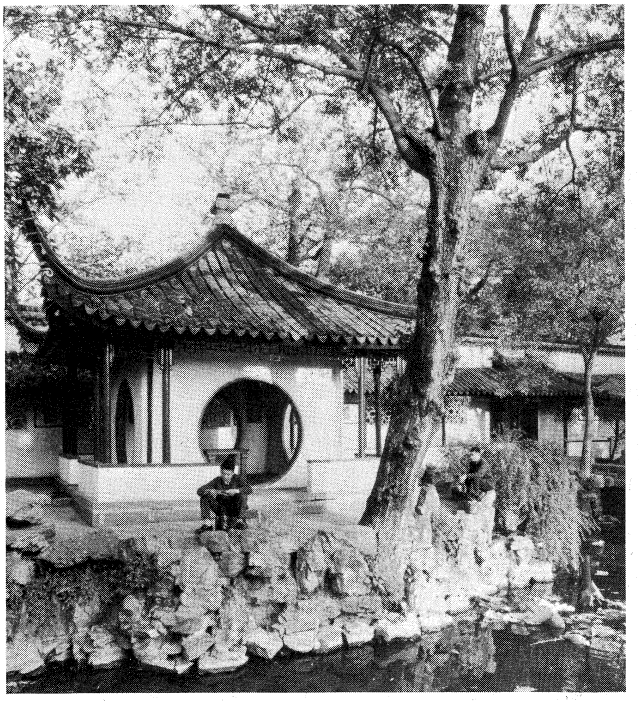
*Industry.* Nanking is the most important heavy industrial centre of Kiangsu proper. Major industrial plants produce trucks and parts for motor vehicles, chemical fertilizers, and detergent raw materials. The Nanking industrial area also produces steel, petrochemicals, electronics, machine tools, cameras, textiles, cement, and sundry building materials. Wu-hsi, near Shanghai, is a rising industrial centre with good inland waterway connections to all parts of the province. Modern manufactures include machine tools, agricultural and transportation equipment, cotton textiles, silk reeling, and food processing. Good deposits of iron and coal have been found at I-hsing and are used in a local ironworks and steelworks.

Kiangsu has become a major exporter since the 1970s. Goods formerly shipped through Shanghai are now handled through the provincial ports of Nan-t'ung, Chang-chia-kang, and Lien-yün-kang. Two of these ports have been designated "open" cities and encouraged to foster foreign trade and investment: Nan-t'ung, with 16 miles of deepwater frontage on the Yangtze, has developed its own economic and technical investment zone; and Lien-yün-kang, as the eastern terminus of the Lung-hai Railway, is a key export outlet for the central and northwestern provinces along the rail line. Kiangsu was incorporated into a larger Shanghai special economic zone in 1984.

*Transportation.* Among the assets of the province is the dense water transport network. With more than 14,000 miles of inland waterways, Kiangsu carries more than 50 percent of its goods by water. In contrast, its railroads carry only about 20 percent of the province's freight shipments. The completion in 1968 of the Yangtze rail and highway bridge at Nanking made the city a key north–south and east–west communications hub. The Grand Canal, which is periodically redredged, continues to play an important role in north–south transport.

**Administration and social conditions.** Kiangsu is divided into 11 prefecture-level municipalities (*shih*). The province also contains municipal districts (*shih-hsia-ch'ü*) and municipal counties (*shih-hsia-hsien*). The provincial capital, Nanking, is the military regional headquarters for eastern China.

Kiangsu has a rich educational tradition, with some of China's top arts and science universities, and a scientific



The Moon Sighting Pavilion in the Humble Administrator's Garden in Su-chou, Kiangsu Province.
Peter Carmichael—Aspect Picture Library, London

and technical work force that forms the basis of one of the country's centres of technology and research. With a predominantly rural population, however, the province's illiteracy levels are slightly higher than the national average. The proportion of the provincial population with a primary level of education or higher is just at the national average. Since 1949 health care has expanded greatly.

Two categories of cities

**Cultural life.** The cities of the province fall into two categories based on the standpoint of historical development—the ancient cities and the modern cities. The former date from ancient or medieval times and include Nanking, Su-chou, Yang-chou, Chen-chiang, and Suchow. Several of them are well known in East Asian history, are rich in cultural heritage, and have a long tradition that has found artistic expression in Chinese traditional architecture, painting, sculpture, flower gardens, stone bridges, and world-renowned handicraft industries, such as silk embroidery and carving of various materials. These cities often possess historical monuments, famous temples, and local shrines and *p'ai-lou* (arches) honouring their illustrious citizens. Many cities have a rich folklore. Nanking, especially, abounds in national monuments and famous historical relics. The most renowned are the simple Ming tombs (of Ming emperors) and the magnificent Sun Yat-sen Mausoleum, at the foot of Tzu-chin Hill. The gastronomic specialty of this ancient capital is the renowned Nanking salted duck. The duck is raised in ponds and lakes nearby. Other products from the Nanking area include handwoven silk (*tzu-ching*), particularly cloud brocades, which use every conceivable shade of colour to portray the clouds of sky at sunset.

The modern cities that sprang up in the 19th century after the Opium Wars and the Treaty of Nanking (1842), which opened China to international trade, include Shanghai, Wu-hsi, Nan-t'ung, and Lien-yün-kiang. Most of them are seaports, river ports, or railway junctions.

HISTORY

The state of Wu

In antiquity, the Kiangsu region was within the jurisdiction of the ancient state of Wu. During the Chou dynasty (*c.* 1111–255 BC) much of the area was called Kou-wu and was considered outside Chinese borders. During the Ch'un-ch'iu (Spring and Autumn) and part of the Chan-kuo (Warring States) periods from the 8th to the 3rd century BC, it was brought into the Chinese empire as one of the "outer states." Known as the Wu region during the Han dynasty (206 BC–AD 220), it became the independent state of Wu during the succeeding San-kuo (Three Kingdoms) period, with its capital at Chien-yeh, the site of modern Nanking. The golden age of culture in the region was during the Six Dynasties (AD 220–589), when it received a major influx of immigrants from the north. After the fall of the T'ang dynasty in 907, Yang-chou in Kiangsu became the capital of the Nan (Southern) T'ang state, which lasted from 937 to 975/976.

Another period of major cultural and commercial development occurred during the Nan Sung dynasty (1127–1279). In the early Ming dynasty Nanking became capital for the entire empire, and even after 1420, when the Ming capital shifted to Peking, Nanking remained as subcapital for South China. During the Ming and Ch'ing, or Manchu, dynasties, Chiang-nan was a major rice surplus region, supplying 40 percent of tribute tax grain to the capital by means of the Grand Canal. Chiang-nan merchants were among the most influential in China during this period. In the mid-19th century there was significant foreign commercial intervention, based on treaty port privileges. The region was seriously affected during the Taiping Rebellion of the 1850s and '60s, and Nanking became the Taiping capital in 1853, remaining under Taiping control until 1864.

In the 20th century Kiangsu became an important power base for the Nationalist Party (Kuomintang) of Chiang Kai-shek, and Nanking was made capital of the Nationalist government in 1928. It remained the puppet government capital under the Japanese occupation, after the Nationalist government moved to Chungking. During World War II the region was the locus of Communist-led guerrilla forces of the New 4th Army, from whose ranks many of Kiangsu's post-1949 leaders came.

## Shanghai

For coverage of Shanghai, see the *Macropædia* article SHANGHAI.

# SOUTH CHINA

## Chekiang

Chekiang (Che-chiang in Wade–Giles romanization, Zhejiang in Pinyin), the third smallest province of China, is also one of the most densely populated and affluent. Its area is 39,300 square miles (101,800 square kilometres). A coastal province, it is bounded by the East China Sea on the east and by the provinces of Fukien on the south, Kiangsi on the southwest, Anhwei on the west, and Kiangsu on the north. The provincial capital is Hang-chou.

Chekiang has for many centuries been one of the great cultural and literary centres of China. Its landscape is renowned for its scenic beauty. The name of the province derives from its principal river, the Che ("Crooked") River, formally known as the Ch'ien-t'ang River at the estuary of Hang-chou Bay, and known as Fu-ch'un River inland. Chekiang is among the leading Chinese provinces in farm productivity and leads in the tea and fishing industries.

PHYSICAL AND HUMAN GEOGRAPHY

**The land.** *Relief.* The northwestern section of the province (Che-hsi) lies within the fertile Yangtze River Delta, with its labyrinth of rivers and canals; its coastal lowlands are protected by dikes. The southern edge of T'ai Lake forms part of its northern border with Kiangsu. The greater part of Chekiang Province lies to the south of Hang-chou Bay and is largely mountainous. It has a rocky and deeply indented coast, dotted with more than 18,000 islands, forming numerous natural harbours. This part is in fact a continuation of the mountain ranges of

The delta and the mountains

Fukien, which run roughly parallel to the coast. In eastern Chekiang, mountains occupy 93 percent of the land surface, while another 1 percent consists of low hills. Only 6 percent is level land, distributed along Hang-chou Bay and the Fu-ch'un and Ta river valleys in southern Chekiang. Most of the province's arable lands—consisting of alluvial plains of great fertility—are found in these three areas.

*Drainage.* The chief river of the province is the Fu-ch'un (Ch'ien-tang) River, the drainage basin of which constitutes 40 percent of the total area of the province. The river has, in fact, two headstreams, one coming down from the southwestern highlands and flowing through the broad Lan River valley and the other rising in Anhwei Province and passing through Chien-te in Chekiang and other cities. On the latter tributary is located the Hsin-an River Dam and hydroelectric power plant, which is one of the largest in East Asia. The Ch'ien-t'ang Estuary tidal bore takes the form of a high wall of water that rushes upstream with a thunderous roar. Best seen just after the full moon and at its highest in the autumn (late September–early October), the bore is a famed tourist attraction. Along the estuary are miles of sea dikes that have been built throughout the ages to protect the rich rice lands of the delta. The other rivers of some importance are the Yung, Ling, and Ou; the Ou and its four principal tributaries together form the second largest river system of the province. Although these mountain streams flow swiftly through rocky channels and gorges, they are navigable to skillful boatmen using sampans (small, roofed boats propelled by sculling) right up to the mountains.

Coastal islands

Hang-chou Bay is almost as broad at its entrance as the Yangtze Estuary but is obstructed by a cluster of some 400 islands known as the Chou-shan (Chusan) Archipelago. The largest island in the group, Chou-shan, is a major coastal fishing centre. On P'u-t'o Mountain, a renowned scenic island east of Chou-shan, is one of the sacred mountains of Buddhism that once attracted pilgrims from all over East Asia.

*Climate.* Chekiang has a humid subtropical climate, controlled chiefly by monsoonal airflows, modified by local influences. Considerable differences exist between the coast and the hinterland, between the lowlands and the highlands, and between the north and the south, particularly in winter. Thus, Hang-chou, in western Chekiang, has an average January temperature of 39° F (4° C), while that of Wen-chou, on the coast, is about 46° F (8° C). Summer is hot throughout the province; the average July temperature at Hang-chou is 82° F (28° C), while that at Wen-chou is 84° F (29° C). Annual rainfall throughout the province is more than 40 inches (1,000 millimetres). The hilly interior has more precipitation than the coast, which is frequently visited by devastating typhoons, particularly during late summer and early autumn.

*Plant and animal life.* The vegetation of the northern, or T'ai Lake, plain differs from that of the rest of the province. Formed from a lake, it is covered with rich alluvial soil and is an open land of rice fields and rural settlements, dotted with some shade and ornamental trees. The original or natural vegetation disappeared centuries ago when the land was cleared for cultivation.

Mountain vegetation

The vegetation of the hilly and mountainous parts of the province, south of the plain, consists primarily of mixed evergreen broad-leaved and coniferous forests that grow on gray-brown podzolic (infertile forest) soils at the higher elevations and on red and yellow lateritic (leached, iron-bearing) podzolic soils on the lower slopes. There is an abundance of such trees as the laurel, pine, cypress, and beech. Besides the ubiquitous bamboo, the tung tree, which supplies valuable oil, is widely distributed in the upper Fu-ch'un Valley.

The province has an animal life typical of the subtropical forest zone and characterized by great diversity; it includes monkeys, anteaters, the southern heron, water turtles, many frogs, and numerous southern birds. There are many invertebrates, among which subtropical insects predominate, although tropical insects characteristic of southern Asia are also found.

*Settlement patterns.* There is a marked contrast between the densely populated plains and the sparsely populated uplands. Thus, two-fifths of the population in the province is concentrated in the T'ai Lake plain and in the coastal region of Hang-chou Bay. About 25 percent of the population lives in cities and towns. The capital and largest city is Hang-chou, located in western Chekiang; it is followed in size by the port cities of Ning-po and Wen-chou, both in eastern Chekiang. Other important cities are Shao-hsing and Chin-hua, in eastern Chekiang, and Chia-hsing and Wu-hsing (locally known as Hu-chou), in western Chekiang. All of these urban centres have a long history; the oldest, Shao-hsing, dates to the 6th century BC. Hang-chou was the capital of the Chinese empire during the 12th and 13th centuries. It was, however, only after the first Opium War (1839–42) and the opening of Ning-po to foreign trade that the modernization of the cities—particularly Hang-chou, Ning-po, and Wen-chou—began. Scores of other towns are distributed throughout the province. They include the county (*hsien*) capitals, which are located mostly on the agricultural plains and valley bottoms. Most of them are also local commercial centres, and some are developing into larger and more modern towns.

**The people.** The ethnic composition of the population is overwhelmingly Han (Chinese). Those belonging to ethnic minorities consist chiefly of She tribesmen living in the mountainous area of southern Chekiang, in the Wen-chou and Chin-hua prefectures along the Fukien border. The She tribesmen, of whom a greater number live in Fukien Province, have their own language, although most of them also understand Chinese. They grow paddy rice in terraces on hillslopes; farm work is done by both men and women. There are also small numbers of Manchu and Hui (Chinese Muslims) scattered in the cities and towns. The former are mostly descendants of Manchu soldiers garrisoned in Chekiang before the overthrow of the Manchu dynasty in 1911/12. With the exception of the Muslims and some Christians, the religious affiliation of the entire population in the province may be characterized as a complex of Taoism, Confucianism, and Buddhism. This form of religion is generally tolerated by the People's Republic, though the monks are required to engage in productive work to earn a living.

Religious affiliations

**The economy.** *Agriculture.* Chekiang is one of the more prosperous of China's provinces, leading the country in farm productivity in the tea industry and second only to Szechwan in sericulture (the raising of silkworms to produce raw silk). Its agriculture is among China's most diversified, with less than half its farm output by value coming from food or cash crops.

Because of the province's hilly topography, only about

Emil Schulthess—Black Star



Picking Lung-ching (Dragon Well) green tea near Hang-chou in Chekiang Province.

one-fifth of its land surface is arable. Two-fifths of the cultivated land lies in northern Chekiang, in the Yangtze Delta and on the southern shore of Hang-chou Bay. About four-fifths of Chekiang's arable land is irrigated—one of the highest ratios in eastern or southern Asia—and about two-thirds of the arable land is used to grow staple food crops—rice, wheat, barley, corn (maize), and sweet potatoes. The rest of the farmland grows either green fertilizer crops or such industrial crops as cotton, jute, ramie (a shrub yielding a fibre used for textiles), rapeseed, sugarcane, and tobacco. Most farmers also raise pigs and poultry on their small private plots, and many also raise fish in village ponds, reservoirs, or lakes and rear silkworms during the slack farming season in spring. In the well-watered hilly areas, tea is grown. All these activities provide a second income for peasant households.

Rice is the chief staple food and is grown widely throughout Chekiang Province, although the well-watered northern plains constitute the most productive area of cultivation. Both single-cropping and double-cropping systems are followed in paddy (rice field) cultivation. Since 1949 double-cropping of rice has been vigorously promoted, and its share in the rice acreage has increased to one-half of the total.

Rice and tea cultivation

Chekiang has four principal tea districts. The Hang-chou district produces the famous Lung-ching (Dragon Well) green tea. The P'ing-shui district has the largest tea acreage and the highest production of processed tea. The other two districts are Chien-te, in the southwest, and Wen-chou, in the southeastern hilly region. World War II caused serious damage to the tea industry as tea gardens were abandoned and aging shrubs were not replaced. During the 1950s a systematic rehabilitation and development program was undertaken. Improved methods of tea cultivation and processing were introduced and new orchards established, and the province resumed its position as China's leading tea producer.

Sericulture is another of Chekiang's traditionally famous industries. The principal silkworm-rearing areas are on the T'ai Lake plain. Secondary districts are located in the northeast and the northwest. These areas, which have a long history of sericulture, yield a consistently high quality of silk from the cocoons. The industry, like the tea industry, suffered serious damage during World War II and the civil war that followed, but vigorous measures to restore production have raised output.

The Chekiang coast lies at the convergence of western Pacific warm and cold currents. Its rivers carry rich organic material into the shallow waters above the continental shelf. As a result, many kinds of fish come there to spawn. More than 100 varieties of fish are found there. Important commercial catches include drums (or croakers), cutlass fish, and cuttlefish. The rapid growth of fishing has required readjusting fishing quotas to protect the fishing banks from overexploitation. A flourishing aquaculture industry has been developed, producing kelp, the edible red algae Porphyra (used in making soups and condiments), shellfish, and other marine products.

Fishing and aquaculture

*Industry.* Most of Chekiang's wealth derives from light industry. This in part reflects the province's historic role as a commercial and handicraft centre and a significant textile producer since the 1890s.

The province has few exploitable minerals, although local low-grade coal deposits are mined and consumed in a number of locations. China's largest fluorspar mine is located in southwest Chekiang and has been worked since the early 1930s. Oil exploration has been undertaken in the East China Sea off Wen-chou. Industrial development has been stimulated by the growth of electric power generation based on Chekiang's fast-flowing rivers. The Hsin-an River hydroelectric plant is one of the largest in China.

Hang-chou has become a major industrial city since 1949 and produces a wide range of industrial and consumer goods, including machinery, textiles, agricultural implements, chemicals, radios, and televisions. Ning-po is also a major industrial centre, producing tractors, electronics, and petrochemicals. The province has become a major exporter with a number of specialized export centres for light industrial products and handicrafts. The designation of Ning-po and Wen-chou as two of China's "open" cities has stimulated the planning of foreign investment and technology transfer programs, and Chekiang has been included in the Shanghai special economic zone.

A flourishing handicraft industry is located mostly in rural villages. Nationally and internationally known products include the porcelain of Lung-ch'üan, the silk umbrellas and tapestry of Hang-chou, embroideries, laces, wood and stone carvings, inlay ware, and a host of other products of Chinese folk art.

*Transportation.* The rivers play an important role in the province's transport; about half of the total freight volume travels on these inland waterways. The remainder of the freight volume is moved mostly by road, though heavier goods are often moved by rail, especially for longer distances. Although there are numerous harbours along the Chekiang coast, coastal shipping accounts for only a small percentage of the total freight volume. The Shanghai–Hang-chou railway is the most important trunk line, connecting western Chekiang with east and North China. The Chekiang–Kiangsi line links Chekiang with South and central China. The Hang-chou–Ning-po railway connects the southern littoral of Hang-chou Bay with the Chekiang–Kiangsi and the Shanghai–Hang-chou lines. A modern highway network with its primary centre at Hang-chou connects the province with the cities of Shanghai and Nanking and with the provinces of Anhwei and Fukien.

**Administration and social conditions.** *Government.* Chekiang Province was governed as part of the East China greater administrative region from 1950 until it came under the direction of the central government in 1954. It is divided into four prefectures (*ti-ch'ü*) and six prefecture-level municipalities (*shih*). Below this are counties (*hsien*) and county-level municipal districts (*shih-hsia-ch'ü*).

From 1958 to the late 1970s the administrative unit below the county was the commune. All administrative units, from the province downward, were theoretically governed by assemblies elected through indirect elections but actually run by local party leaders. In 1980 the People's Government and People's Congress were created to take over functions from Cultural Revolution-era governments.

*Education.* Chekiang has a strong tradition of locally supported education. Its levels of adult literacy and primary-level educated citizens are above the national average. The province boasts more than 20 institutions of higher learning.

**Cultural life.** During the Nan (Southern) Sung dynasty (1127–1279) the political and cultural centre of China moved from the North to western Chekiang. The Hang-chou area became the homeland of a galaxy of famous painters (including a Sung emperor), as well as of calligraphers, poets, essayists, philosophers, and historians. The beauty of Lin-an (modern Hang-chou), the Nan Sung capital, was immortalized by the landscape painters Hsia Kuei and Ma Yüan. The cosmopolitan legacy has lingered on in provincial pride and national stereotypes that often depict the Chekiang people as both cultured and affected. Various national and regional operatic traditions flourish, including the famous Yüeh opera of Shao-hsing. There are many distinct regional subcultures, with their own musical and culinary traditions.

Role as cultural centre

## HISTORY

Before the 8th century BC western Chekiang was a part of the ancient state of Wu, while eastern Chekiang was the land of Yüeh tribes. In about the 6th century BC the two subregions became the rival kingdoms of Wu and Yüeh. The heartland of the Wu state lay in southern Kiangsu Province, whereas that of Yüeh occupied the coastal area to the south of the Ch'ien-t'ang Estuary where it merges into Hang-chou Bay. Yüeh and Wu engaged in constant warfare from 510 until 473 BC, when the Yüeh conquered Wu, after which the victorious kingdom became a dominant power in the Chinese feudal empire, nominally headed by the Tung (Eastern) Chou dynasty. Yüeh was itself subsequently subjected, first by the kingdom of Ch'u in 334 BC and then by the kingdom of Ch'in in 223 BC.

Yüeh (consisting of Chekiang and Fukien) was quasi-independent during the Han dynasty (206 BC–AD 220).

Chekiang later formed a part of the kingdom of Wu (220–280). During the T'ang (618–907) and Sung (960–1279) dynasties, Chekiang was divided into Che-hsi (Western Chekiang) and Che-t'ung (Eastern Chekiang), which became the traditional geographic divisions of the province.

Lin-an

Lin-an was made capital of the Chinese empire during the Nan Sung dynasty, and its population in 1275 was estimated at about 1,000,000. Marco Polo, who visited the city, described it as the finest and noblest in the world. Odoric of Pordenone also visited the city, which he called Camsay, then renowned as the greatest city of the world, of whose splendours he, like Marco Polo and the Arab traveler Ibn Baṭṭūṭah, gave notable details. Chinese, Mongols, Nestorian Christians, and Buddhists from different countries lived together peaceably in the city during this period. Hang-chou continued to be a great cultural centre until 1862, when it was destroyed during the Taiping Rebellion. Of its citizens, 600,000 were slaughtered, while the rest either drowned themselves or else perished from starvation and disease. Hang-chou did not fully recover from this disaster, but it was eventually rebuilt and underwent gradual modernization.

Foreign penetration of Chekiang began in the 1840s with the opening of Ning-po as a treaty port city. Ning-po merchants gradually established commercial networks in Shanghai and along the coast. In 1913 a railroad linking Hang-chou to Shanghai was built. During the Chinese Revolution of 1911–12 the moderate landed elite seized power, but the province soon fell into the hands of warlords and became in the mid-1920s the power base of Sun Ch'uan-fang. In the late 1920s the province became a base of power for the Nationalist Party (Kuomintang) of Chiang Kai-shek, who was born at Feng-hua near Ning-po. The Chekiang elite came to dominate the Nationalist regime, and the province benefited from modernization programs introduced between 1928 and 1937. The Japanese occupied much of Chekiang after 1938, but the harbour at Wen-chou remained in Chinese hands from 1938 to 1942.

(F.Hu./V.C.F.)

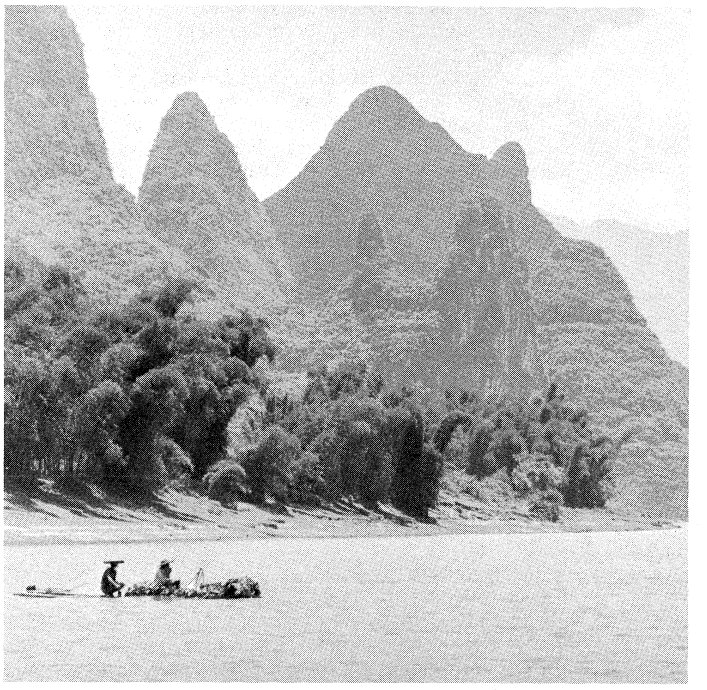
## Chuang Autonomous Region of Kwangsi

The Chuang Autonomous Region of Kwangsi (Kuang-hsi in Wade–Giles romanization, Guangxi in Pinyin) is bounded by the Chinese provinces of Yunnan on the west, Kweichow on the north, Hunan on the northeast, and Kwangtung on the southeast, and by Vietnam and the Gulf of Tonkin on the southwest. It covers an area of 85,100 square miles (220,400 square kilometres). Nanning, the capital, is about 75 miles (121 kilometres) southwest of the region's geographic centre. The name Kwangsi dates to the Sung dynasty (960–1279), when the region was known as Kuang-nan Hsi-lu, or "Wide South, Western Route" (western half of all territory south of the Nan Mountains). The Yüan dynasty (1206–1368) contracted the name to Kwangsi when it created a province out of the western half. In 1958 the province was transformed into the Chuang Autonomous Region of Kwangsi—a step designed to help foster the cultural autonomy of the Chuang, or Chuang-chia, people, who constitute the largest minority living in the region.

### PHYSICAL AND HUMAN GEOGRAPHY

**The land.** *Relief.* Kwangsi forms a tableland that descends in elevation from the north and northwest to the south and southeast. Elevations between 3,000 and 6,000 feet (900 and 1,800 metres) above sea level are reached at the edge of the Yunnan–Kweichow plateau in the northwest, the Chiu-wan and Feng-huang mountains in the north, and the Yüeh-ch'eng Mountains in the northeast. The greater part of the region is composed of hilly country lying at a height of between 1,500 to 3,000 feet. In the west, the Tu-yang Mountains rise to 6,500 feet. In the southeast, lowlands are situated at a height of between 300 and 1,500 feet.

The predominance of limestone gives many parts of Kwangsi a spectacular type of landscape known as karst, in which pinnacles and spires, caves and caverns, sinkholes, and subterranean streams abound. Picturesque rocky hills, spires of grotesque proportions, and strangely shaped caves with all types of stalactites and stalagmites are found in many parts of this region.



The karst landscape along the Li River, a tributary of the Kuei River, near Kuei-lin, Chuang Autonomous Region of Kwangsi.

© Lynn Lennon 1986—Photo Researchers, Inc.

*Drainage and soils.* The Ch'in and Nan-liu rivers flow into the Gulf of Tonkin. The headwaters of the Hsiang River flow into Hunan Province. The remainder of the region's numerous rivers—including the Hung-shui, Liu, Ch'ien, Yu, Tso, Yü, Hsün, and Kuei—follow the general southeastward slant of the terrain. They rise from a profusion of sources and flow into one another in a succession of convergences until they merge into one major river, the Hsi. This mighty river rises in Yunnan Province and cuts across the entire width of Kwangsi before emptying into the South China Sea near Canton in Kwangtung Province.

The Hsi River Basin

The hilly areas are composed of red soil, while the lowlands are characterized by alluvial soil brought down by the many rivers.

*Climate.* Throughout the region, temperatures are warm enough to assure agricultural production throughout the year. The summer, which lasts from April to October, is marked by enervating heat and high humidity. Winters are mild and snow is rare. July temperatures vary between 80° and 90° F (27° and 32° C), while January temperatures range between 40° and 60° F (4° and 16° C).

Because of the influence of the rain-bearing monsoon wind, which blows from the south and southwest from late April to the end of September, precipitation is abundant. Drier areas are in the northwest, while the wetter areas are in the south and east. The average annual rainfall varies from 35 inches (890 millimetres) in the drier areas to 68 inches in the wetter zones. Most of the precipitation occurs in the period between May and August. In the extreme south, rain bursts caused by typhoons occur between November and February.

*Plant and animal life.* Stands of fir, red pine, cedar, camphor, and rosewood are found in the north and west; oranges grow in profusion in the south; while the cassia tree, anise, and betel palm flourish in many parts of the region. In central and south Kwangsi, many denuded hillsides have been taken over by tall coarse grasses, which are used for fuel or as pasturage for young water buffalo. Prominent types of wildlife include the bison, boar, bear, gibbon (a kind of ape), hedgehog, and cockatoo.

*Settlement patterns.* Almost 90 percent of Kwangsi's population lives in rural areas. The population density is unevenly distributed. Approximately two-thirds inhabit

the eastern third of the region, while only one-third occupies the remainder of the territory to the west. The principal cities of the region are the capital city of Nan-ning, the major city and industrial centre of the southwest; Liu-chou in the north, a hub of water and rail transport, the trading centre for the region's forest products, and a burgeoning industrial area; Kuei-lin in the northeast, which lies on the traditional trade route to central China and is a leading educational and commercial centre; Wu-chou in the southeast, the gateway to trade along the Hsi River; Pei-hai on the Gulf of Tonkin, one of China's designated "open" coastal cities; and P'ing-hsiang on the China–Vietnam border, which is a major centre of regional and international trade.

Chuang, Yao, Miao, and Tung peoples

**The people.** The population includes Han (Chinese), Chuang, Yao, Miao, and Tung. The Chuang are found largely in the western two-thirds of the region, while the Han are concentrated in the eastern third. Two distinct Chinese linguistic influences can be noted—Southwest Mandarin is spoken in the Kuei-lin district in the northeast as well as in the north, while Cantonese is spoken throughout the remainder of the region. The Yao, Miao, and Tung settlements are widely scattered.

The Chuang, a Tai people, have inhabited Kwangsi since classical antiquity. Living on the plains and in the river valleys of the hilly west, they cultivate paddy rice and practice an economy that easily merges with that of the Chinese. They are often referred to as "water dwellers" because their settlements are close to water and their dwellings are constructed on piles or stilts. For two millennia the Chuang have coexisted with the Han. The Chuang have absorbed Chinese culture, speaking both their own dialects and Cantonese. A romanized Chuang alphabet has been created and is one of the four writing systems to be printed on Chinese bank notes.

The origins of the Tung are not clear, but they are generally considered to be a branch of the Chuang, whom they resemble closely. They live in the high mountains close to the Kweichow border to the north. The Miao and the Yao, however, have long resisted the absorption of Chinese culture. They belong to a separate linguistic branch of the vast Sino–Tibetan language family. Neither the Miao nor the Yao dialects were written until alphabets based on adaptations of the Latin script were introduced in the late 1960s.

Upland dwellers who suffer from a scarcity of arable land, the Miao and the Yao practice subsistence agriculture. Characteristically, the Miao–Yao settlements are removed from transportation routes and are walled for defense. Besides farming and lumbering, which form the basis of their economy, the Yao make charcoal and bamboo basketry.

**The economy.** Since 1949 the region has made considerable progress in its economic development. Dams, canals, and reservoirs have been built to help irrigate dry lands; hydroelectric stations have been constructed and mineral resources exploited to stimulate modern industry; and rural industries have been developed in an effort to diversify village economy. Kwangsi has become self-sufficient in rice and, in fact, exports surplus rice to Kwangtung.

*Resources.* The region has sufficient coal and iron deposits to support moderate industrial development. Coal is mined north of Kuei-lin and south of Liu-chou. Iron is mined in the area near the Kwangtung–Hunan border as well as in southeastern Kwangsi. Other exploited mineral resources include tin (of which Kwangsi is a major producer), tungsten, manganese, and antimony. Moderate amounts of bismuth, zinc, and lead are also produced.

*Agriculture.* Only small areas of the region are under cultivation. Agriculture is concentrated in the river valleys and on the limestone plains. The hillsides are terraced wherever feasible. Since the 1950s the government has been seeking to bring new land under cultivation and to increase the yield of areas already cultivated by the use of irrigation and tractors. Major food crops include rice, corn (maize), wheat, and sweet potatoes. The leading commercial crop is sugarcane; other important commercial crops include peanuts (groundnuts), sesame, ramie (China grass), tobacco, tea, cotton, and indigo. Kwangsi is also a rich producer of a wide variety of fruits. The raising of livestock in Kwangsi is ancillary to farming. Water buffalo are used as draft animals in the paddy fields. Pigs, chickens, and ducks are raised on farms, and goats are raised in the hills. In many areas silkworms are also raised.

*Fishing.* Fishing is extensive. Both inshore and deep-sea fishing are carried on in the Gulf of Tonkin, which contains some of the world's richest fishing grounds. Catches include croaker (a fish that makes a croaking noise), herring, squid, prawns, eels, perch, mackerel, sharks, and sturgeon. The catching of fish fry in the region's many streams is characteristic of the freshwater fishing industry. Fish culture and the production of silkworms are complementary; the waste cocoons of silkworms are fed to the fish, and mud from fishponds is used as fertilizer for the mulberry bushes on which the silkworms feed.

Use of forest products in traditional Chinese medicines

*Forestry.* Kwangsi is an important producer of timber and forest products. In the north, large quantities of pine, fir, cedar, and giant bamboo are exploited. Red and black sandalwood are also produced in the west. More important, however, are sandarac (a resin used in making varnish and incense), star anise (Chinese anise), cassia bark (Chinese cinnamon), nutgall (a swelling on oak trees that produces tannin), and camphor. Tung oil, tea oil, and fennel oil are also produced. Some of these and other products are vital to traditional Chinese medicine.

*Industry.* Light industries produce textiles, paper, flour, silk, leather, matches, chemicals, and pharmaceuticals as well as sandarac gum, sugar, dyestuffs, and oils and fats. Pine resin is a particularly notable export of Wu-chou. Heavy industries include the ironworks and steelworks at Liu-chou and Lu-chai, machinery production at Nan-ning and Wu-chou, and the cement works at Liu-chou. The numerous handicraft products made in the region include cotton and ramie cloths, bamboo paper and rice paper, and bamboo combs. About one-tenth of the region's light industrial enterprises were granted autonomy in management in the late 1970s, and this restructured group accounted for nearly half of production and more than four-fifths of all industrial projects by the early 1980s. Tourism, especially oriented toward the city of Kuei-lin, has increased sharply and has become a significant source of income for the region.

*Transportation.* The elaborate system of waterways provides transportation throughout almost all of the region. A large proportion of the traffic is by junk, although portions of many rivers are navigable by motor launches and occasionally even by small steamers. The Hunan–Kwangsi railroad runs diagonally across the region from the northeast to the southwest. It forms a vital continental artery that connects with the Canton–Han-k'ou railroad and, south of P'ing-hsiang, with the Vietnamese railroad. The Kwangsi–Kweichow railroad links Liu-chou with Kuei-yang, Kweichow Province, and, along with the Liu-chou–Chih-chiang line, opened in 1983, is an impetus to the development of northern Kwangsi. The highway system has been substantially extended and improved since 1949. The highway network forms a central rectangle—with Nan-tan (in the northwest), Liu-chou, Nan-ning, and Pai-se (in the west) at its four corners—from which other roads radiate. Running almost due north and south, a trunk road connects Tu-yün in Kweichow Province, Nan-tan, and Nan-ning with the coast of the Gulf of Tonkin.

**Administration and social conditions.** *Government.* The region's administration is organized in a series of hierarchical levels. The top is the autonomous regional level, directly under the central government in Peking. At the second level there are five prefecture-level municipalities (*shih*) and eight prefectures (*ti-ch'ü*). Below these are county-level municipalities (*shih*), counties (*hsien*), and autonomous counties (*tzu-chih-hsien*). The lowest administrative units are the townships.

National minorities' education

*Education.* A special educational feature in Kwangsi is the program for the education of national minorities. Minority languages are used for instruction in primary and middle schools, written scripts, such as that for Chuang, are developed for spoken minority languages wherever needed, minority teachers are trained, and government subsidies are provided for minority students. Instruction in the Chuang language is offered where the size of the

Chuang population warrants it. The Institute for Minorities in Nan-ning trains both intellectuals and technical specialists of minority descent to work among the minority peoples below the county level. Institutions of higher education include the Kwangsi Normal College at Kuei-lin, as well as the Kwangsi Agricultural Institute and the Kwangsi Medical College, both at Nan-ning. The Kwangsi Provincial Museum and the Provincial Library of Kwangsi are located in Kuei-lin.

*Health and welfare.* Since the 1950s Kwangsi has made significant progress in public health and medicine. Such widespread diseases as malaria, smallpox, measles, and schistosomiasis (a parasitic infestation of the bladder or intestines) have been brought under control. The addition of iodine to water has ended the once-frequent occurrences of goitre, and the liver-fluke disease has been overcome by filling in old canals that were sources of infection and digging new ones. There is also a mass program to combat leprosy. Traditional Chinese medicine has been promoted to compensate for the shortage of Western medicine.

A basic social welfare system is available. Welfare funds guarantee care for the sick, disabled, and aged and provide relief in times of drought or flood. For industrial workers, there are accident prevention and insurance programs that provide for hospital treatment, sick leave, disability compensation, maternity leave, old-age benefits, and death benefits. Supplementary benefits are offered to those who participate in government programs such as birth control. The government has improved housing, expanded recreational facilities, and provided public-health centres.

**Cultural life.** The primacy of Chinese culture is widely recognized. Because the minorities in Kwangsi possess neither a unified organization nor support by fraternal groups, their assimilation by the Chinese is far more advanced than in the other autonomous regions. The underlying causes of what appear to be the region's ethnic tensions are economic and geographic factors that have exerted a powerful influence on cultural trends.

## HISTORY

**Early history.** Kwangsi was known as the land of Pai-Yüeh (the Hundred Yüeh—referring to the aborigines of South China) during the Chan-kuo (Warring States period) of the Tung (Eastern) Chou dynasty (475–221 BC). A subgroup of the Tai people, known as the Chuang, inhabited the region and had an economy based on wet (irrigated) rice. Eastern Kwangsi was conquered by the Han people in 214 BC under the Ch'in dynasty, and the Ling Canal was dug to link the Hsiang and Kuei rivers to form a north–south waterway.

The invasion of the Han

An independent state known as Nan Yüeh (Southern Yüeh) was created by Gen. Chao T'o, with Chuang support, at the end of the Ch'in dynasty and existed until it was annexed in 112–111 BC by the Han dynasty (206 BC–AD 220). The Han rulers reduced the power of the Chuang people by consolidating their own control in the areas surrounding the cities of Kuei-lin, Wu-chou, and Yü-lin.

In AD 42 an uprising in Tonkin was quelled by an army under Gen. Ma Yüan, who not only sought victory on the battlefield but also showed concern for the well-being of the people. He reorganized Kwangsi's local government, improved public works, dug canals, and reclaimed land to increase production. Temples erected to his memory can still be seen in many places.

From the end of the Han to the beginning of the T'ang dynasty (618–907), the influx of Yao tribes from Kiangsi and Hunan added to racial tensions in Kwangsi. Unlike the Chuang, the Yao resisted Chinese culture. The hill country of Kuei-p'ing, Chin-hsiu, and Hsiu-jen in central eastern Kwangsi (the Ta-yao-shan region) where they settled became a centre of chronic unrest. In subsequent dynasties there were further migrations of the Yao from Hunan and Kweichow provinces.

Under the T'ang dynasty, Kwangsi became a part of the Ling-nan Tao (large province). The noted scholar Liu Tsung-yüan was prefectural administrator at Liu-chou. Irked by Chinese expansion, however, the Chuang people moved to support the Tai kingdom of Nanchao in Yunnan. Kwangsi was then divided into an area of Chuang ascendancy west of a line from Kuei-lin to Nan-ning and an area of Chinese ascendancy east of the line. After the fall of the T'ang, an independent Chinese state of Nan (Southern) Han was created, but it was liquidated by the Sung dynasty in 971.

The Sung governed Kwangsi from 971 to 1279 by the alternate use of force and appeasement—a policy that neither satisfied the aspirations of the Chuang nor ended the savage warfare waged by the Yao against the Chinese. In 1052 a Chuang leader, Nung Chih-kao, led a revolt and set up an independent kingdom in the southwest. The revolt was crushed a year later, but the region continued to seethe with discontent. The Yüan dynasty imposed direct rule and made Kwangsi a province, but relations between the government and the people did not improve. To further complicate race relations, another aboriginal people—the Miao—migrated from Kweichow, and more Chuang also came from Kiangsi and Hunan.

Independent Chuang kingdom

Confronted with a complex situation, the Ming dynasty (1368–1644) actively promoted military colonization in an effort to undermine the tribal way of life. It governed the minority peoples through the hereditary *t'u-ssu* (tribal leaders serving as the agents of Chinese government). This led to some of the bloodiest battles in Kwangsi history—notably, the war with the Yao tribesmen at Giant Rattan Gorge, near Kuei-p'ing, in 1465.

The Ch'ing (Manchu) dynasty (1644–1911/12) placed the minorities under direct Imperial rule in 1726. This, however, did not bring peace. Following a Yao uprising in 1831, the great Taiping Rebellion broke out in 1850—again near Kuei-p'ing and under minority leadership—lasting for more than a decade.

Meantime, the execution of a French missionary in western Kwangsi led to an Anglo-French War against China that was concluded by the humiliating treaties of Tientsin in 1858. Then, following the Sino-French War of 1883 to 1885, French supremacy in Vietnam exposed Kwangsi to foreign encroachment. Lung-chou was opened to foreign trade in 1889, Wu-chou in 1897, and Nan-ning in 1907; while in 1898 France obtained a sphere of influence that included Kwangsi.

**The revolution.** Together with neighbouring Kwangtung, Kwangsi in the early years of the 20th century became the base of the nationalist revolution led by Sun Yat-sen. Between 1906 and 1916 the provincial leaders of Kwangsi supported the establishment of a republic, and during the following decade played an active role in the reorganization of the Chinese Nationalist Party (Kuomintang). Following the rise of Chiang Kai-shek to power in 1927, the Kwangsi leaders (notably Li Tsung-jen and Li Chi-shen) formed the Kwangsi Clique in opposition to Chiang. The group did much to modernize Kwangsi but maintained a defiant posture against the central government. Although Chiang crushed their revolt in 1929, he was unable to end the semi-independent status of the region. The Chuang, on their part, formed a string of revolutionary soviets (elected Communist organizational units) between 1927 and 1931 that gave rise to new Communist leaders.

During World War II Kwangsi was a major target of Japanese attack. The Japanese invaded southern Kwangsi in 1939 and occupied Nan-ning and Lung-chou. In this period Kuei-lin became the principal base for the Chinese and Allied air forces, as well as the home of the patriotic press, the *National Salvation Daily News.* In 1944 the Japanese made a determined drive into Kwangsi; although they briefly took Kuei-lin, Liu-chou, and Wu-chou, they were unable to maintain their position. Chinese forces subsequently recaptured the major cities. In the civil war that followed World War II, the Chinese Communist forces took Kuei-lin in November 1949, and Kwangsi became a province of the People's Republic; the autonomous region was created in 1958 in an effort to satisfy local aspirations.

World War II

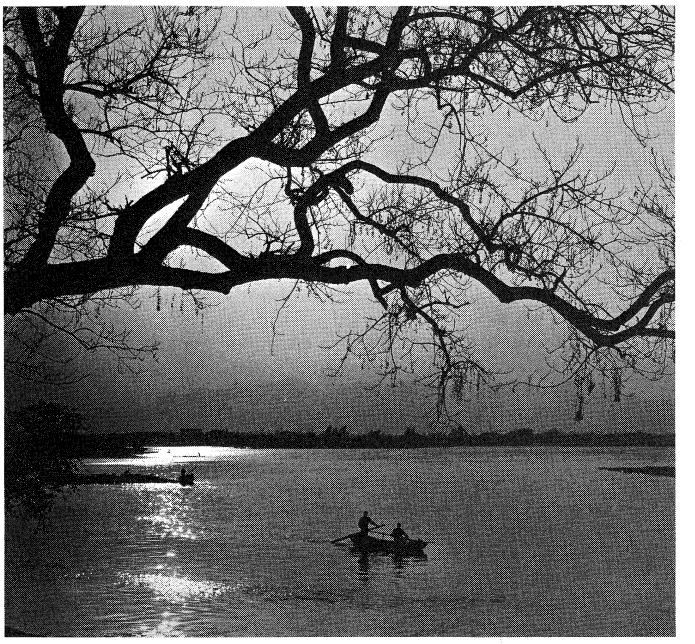
(P.-c.K./V.C.F.)

# Fukien

Fukien (Fu-chien in Wade–Giles romanization, Fujian in Pinyin) is a province on the southeastern coast of China to

the northwest of the island of Taiwan. It is bordered by the provinces of Chekiang to the north, Kiangsi to the west, and Kwangtung to the southwest; and by the East China Sea to the northeast, the Taiwan Strait to the east, and the South China Sea to the southeast. It occupies a strategic maritime position linking the two sections of the China Sea. One of the smaller Chinese provinces, Fukien has an area of 47,500 square miles (123,100 square kilometres). Its capital and largest city is Fu-chou ("Happy City").

The name of the province, Fukien, means "Happy Establishment." The province is also known as Min Sheng (Min Province), after the "seven Min tribes" that inhabited the area during the Chou dynasty (1111–255 BC). It was, however, during the Sung dynasty (AD 960–1279) that the name Fukien was adopted and the basic geographical boundaries of the province were established. The region is one of the most picturesque in Asia, with wooded hills and winding streams, orchards, tea gardens, and terraced rice fields on the gentler slopes.



Small boats on the Min River near Fu-chou in Fukien Province.
© Leon Lee

## PHYSICAL AND HUMAN GEOGRAPHY

**The land.** *Relief.* About 95 percent of Fukien is mountainous. The province is crossed by several ranges of moderate elevation that run roughly parallel to the coast. They constitute a part of a system of ancient blocks of mountains trending from southwest to northeast. The Fukien–Chekiang section forms a part of a raised massif that has been subjected to folding and refolding. A sharp natural boundary exists to the west and northwest between this uplifted block, on the one hand, and the low-lying Kiangsi Basin and the southwest part of Chekiang Province, on the other. Along this boundary run the Wu-i Mountains, which, in the extreme north, include the Hsien-hsia Mountains on the Chekiang–Fukien border.

The Wu-i Mountains, which form a formidable natural barrier between Fukien and the interior of China, reach a height of about 6,000 feet (1,800 metres) in western Fukien and in adjacent parts of southwest Chekiang. The range forms the watershed between the Min River system to the southeast and the Kan River system—a tributary to the Yangtze—to the northwest. The few passes across the Wu-i Mountains are high and difficult.

The mountain ranges tend to be more compressed in the interior and to broaden out toward the coast. Faults occur both along the axes of the mountains and across them, thus causing an extreme fragmentation of the land surface, so that local relief forms a complicated pattern.

Fukien has a submerged rocky coast that abounds in islands and islets, capes and peninsulas, and bays and havens. The shoreline is extremely irregular, with a total length estimated to be 1,680 miles (2,700 kilometres). The chief offshore islands are Quemoy (Chin-men, under Chinese Nationalist control), Hsia-men, and Tung-shan in the south; and Hai-t'an and Matsu (Ma-tsu, also under Nationalist control) in the north.

*Drainage.* Rivers are of great importance in Fukien, having for centuries provided the only means of transport. They flow into estuaries that form natural harbours, and their abundant water supplies are used for domestic consumption as well as for the irrigation of the myriad rice fields in the alluvial plains along their courses.

The general slope of the land descends from the northwest to the southeast. The main rivers cut across the intermediate ranges in deep gorges, while their tributaries drain broader intermontane valleys that follow the grain of the relief. The result is an almost perfect example of the trellis pattern of stream drainage, particularly well illustrated in the Min River system.

Min River system

The drainage area of the Min River of Fukien (to be distinguished from the Min River of Szechwan Province) occupies about half of the province. It is formed by the confluence upstream of three rivers, the largest of which is the Chien, which flows from its source near the Fukien–Chekiang border. The Chien has its own subsystem of tributary streams that drain the famous Wu-i tea district. The second source stream of the Min, the Fu-t'un, is also called the Shao-wu for the chief city of the region; it flows down the eastern slopes of the Wu-i Mountains. The third source, the Sha, flows from the south and southwest, arising on the eastern slopes of another section of the Wu-i range. The three streams, converging from the north, south, and west, meet at Nan-p'ing, their waters uniting to form the Min, which flows southeast past the city of Fu-chou to the sea.

To the south of the Min is the Chiu-lung River, which has its outlet to the sea at Amoy (Hsia-men). To the southwest of the Min is the Han River, which crosses the southwestern border of Fukien Province to empty into the sea at Swatow, the main port of eastern Kwangtung.

*Soils.* After centuries of rice cultivation, soils in the valley plains have been greatly modified. Well-developed gray-brown forest soils are widely distributed in the forest areas of the interior mountains, whereas mature red soils are common in the low hills and on high terraces. White saline soils and salt swamps are found in the coastal flatlands. Their parent rocks are marine saline deposits, penetrated by seawater. Attempts at desalination appear to have been successful, and some soils that were formerly saline are now used for rice cultivation.

*Climate.* Fukien lies just north of the Tropic of Cancer. The climate along the coastal area of the province is semitropical—very hot in summer but cool in winter. The mean temperatures in Fu-chou are about 84° F (29° C) in July and about 52° F (11° C) in January. There are three seasons in the year. November through February is the cool season; March through May, the warm season; and June through October, the hot season. The growing period lasts throughout the year. The northwestern mountains have a temperate climate but can become very cold in winter.

Summer is dominated by a monsoonal (rain-bearing) tropical airflow from the sea. Rainfall increases from the coast to the western mountains and averages between 50 and 80 inches (1,270 and 2,030 millimetres) a year. Some precipitation occurs in winter, occasionally in the form of snow in the northwest. The coast is subject to typhoons during late summer and early autumn.

*Plant and animal life.* The province has extremely varied vegetation, ranging from tropical species to forest and plant types associated with a cold temperate climate. Commercial forests are located upstream in the mountainous and rainier interior, away from rural settlements. The province has subtropical, laurel-leafed forests, as well as many kinds of conifers. In western Fukien the lower elevations support tropical mountain forests. The lianas are purely tropical. Tree ferns grow in the ravines. Higher up, where altitude modifies the climate, deciduous trees, conifers, and rhododendrons occur. Animal life in Fukien

is of the subtropical forest variety and is characterized by great diversity, with many kinds of birds, amphibians, and reptiles.

*Settlement patterns.* About one-fifth of Fukien's population lives in cities and towns, the rest in rural areas. Population densities are lowest in the mountain uplands, increase in the river valleys, and are highest on the coastal plains and estuaries near Amoy, Ch'üan-chou, and Fu-chou.

**The people.** Han (Chinese) make up nearly all of the population. The largest ethnic minority group consists of She tribesmen (She Min). Those who live in Fukien are located in the hilly hinterland of the northern coast. Most of them are distributed in the four counties of Fu-an, Hsia-p'u, Fu-ting, and Ning-te; all are engaged in farming.

Minority groups

Other minority groups include the Miao, Hui (Chinese Muslims), and Manchu. The Miao are distributed in the mountainous interior of northern Fukien; the Hui live in the cities of Fu-chou, Amoy, and Ch'üan-chou; and the Manchu live principally in Fu-chou, being descendants of Manchu soldiers who garrisoned Chinese cities during the Ch'ing (Manchu) dynasty. The She people, culturally affiliated with the Miao and the Yao, are not officially recognized as an ethnic group. They are distributed in the northern mountains, from the coast to the interior, and are even found beyond the Fukien border in Kiangsi and southern Chekiang. Nor are the "boat people" (Tanka or Tang-chia), who live on boats in the streams and estuaries, recognized as a separate group.

The four principal dialects

There are four principal local dialects in Fukien. Hokchiu (the Fu-chou dialect) is spoken principally in Fu-chou and in the Min-hou area corresponding roughly to the area of the former Fu-chou Fu (prefecture). Hokkien, the Amoy dialect, is spoken in southern Fukien (thus, it is also known as the Min-nan, or south Fukien, dialect). The Hokchia, or Hakka, dialect of Fukien is spoken in the upper Han Valley of southwestern Fukien. Lastly, the Henghua dialect is spoken in the Henghua district between Fu-chou and Amoy. There are also literally hundreds of subdialects, making the province one of the most linguistically fragmented in China.

**The economy.** *Agriculture.* Since the 1950s Fukien has largely been a net importer of food grains despite significant growth in output. Its major crops are sugarcane, peanuts (groundnuts), citrus fruit, rice, and tea. Fukien's sugarcane yields are among the highest in the nation. Much of the province's productivity comes from its use of chemical fertilizer. A growing proportion of agricultural output has also come from noncrop sources, particularly from fisheries, animal husbandry, and forest products. The most important woods are fir, pine, and rosewood, mostly floated in the form of big rafts down to Fu-chou, a great timber emporium. Plans emphasize the more intense exploitation of Fukien's hilly uplands as the key to its more rapid agricultural diversification.

Two crops of rice are harvested each year, the first in June, the second in September. The export of tea from Fu-chou to the European market has become insignificant, but Fukien remains a great tea-growing province with a large domestic market. A special feature is its production of flower-scented teas, for the manufacture of which there are factories in Fu-chou. There are also factories for the manufacture of paper from bamboo pulp.

Fukien also has considerable mineral wealth, including coal, iron, copper, gold, graphite, and kaolin (china clay) for making porcelain. Mines are widely scattered over the province.

*Industry.* Before 1949 Fukien had little modern industry. The modern Fu-chou Shipyard was built in 1866, but it was largely destroyed in the 1880s. There was some Russian and Japanese investment in tea and textiles in the 1870s and a spurt of overseas Chinese investment in food-processing industries in the coastal areas in the early 1900s, but overall, the modern industrial base was negligible at the establishment of the People's Republic in 1949.

During the 1950s investment in the province was hampered by Peking's decision to emphasize inland rather than coastal provinces and by conflict in the Taiwan Strait, which made the national government hesitant to invest in a potential war zone. Fukien's share of investment was smaller than that of any other province. Gradually, however, for strategic and developmental reasons, the economy began to grow, and regional centres, particularly in the Min Valley, developed. Nan-p'ing became a key forest-products centre, acquiring one of the country's advanced pulp and paper plants. San-ming became the site of a medium-sized iron and steel plant drawing on local coal and iron reserves. The development of a major cement plant at Shun-ch'ang laid the foundation for the local building materials industry.

Economic growth

Provincial economic growth increased markedly with the shift in government policy toward favouring the development of coastal trading cities. Fukien and Kwangtung were given special powers in 1979 to attract foreign investment, particularly in export industries, and to establish special economic zones for that purpose. One such zone was set up in northwest Amoy in the 1980s to develop industrial sites and support infrastructure for the zone. The effect was to double the harbour capacity of Amoy.

A second reform was the creation of a south Fukien special economic zone, similar to the Shanghai and Pearl River Delta zones, designed to orient regional development in southernmost Fukien toward the production of light industrial goods for export. A similar pattern of development is also affecting Fu-chou, which was designated one of China's "open" cities in 1984 and which has been working to establish an economic and technical development zone near the port city of Ma-wei.

*Transportation.* From the Korean War and the partial blockade of the Fukien coast by the United States and by Chinese Nationalist forces based on Taiwan, overseas trade was virtually halted. Fukien's trade patterns consequently turned inland, especially after the completion in 1955 of the Amoy–Ying-t'an railway, which crosses the Wu-i range to link the province with the Chinese national rail network. Fukien's traditional isolation has also been breached in recent years by the construction of modern highways, linking it to neighbouring Kiangsi and Chekiang provinces. Air services centre on the chief airport at Fu-chou.

Fukien's rivers are still in use for transportation. The headwaters of the Chin River, a tributary of the Fu-t'un River, are navigable for small boats right up to the Wu-i Mountains, despite the river's rocky channel and many rapids; boats bring downstream the tea grown on the slopes of the mountains. Below Chien-ning, larger boats of special construction are employed for the tea trade.

**Administration and social conditions.** *Government.* The original leadership of the province was drawn from the ranks of the 1920s local guerrilla movement and from the soldiers of the 3rd Field Army, which took control in 1949. They were displaced in the turmoil of the Cultural Revolution, which badly affected the province. After the late 1970s a new leadership emerged with a more technocratic and development-oriented character. Within provincial jurisdiction are five prefectures (*ti-ch'ü*) and four prefecture-level municipalities (*shih*). Below that level are counties (*hsien*) and county-level municipalities (*shih*).

*Education.* One of the most notable institutions of higher learning in Fukien is Amoy University. Fu-chou, famous since the Sung dynasty as a cultural centre, is the site of Fu-chou University, Fukien Medical College, Fukien Agricultural Institute, Fukien Normal College, and the Fukien Institute of Epidemiology of the Chinese Academy of Medical Sciences. Illiteracy has generally been eliminated among those persons born since 1950.

*Health and welfare.* Public health has improved considerably since the establishment of the People's Republic, and malnutrition has not been reported for decades.

**Cultural life.** Traditional Chinese culture reached a high level in Fukien during the Sung period (960–1279). Certain unique traditional customs evolved that gave women a stronger social position than that of the women in North China. The province's long literary tradition centres about local history recorded during the last thousand years.

At least two distinct provincial subcultures persist to this day, reflecting linguistic and historical differences among Fukien's regions. The Min-pei, or northern section of

Fukien centred on Fu-chou, was an early centre of Buddhism and, because of close contact with Japanese culture through the Ryukyu Islands, still shows some of those influences in culture and cuisine. As the centre of administration, the Min-pei has a more conservative tradition and, with its seafaring history, has historically supplied many of China's greatest naval officers.

In contrast, the Min-nan, or southern Fukien, centred on the Amoy–Chang-chou–Ch'uan-chou triangle, has the reputation of being more commercial, adventurous, and hardworking. With strong linguistic differentiation, it is home to a rich operatic and balladic tradition of its own. Much of the modern history of the region has been shaped by the close continuing contact between Min-nan peoples and their overseas relatives, who set down roots in Southeast Asia from the 16th century onward.

Fukien cuisine is considered to be one of China's five main regional cooking styles, though it is not well known outside China. Fukien is also known for its strong educational tradition. During the Ming and Ch'ing dynasties many of China's great statesmen and scholars came from the province.

### HISTORY

Early history

The area now called Fukien was first referred to in the *Chou li,* a classic that may date to the 12th century BC, although modern scholars believe it to have been written at a much later date. In this classic the seven Min tribes are mentioned together with "eight barbarian peoples" in the south.

During the latter part of the Ch'un-ch'iu (Spring and Autumn period; 770–476 BC) one of the feudal states within the China area was the kingdom of Yüeh, located south of Hang-chou Bay; it included what is now Fukien Province. The lord of Yüeh was nominally a vassal (viscount) of the Chinese king. The Yüeh and their culture are considered by some to have constituted one of the principal elements that merged to form the contemporary Chinese ethnic and cultural complex.

During the last quarter of the 5th century BC, Yüeh became a powerful kingdom after its conquest of the state of Wu (473 BC) to its north. During the era known as the Chan-kuo (Warring States) period, Yüeh was, in turn, conquered by the kingdom of Ch'u (*c.* 334 BC) to the northwest. Wu-chu, one of the sons of the vanquished Yüeh king, fled by sea and landed near Fu-chou to establish himself as the king of Min-yüeh. When the first emperor of the Ch'in dynasty conquered the kingdom of Ch'u in 223 BC, the Chinese domain was finally unified within the bounds of a monolithic state. Li Ssu, the famous prime minister of Ch'in, deposed the King of Min-yüeh, establishing instead a paramilitary province there called Min-chung Chün. The collapse of the Ch'in dynasty (206 BC) was followed by the war between the famous general Hsiang Yü and the crafty Kao-tsu, the founder of the Han dynasty. Wu-chu, the deposed king of Min-yüeh, sided with Kao-tsu, who defeated his rival and became emperor of China; he reestablished Wu-chu as the king of Min-yüeh, which consisted roughly of the present area of Fukien. During the reign of the emperor Wu-ti (141/140–87/86 BC) a rebellion by the Min-yüeh tribes was put down, and the tribes were resettled in the inland region far to the north between the Huai and Yangtze rivers.

During the Six Dynasties period (AD 220–589) the region remained in the Chinese domain, but true Sinicization did not come about until the T'ang dynasty (618–907), when intermarriage between the T'ang settlers from the north and the local people became common.

After the fall of the T'ang, the territory of Fukien reemerged as the kingdom of Min, with its capital in Fu-chou. In the middle of the 10th century it was subdivided into the state of Yin, controlling the Min-pei, and the state of Min, controlling southern Fukien from Chang-chou. The province grew rapidly in importance as the economic hinterland of the Nan (Southern) Sung capital, Lin-an (modern Hang-chou). The province became a key supplier of rice to the region following the introduction of a fast-ripening variety called Champa rice from Southeast Asia. It also became the major producer of sugar, fruit, and tea. Because of the importance of trade to the Nan Sung, the province also was important as a shipbuilding and commercial centre for both overseas and coastal trade. The port of Ch'üan-chou, known to Marco Polo as Zaitun, was one of the world's great ports in this period, with more than 100,000 Arab traders living in the area.

Marco Polo's port of Zaitun

The province's decline began with the Ming dynasty ban on maritime commerce in 1433 and was reinforced by the Ch'ing dynasty's policy of isolation, which particularly affected the province in the late 17th century, when Ming dynasty loyalists occupied Taiwan and the islands off Fukien. There was some revival of the economy in the mid-19th century with the opening of Fu-chou and Amoy as treaty port cities, but the modern shipbuilding industry established at Ma-wei by the Ch'ing was destroyed by a French fleet during the Sino-French War of 1883–85.

In the aftermath of the revolution of 1911–12, Fukien was a pawn in local warlord struggles and was divided into political and military fiefdoms. In the early 1930s part of western Fukien was incorporated into the Communist-controlled territory of the Kiangsi Soviet. In 1933 a revolt of government troops stationed in the province against the Nanking government led to assertion of Nanking government control over the province and to the expulsion of Communist forces. After 1938 the Japanese occupied the coastal centres of the province, while the provincial government retreated inland to Yung-an in central Fukien, where it administered the interior of the province for the remainder of the war. In 1949 the Communist-led 3rd Field Army took control of the province. (F.Hu./V.C.F.)

## Hai-nan

The province of Hai-nan (Hai-nan in Wade–Giles romanization, Hainan in Pinyin), the name of which means "South of the Sea," is coextensive with Hai-nan Island. Hai-nan is located in the South China Sea, separated from Kwangtung's Lei-chou Peninsula to the north by a shallow and narrow strait. It is the southernmost province of China and, with an area of about 13,200 square miles (34,300 square kilometres), is also the smallest. For centuries Hai-nan was part of Kwangtung Province, but in 1988 this resource-rich yet economically underdeveloped tropical island became a separate province. The capital is Hai-k'ou (Pinyin, Haikou).

### PHYSICAL AND HUMAN GEOGRAPHY

**The land.** Hai-nan's long coastline of more than 930 miles (1,500 kilometres) contains numerous bays and natural harbours. Alluvial plains cover a narrow coastal margin, reaching their broadest point in the northeast. The southern interior of the island is mountainous, with Mount Wu-chih rising to over 6,000 feet (1,830 metres) above sea level. Numerous rivers and streams cascade out of the mountains; the longest, the Nan-tu River, flows northeastward.

Climate

Hai-nan's climate is tropical. Temperatures average about 64° F (18° C) in January and 84° F (29° C) in June. Rainfall is heavy, with an average annual precipitation of about 70 inches (1,800 millimetres) in the south and 60 inches in the north. The northeastern lowlands can sustain triple-cropping of rice.

The island is covered with mature red soils. The natural vegetation, which has been much reduced, includes many palms, bamboos, rattans, and tropical hardwoods. The mountain belt, especially in the east, is covered with dense tropical rain forest up to an elevation of about 2,600 feet. Animal life is rich and varied and includes deer, gibbons, and blind snakes; Hai-nan's streams and offshore waters abound in fish.

**The people.** Two-thirds of Hai-nan's predominantly rural population is concentrated in the northeastern lowlands. Most of the people are Han (Chinese), but about one-sixth are ethnic minorities. The Li, concentrated in the south, constitute the largest minority group, followed by the Miao. The largest cities are Hai-k'ou in the north and the port city of Ya-hsien (locally called San-ya) in the south. The lingua franca of Hai-nan, Hainanese, is a variant of the southern Fukien dialect.

**The economy.** Hai-nan's economy is predominantly agricultural, and more than two-thirds of the island's exports are agricultural products. Hai-nan's elevation to province-level status, however, was accompanied by its designation as China's largest "special economic zone," the intent being to hasten the development of the island's plentiful resources. The central government has encouraged foreign investment in Hai-nan and has allowed the island to rely to a large extent on market forces.

Hai-nan has commercially exploitable reserves of more than 30 minerals. Iron, first mined by the Japanese during their occupation of the island in World War II, is the most important. Also important are titanium, manganese, tungsten, bauxite, molybdenum, cobalt, copper, gold, and silver. There are large deposits of lignite and oil shale on the island, and significant offshore finds of oil and natural gas have been discovered. Virgin forests in the interior mountains contain more than 20 commercially valuable species, including teak and sandalwood.

*Agriculture and fishing.* Paddy rice is cultivated extensively in the northeastern lowlands and in the southern mountain valleys. Leading crops other than rice include coconuts, palm oil, sisal, tropical fruits (including pineapples, of which Hai-nan is China's leading producer), black pepper, coffee, tea, cashews, and sugarcane. In the early 20th century Chinese emigrants returning from Malaysia introduced rubber trees to the island; after 1950, state farms were developed, and Hai-nan now produces most of China's rubber.

Rubber production

Pan Jiamin—New China Pictures



Harvesting coconuts on Hai-nan Island, Hai-nan Province.

Marine products contribute a significant share to the provincial economy. Shrimps, scallops, and pearls are raised in shallow bays and basins for local use and export. Grouper, Spanish mackerel, and tuna constitute the bulk of the catch from the rich offshore fishing grounds.

*Industry.* Hai-nan's industrial development largely has been limited to the processing of its mineral and agricultural products, particularly rubber and iron ore. Since the 1950s machinery, farm equipment, and textiles have been manufactured in the Hai-k'ou area for local consumption. A major constraint on industrial expansion has been an inadequate supply of electricity. Much of the island's generating capacity is hydroelectric, and it is subject to seasonal fluctuations in stream and river flows.

*Transportation.* Before 1950 there were practically no transportation links with the interior of the island. The Japanese built a railroad from the iron mines in the southwestern mountains to the coast, which subsequently was upgraded and extended around the southern coast, but there is no link with Hai-k'ou in the north. The first roads were built in the early 20th century, but no major road construction was undertaken in the mountains until the 1950s. Parallel north–south roads along the east and west coasts and through the interior of the island constitute most of Hai-nan's road network. The freight-handling facilities of the island's ports have been improved, and Hai-k'ou's airport has been upgraded to accept international air traffic.

**Government.** Even while Hai-nan was a part of Kwangtung it had a considerable amount of local autonomy; the southern half of the island was an autonomous prefecture (*tzu-chih-chou*). Hai-nan's elevation to provincial level increased its accountability to the central government, but by designating the new province a special economic zone the central government expressed its intent to allow Hai-nan maximum flexibility in devising programs to facilitate foreign investment and economic growth. Administratively, the province has been divided into five economic districts.

**Cultural life.** Hai-nan has always been on the fringe of the Chinese cultural sphere. Traditionally, the island was a place of exile for criminals and disgraced officials. As a frontier region celebrated by such exiled poets as Su Tung-p'o, Hai-nan acquired an air of mystery and romance. The influx of large numbers of mainlanders after 1950—particularly in the 1970s, when young Chinese from southern Kwangtung were assigned to state farms to help develop Hai-nan, and in the 1980s, when thousands more came to take advantage of the economic opportunities offered—has perpetuated the frontier atmosphere on the island. The level of primary and secondary education has improved since 1949, but facilities for higher education remain inadequate.

Education

HISTORY

Hai-nan was formally incorporated into the Chinese empire in 110 BC, when the Han government established a garrison in the north. Chinese sovereignty remained nominal, however, until the T'ang dynasty (AD 618–907). During the Yüan (Mongol) dynasty (1206–1368) it became an independent province, at which time it acquired the name Hai-nan. In 1370, however, it became a part of Kwangtung. Hai-nan's first major period of settlement occurred in the 16th and 17th centuries, when large migrations from Fukien and Kwangtung pushed the island's indigenous peoples into central and southern Hai-nan.

In 1906 the Chinese Republican leader Sun Yat-sen proposed that Hai-nan become a separate province, and for a short time (1912–21) it was nominally independent under the name Ch'iung-yai Island. The Japanese occupied Hai-nan (1939–45) during World War II and began developing the island's resource potential. Hai-nan reverted to the Chinese Nationalists in 1945 and was one of the last places to fall to the Communists.

After 1950 Hai-nan served as a military outpost and as a source of raw materials, but because of its strategic vulnerability the central government was reluctant to make it an investment priority. With China's shift in economic policy at the end of the 1970s, Hai-nan became a focus of attention. In 1984 the island was designated as a special zone for foreign investment; and, though it was still part of Kwangtung, it was upgraded to the status of a self-governing district, a prelude to its establishment as a province in 1988. (V.C.F.)

## Hunan

Hunan (Hu-nan in Wade–Giles romanization, Hunan in Pinyin) is a landlocked province of China, covering an area of 81,300 square miles (210,500 square kilometres). A major rice-producing area, Hunan is situated to the south of the Yangtze River Basin. It is bounded by the provinces of Hupeh to the north, Kiangsi to the east, and Kwangtung to the southeast, by the Chuang Autonomous Region of Kwangsi to the southwest, and by the provinces of Kweichow and Szechwan to the west. The name Hunan is formed from the Chinese words *hu* ("lake") and *nan* ("south"), meaning the land to the south of the lakes

that reach from Sha-shih, Hupeh, to Chiu-chiang, Kiangsi. The capital and most important city of the province is Ch'ang-sha, situated in the northeast, on the banks of the Hsiang River.

## PHYSICAL AND HUMAN GEOGRAPHY

**The land.** *Relief.* More than one-quarter of the terrain lies at a height of more than 1,640 feet (500 metres), and much of it is well over 3,000 feet above sea level. The highlands in the west run from northeast to southwest, forming the eastward edge of the Kweichow Plateau, whose extension, the Hsüeh-feng Mountains, lies in the heart of the province. These mountains are composed mainly of slate, quartzite, and sandstone, deeply incised by river valleys. The Nan Mountains in the south run from east to west at altitudes of between 500 and 3,300 feet, forming a broad mountain border between Hunan, Kwangtung, and Kwangsi. They are largely dome-shaped and granitic, although limestone and red clay are found in lower-lying areas. In the east the mountain ranges of Chu-kuang and Wu-kung form the border with Kiangsi. The Chu-kuang Mountains, in the extreme southeast of the province, rise to a height of 6,600 feet.

The uplands of the west, south, and east fall steadily in altitude toward the plain of Tung-t'ing Lake in the north, which is contiguous with the Hupeh Plain and forms part of the floodplain of the Yangtze River. The part of the plain within the borders of Hunan has an area of 3,800 square miles; it has been formed by silt carried down from the mountains by the Yangtze and its right-bank tributaries. Tung-t'ing Lake is a broad and shallow seasonal lake, consisting of the remnants of a former inland sea that once filled the Yangtze Basin. Its area varies considerably between summer and winter, and the lake has dried up considerably since the 1950s.

Tung-t'ing Lake

*Drainage.* Hunan's entire river system drains into Tung-t'ing Lake, with the exception of the Lin Stream, which divides into two parts, with one distributary draining directly into the Yangtze River and the other into Tung-t'ing Lake. The western highlands are drained by the Yüan River and by the Tzu and Li streams. The Yüan and Tzu are torrents in their upper courses; fast-flowing in summer, they run through deep gorges, broadening out to wider valleys in their lower courses. Hunan's largest river, the Hsiang, rises in the heart of the Nan Mountains, as do its tributaries. Many smaller rivers that rise in the mountains along the eastern border flow westward to join the Hsiang in its northward course.

Much of the low-lying land around Tung-t'ing Lake is subject to flooding when the rivers come down in spate during the summer months. The system of dikes built to contain the floodwaters is supplemented by a vast network of electric pumping stations. These pumps drain the fields when waterlogged and irrigate them in times of drought. In the dry hill lands, numerous large and medium-sized water-control projects have been built. In these projects, valleys are dammed and "mountain pools" formed, from which channels are led to the arid land. One of these schemes—the Shao-shan Irrigation System—diverts some of the upper waters of the Lien Stream, thus irrigating the dry hill land, and also controls flooding in the river's lower reaches; the irrigated area has been converted from single-crop to double-crop rice land.

The Shao-shan Irrigation System

*Soils.* The soils of the province are largely pedalferic (rich in alumina and iron) and are mainly lateritic (leached, iron-bearing) yellow earths or red clays. In the hilly regions of central and southern Hunan, the soils are for the most part lateritic clays that are strongly acidic and poor in organic material. The alluvial soils of the northern plains are less acidic and are used for growing rice.

*Climate.* The north generally experiences more extreme weather conditions than does the south. In winter, occasional waves of cold air from Mongolia sweep southward, injuring tea bushes and fruit trees in northern Hunan. The average minimum winter temperature is 43° F (6° C). Summers are long and humid, the average maximum summer temperature being 86° F (30° C). The north has an average of 260 frostless days a year, while in the south the average is 300 days. Rainfall is ample, with the maximum precipitation occurring between spring and summer. The total annual rainfall of 56 inches (1,422 millimetres) decreases from south to north. Hunan lies in the path of cyclones that pass from west to east along the Yangtze Basin in summer, bringing with them at times long periods of heavy rain, resulting in extensive flooding of low-lying lands.

*Plant life.* The natural vegetation of Hunan was originally dense deciduous and coniferous forest. Over the centuries, as the population has increased, all the lowlands and much of the highlands have been cleared to make way for cultivation. Despite this vast deforestation, however, large stands of pine, cedar, bamboo, and camphor are found in the western highlands. Other important trees and shrubs include tung (from which tung oil is obtained), tea (from which tea seed oil is obtained), and liquidambar. Bamboo groves planted along the roadsides are characteristic of Hunan and supply the province's craft industries. As elsewhere in central and southern China, groves of bamboo, camphor, and cedar are usually found around villages, contributing greatly to the charm of the countryside.

*Settlement patterns.* Villages are usually small, and it is not unusual for an entire village to belong to one extended family, from which the settlement takes its name. Most of the farms on the plain south of Tung-t'ing Lake are built on islands of Yangtze alluvium, protected by dikes from summer flooding.

Three of the largest cities, Ch'ang-sha, Hsiang-t'an, and Chu-chou, lie close together at the intersection of road, rail, and river communications along the Hsiang River. Other large cities include Heng-yang, the economic and communications centre of southern Hunan, and Ch'ang-te, the marketing centre for the Yüan River Basin.

**The people.** The overwhelming majority of the population is rural. The population is primarily concentrated on the Tung-t'ing Plain and in the main river valleys. Almost all of the people are Han (Chinese). In addition, some minority peoples, mainly of four tribes—the Miao, T'u-chia, Tung and Yao—live in the western highlands. The way of life and economy of the Miao and the T'u-chia are similar, and much intermarriage has occurred between them. They live in the southwest, where their economy is based on the cultivation of terraced fields in the foothills and narrow valleys. The Tung inhabit their own autonomous counties in the extreme southwest, with their centres at T'ung-tao and Hsin-huang. Their language, economy, and way of life are similar to those of the Han. The Yao are widely scattered over the mountainous regions of the south and west. They practice dry farming and are known for their expertise in cedar tree culture. Much of their livelihood comes from forestry.

Minority peoples

The Han of the province speak a dialect—Hunanese—that approximates Mandarin quite closely. Radio broadcasting has had the effect of slowly reducing differences in local dialects, which can be considerable. The minority languages were unwritten until missionaries devised scripts for some of them, such as the Samuel Pollard script for the Miao language. Since 1949 these scripts have been revised, extended, or replaced by a phonetic script, based on the Latin alphabet, that is akin to the Pinyin script adopted for the Mandarin language of the Han. There is growing literacy among the Miao and Tung peoples. The interweaving of Confucianism, Buddhism, and Taoism, as well as Islām and Christianity, are complicated.

**The economy.** *Agriculture.* Although mining and industry have been developed since 1949, Hunan's economy remains mostly agricultural. Hunan ranks first among China's provinces in rice production. Most of Hunan's arable land is farmed using modern techniques, including mechanized irrigation and chemical fertilization. Most farms are small, however, and mechanization has been confined to the use of simple machines and tools, such as rice transplanters, foot-operated rice-threshing machines, and a tube water raiser that is replacing the old wooden trough and paddles.

Hunan consistently ranks first nationally in rice output and exports a large surplus to other provinces. It is estimated that most of the province's cultivable land is

Rice production



Terraced paddies in Hunan Province, one of the leading producers of rice in China.
© Leon Lee

devoted to paddies (rice fields), a great many of which in the south produce two crops of rice per year and demand careful cultivation. The first crop is planted at the end of April and harvested at the end of July; the second crop is harvested in November. Autumn is the most difficult period, as decreasing rainfall and increasing evaporation necessitate continuous irrigation. With improved irrigation, a decreasing amount of rice is grown on fields where the crop relies totally on rainfall. Introduction of hybrid rice varieties has further increased production. Other food crops include sweet potatoes, corn (maize), barley, potatoes, kaoliang (a variety of grain sorghum), buckwheat, garden peas, millet, and horsebeans.

Among the industrial crops, rape—an herb grown for its seeds—is cultivated mainly in the upper valleys of the Hsüeh-feng Mountains, while ramie (a shrub that yields a fibre used in textile manufacture), cotton, and jute are produced around Tung-t'ing Lake. Red and black tea are grown on the foothills of the Hsüeh-feng Mountains and on Mount Mu-fu on the eastern border. Peanut (groundnut) cultivation is widespread, and tung trees and tea seed shrubs are grown for their oils in the western and southern highlands. A variety of fruits is grown throughout the province, including citrus, pears, peaches, and chestnuts.

During the early 20th century, heavy and wasteful cutting of Hunan's timber reserves occurred. Since then, stricter control of cutting has been enforced, and some reforestation has been carried out. Fish are taken in large quantities from lakes, rivers, and village ponds. The most common varieties are carp, silver carp, and "silver fish." The full exploitation of fishpond culture was developed in the early 1970s. Cattle, including water buffalo, are used almost exclusively for draft purposes. Hogs are concentrated mainly in the central and eastern areas, where the population is densest. The swine industry is a significant source of rural cash income.

*Industry.* Hunan's considerable mineral wealth includes ample coal reserves; iron ore, tin, and manganese deposits; rich deposits of antimony; and lead, zinc, tungsten, molybdenum, bismuth, niobium, and tantalum. More than half of the province's electric power is produced by hydroelectric power stations.

The main coal measures are located in the south. Coal was little developed before 1949, but production rose substantially as a result of the opening of large mines north of Ch'en-hsien in the extreme south. These mines serve the ironworks and steelworks at Wu-han in Hupeh. Iron ore is widely distributed, and there is a long-established local industry that produces iron cookware. The main iron mines are located in the hills east and south of Ch'ang-sha and Hsiang-t'an. Development of the iron and steel industry is centred in the triangle formed by the three large cities of Ch'ang-sha, Hsiang-t'an, and Chu-chou. Antimony production is centred on Hsin-hua, northwest of Shao-yang. Hunan is one of China's largest producers of tungsten; it is chiefly mined in the hills between the Tzu and Yüan rivers around Yüan-ling.

Plants producing iron and steel, processed foods, and electrical equipment are located in Hsiang-t'an, while Chu-chou is the centre of large-scale heavy industry and hog exports. Ch'ang-sha is Hunan's centre for light industry, which includes rice milling, food processing, aluminum smelting, and the manufacture of machine tools, bearings, and textiles. It is also famous for its handicrafts, which include *hsiang* (border) embroideries, duck-down quilts, umbrellas, and leather goods. I-yang—known as "the Bamboo Town"—is typical of many of the smaller towns specializing in one particular handicraft. Nearly everything required in domestic life—from beds to scrubbing brushes—is made from bamboo.

I-yang, the Bamboo Town

There are several famous pottery kilns in the province that date back to the T'ang dynasty (618–907). Situated at Yüeh-yang (Yoyang), Hsiang-yin, and near Ch'ang-sha, these kilns have at different epochs produced all sorts of wares, according to the market of the period. Their fortunes have fluctuated through the centuries. More recently, they have increased their output, especially in the Hsiang-yin kilns, which produce large quantities of crockery for the general market.

*Transportation.* Hunan stands at the crossroads of China's historical lines of communication—the great waterway of the Yangtze River, which flows from Szechwan Province to the sea, and the Imperial Highway, running from Canton northward to Peking.

Railway construction began in 1912, and the first line was between Wu-han and Chu-chou; it was eventually extended south to Canton and is now part of the major Peking–Canton trunk line. There is a junction at Heng-yang leading to Kwangsi, from where the line continues south to Hanoi. From Chu-chou, the Che–Kan Railway runs east to Kiangsi Province and also to Chekiang and Fukien on the east coast. Another railway runs from Hsiang-t'an westward to Kweichow Province, opening up the hitherto remote western lands, and a second north–south line has been built through western Hunan.

Water and road transport

Shipping is another important means of transportation, as about one-fourth of Hunan's goods are moved by water. Traffic on the Hsiang River is the most important. Tung-t'ing Lake has innumerable shallow waterways connecting four main rivers. Yüeh-yang in the northeast corner of the lake is the collecting centre for the timber rafts that sail the Yangtze River to Wu-han. One main trunk road runs

from north to south, following the Peking–Canton railway into Kwangsi. Three other main routes run from east to west and are of growing importance because they open up areas not served by the railways.

**Administration and social conditions.** *Government.* From 1950 to 1954 Hunan was part of the South Central China greater administrative region, which extended from Hupeh in the north to Kwangtung and Kwangsi in the south. In 1954 provincial (*sheng*) government throughout the country was made directly subordinate to the national government. Since then the province's administrative structure has gone through several changes. It is now divided into eight prefectures (*ti-ch'ü*), six prefecture-level municipalities (*shih*), and one autonomous prefecture (*tzu-chih-chou*). Below that level are counties (*hsien*) and county-level municipalities (*shih*).

*Health and welfare.* Emphasis is laid on preventive medicine. After 1949, public health teams were sent into the country to vaccinate and inoculate and to advise on and supervise public hygiene. Debilitating diseases such as malaria and schistosomiasis—a disease of the blood and tissues that is spread by larvae in the droppings of animals in the rice fields—have been attacked, and since 1949 there has been a marked fall in infant mortality and an increase in life expectancy.

**Cultural life.** Although the aim of the government is to promote linguistic uniformity, Hunanese—which was Mao Zedong's (Mao Tse-tung's) own dialect and is fairly akin to Mandarin—persists.

Education

Before the revolution, Western learning was largely acquired through Christian missionary or other Western-style schools, and most of the population remained illiterate. Since 1949 a countrywide literacy drive has been pursued with vigour and enthusiasm and a large measure of success. By the 1980s Hunan had one of the lowest illiteracy rates in the country. Ch'ang-sha has retained its historic role as the province's cultural and educational centre and is the location of technical institutes, teacher-training colleges, and institutes for minorities.

HISTORY

From 350 to 221 BC Hunan formed the southernmost extension of the state of Ch'u, which nominally was ruled by the Chou dynasty. From 221 to 206 BC Hunan was ruled by the Ch'in dynasty, which subdued contending feudal states and joined them into the first unified state of China, of which Hunan formed part of the central area. Most of Hunan at this time was covered with dense primeval forest that was sparsely inhabited by tribes who engaged in hunting, fishing, and clearing land by burning or cutting for temporary cultivation. These tribes also mined the copper and tin that were used in the north for making bronze.

After the downfall of the Ch'in dynasty, the area became quickly incorporated into the Chinese empire ruled by the Han dynasty from 206 BC to AD 220. During this

Early migrations

period persistent waves of migrant Han (Chinese) from the North occupied the land, and the indigenous Miao, T'u-chia, Tung, and Yao were pushed west and southwest into the hills, which they still occupy. By the end of the Hsi (Western) Chin dynasty in AD 316/317, the Tung-t'ing floodplain to the north and the Hsiang Valley in the east were relatively well populated. Han migration from the North continued under subsequent dynasties, with migrants fleeing first from Mongol and then from Manchu invasions. Those who went farther south, crossing the Nan Mountains in the southern part of the province to enter Kwangtung, have since considered themselves T'ang-jen, or southern Chinese, but the Hunanese have remained Han in both culture and speech.

Population pressures on the land increased markedly in the 19th century during the latter part of the Ch'ing, or Manchu, dynasty (1644–1911/12), leading to increased peasant unrest, particularly among the non-Chinese tribes. When the Taiping Rebellion broke out in Kwangsi in 1850, it spread northward into Hunan. Hunan, together with other provinces on the lower Yangtze Basin, was desolated in the subsequent fighting, although the city of Ch'ang-sha withstood a Taiping siege in the mid-1850s. It was a Hunanese, Tseng Kuo-fan, who ultimately was responsible for crushing the rebellion.

Hunan was not opened to foreign trade until 1904, following the conclusion of the Treaty of Shanghai between China and Japan. A foreign settlement was established at Ch'ang-sha, and British and Japanese firms built warehouses. Hunan became a centre of revolutionary activity: the first uprisings against Yüan Shih-k'ai's attempted regency over the Chinese empire occurred in the province in 1910, although the more widespread revolution that finally overthrew the tottering Manchu dynasty and established the Republic of China did not occur until the following year. Thereafter, Hunan remained in a state of unrest from which it had little respite until 1949, when the People's Republic of China was established. Many important Chinese Communist leaders—including Mao Zedong, who was born in Shao-shan, near the border with Kiangsi, and Liu Shaoqi (Liu Shao-ch'i), chairman of the People's Republic (1959–68)—were from Hunan. Mao was largely responsible for encouraging the peasants and miners to make the abortive Autumn Harvest Uprising of 1927. He subsequently held the Communist forces together in the Ching-kang Mountains, where they withstood repeated attacks by the forces of Chiang Kai-shek, the Chinese Nationalist leader. In 1934 Mao set out from the Hunan–Kiangsi border region, leading his forces westward in the difficult northward retreat that later came to be known as the Long March.

The years of unrest (1910–49)

During the Sino-Japanese War Hunan was the scene of bitter fighting between 1939 and 1941. After the fall of Hunan to the Japanese, the Nationalist general Hsüeh Yüeh continued to successfully defend Ch'ang-sha against the Japanese invaders, until it too fell in 1944. Between 1946 and 1949 the province was relatively peaceful. In 1949, despite damage to bridges and communications, the province experienced comparatively little destruction when the Nationalist forces retreated rapidly southward before the advancing Communist armies.

Provincial leaders from Hunan have played an important national role since 1949. Hunan's provincial party leader was purged in 1958 for opposing the economic policies of the Great Leap Forward (1958–60) and was replaced by supporters of Mao Zedong's more ambitious and radical policies. One of Mao's rising provincial supporters, Hua Guofeng (Hua Kuo-feng), was Communist Party chairman (1976–81) after Mao's death.

Hunan supported many of the policies of Mao's Cultural Revolution (1966–76), and it was slower than other provinces at implementing the economic and political reform programs instituted by the post-Mao leadership. Gradually, however, the provincial leadership has been replaced by more technically proficient and younger leaders, who are taking over from the revolutionary generation.

(T.R.T./V.C.F.)

## Kiangsi

Kiangsi (Chiang-hsi in Wade–Giles romanization, Jiangxi in Pinyin), a province of China occupying a south central location, is bounded by the provinces of Hupeh and Anhwei on the north, Chekiang and Fukien on the east, Kwangtung on the south, and Hunan on the west. The area of the province is 63,600 square miles (164,800 square kilometres). On the map its shape resembles an inverted pear. The port of Chiu-chiang, 430 miles (692 kilometres) upstream from Shanghai and 135 miles downstream from Han-k'ou, is the province's principal outlet on the Yangtze River. The provincial capital is Nan-ch'ang.

The name Kiangsi means "West of the (Yangtze) River," although the entire province lies south of it. This seeming paradox is caused by changes made in administrative divisions throughout China's history.

Kiangsi's strategic location

Lying in the midst of a longitudinal depression between China's western highlands and the coastal ranges of Fukien Province, Kiangsi constitutes a corridor linking the province of Kwangtung, in the South, with the province of Anhwei and the Grand Canal, in the North. Throughout China's history, Kiangsi has played a pivotal role in national affairs because of its position astride the

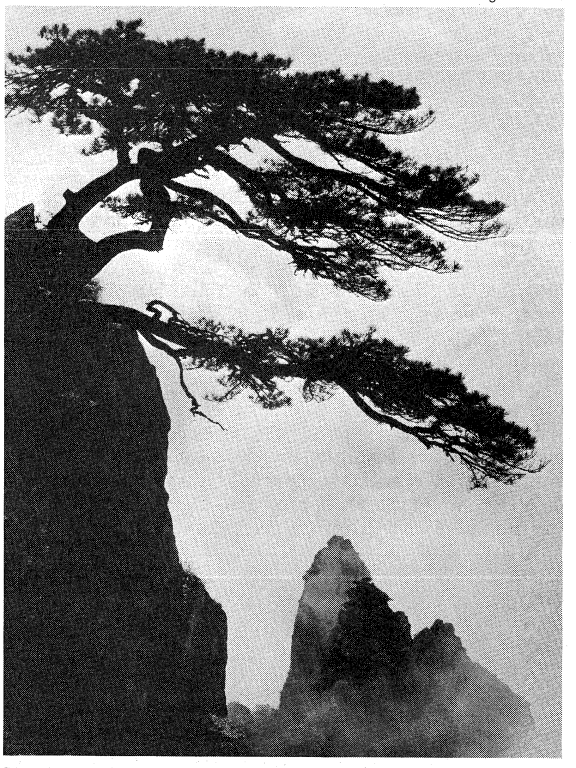
main route of armies, commerce and trade, and large population migrations.

## PHYSICAL AND HUMAN GEOGRAPHY

**The land.** *Relief.* Topographically, Kiangsi corresponds to the drainage basin of the Kan River, which runs northeastward in descending elevation from the southern tip of the province to P'o-yang Lake and the Yangtze in the north. This basin is surrounded by hills and mountains that rim the province from all sides. Among the more important ranges are the Huai-yü Mountains, to the northeast; the Wu-i Mountains, to the east; the Chiu-lien and Ta-yü ranges, to the south; the Chu-kuang, Wan-yang (including Mount Ching-kang), Wu-kung, and Chiu-ling ranges, to the west; and the Mu-fu and Lu ranges, to the northwest. A remarkable feature of these mountains is that they rise in disconnected masses and thus contain corridors for interprovincial communication, especially along the Hunan border. The mountains to the south, too, present no formidable barrier. The Mei-ling Pass is a broad and well-paved gap leading to Kwangtung Province.

Other mountains are found in the centre and north of the province. East of the Middle Kan Valley are the Yü Mountains. Made up of short and moderate hills separated by a network of streams, the country traversed by this range consists of a succession of small valleys with bottomlands from five to 12 miles wide. The Lu Mountains, in the north, rise sharply to almost 5,000 feet (1,524 metres) from the lowlands west of P'o-yang Lake.

You Yungu—Eastfoto



Pine tree overhanging peaks in the Lu Mountains in northern Kiangsi Province.

Rivers

*Drainage.* The principal river of Kiangsi is the Kan, which traverses the entire province from south to north. Its headwaters are two streams that converge to form one river at Kan-chou. Along its course this great river receives several large tributaries from the west and a lesser number of smaller tributaries from the east.

Besides the Kan, other rivers of Kiangsi form distinct basins of their own in the northeastern and northwestern parts of the province. These include the Hsin River, which rises near Yü-shan in the northeast and runs westward to P'o-yang Lake; the Ch'ang and Lo-an rivers, also in the extreme northeast of the province; and the Hsiu River, which, rising in the Mu-fu Mountains in the northwest, drains southeastward into P'o-yang Lake.

Ultimately, all Kiangsi's rivers drain into P'o-yang Lake, which is connected with the Yangtze by a wide neck at Hu-k'ou, a short distance east of the Yangtze port of Chiu-chiang. In summer, when the Yangtze rises, P'o-yang Lake gains in size and depth: its area then covers about 1,800 square miles (90 miles long by 20 miles wide), and its depth averages 65 feet. In winter, when the Yangtze waters recede, it shrinks in size, leaving shallow channels of water in many places. If the high-water stage occurs simultaneously on the Yangtze, the Kan, and other rivers, floods inevitably result. The lake also serves as a useful reservoir.

*Soils.* The soil in the plains of northern Kiangsi is alluvial and permits intensive cultivation. The hilly lands in other parts of the province have red and yellow soils. In these poorer regions few natural forests have been preserved; commercial trees planted instead include the tea, tung, camphor, bamboo, and pine. On farms with clayey red soils, where the rains have washed away the mineral contents as well as the humus, the soil requires working over and the addition of green manure or chemical fertilizers in order to become productive.

*Climate.* Situated in the subtropical belt, Kiangsi has a hot and humid summer lasting more than four months, except in spots of high elevation, such as Lu-shan. High temperatures in Nan-ch'ang in July and August average 95° F (35° C). In winter temperature variations between north and south are greater. January temperatures in the north at times fall to 25° F (−4° C), while those in the south average 39° F (4° C). Most of the province has a growing season of 10 to 11 months, thus making possible two crops of rice. Rainfall is plentiful, particularly during May and June. Average annual rainfall is 47 inches (1,194 millimetres) in the north and 60 inches in the south; in the Wu-i Mountains region it can reach 78 inches.

*Settlement patterns.* Most of Kiangsi's people live in rural areas. The leading city is Nan-ch'ang. Situated on the right bank of the Kan River, a short distance before it enters P'o-yang Lake, Nan-ch'ang is the focal point for rail and river transport, an industrial centre, and a trading centre for agricultural products. Chiu-chiang, on the south bank of the Yangtze 87 miles north of Nan-ch'ang, is the principal port through which the province's products are exported. Just south of Chiu-chiang is the beautiful resort of Ku-ling, perched 3,500 feet atop Mount Lu.

From Nan-ch'ang southward up the Kan are Chi-an, rich in literary lore and the commercial metropolis of the Middle Kan Valley, and Kan-chou, the centre of culture and trade in the Upper Kan Valley. Other cities dot the hinterland on both sides of the river. The leading city in the extreme northeast is Ching-te-chen, the porcelain capital of China. The vast stretch of country east and southeast of Nan-ch'ang contains many cities of historical and commercial importance, the largest of which is Fu-chou. The west and northwest of the province is a focus of heavy and light industry, of which P'ing-hsiang, on the Hunan border, is the major centre.

**The people.** Kiangsi received successive waves of migration from North China through the ages. Its population is virtually all Han (Chinese); minority groups include the Miao, Yao, and Hui (Chinese Muslim) peoples. The Hakka, descendants of a unique group of migrants from North China, have maintained their separate identity with their own dialect and social customs.

Composition of the population

The language usually spoken is Mandarin, with a marked Lower Yangtze accent, although it has an admixture of the Fukien dialect in regions south of Kuei-hsi and is heavily tinged with the Cantonese accent in the Ta-yü region, south of Kan-chou.

**The economy.** *Agriculture.* The beautiful basin of the Kan River, together with the valleys of its many tributaries, was one of the nation's most affluent regions before trade patterns were changed by the opening of treaty ports to the Western powers in the mid-19th century. Nevertheless, Kiangsi is still one of China's richest agricultural provinces. Since 1949 the reclamation of unused land, treatment of red soil to make it more fertile, construction of irrigation projects and hydroelectric power stations, and increased use of chemical fertilizers and mechanization

has increased the amount of arable land to more than one-third of the total area of the province.

Principal crops

Food crops produced in Kiangsi include rice, sugarcane, fruits, peanuts (groundnuts), and sweet potatoes. Of these, rice is by far the most important. The P'o-yang Lake plain and Lower Kan and Hsü valleys are the principal areas of rice production; two crops a year are raised in all parts of the province. Kiangsi also produces a great variety of commercial crops: tea is grown on hillsides in many regions; ramie, used for making a fine, silky fabric, is raised south and west of P'o-yang Lake; cotton is grown on the plains northeast of the lake; tobacco is produced in the Chekiang border area; and sugarcane is raised in the northeast and in the south. Other important commercial crops include soybeans, rapeseed, and sesame seeds. Kiangsi is a great provider of fruit, especially citrus, watermelons, pears, and persimmons. The hills of the province also supply the nation's apothecaries with such important herbs as the three-foliaged orange, the greater plantain (*Plantago major*), and the gallnut; and the indigo plant is grown in the valleys east of P'o-yang Lake.

Lush forests in the region from Chi-an southward contain pine, fir, cedar, oak, and banyan. The timber produced there—used for building material and for furniture—is floated down to Chiu-chiang for export to all parts of China. No less important are the camphor tree and the giant bamboo. The timber industry also yields valuable by-products, especially tung oil, resin, turpentine, lampblack (for making Chinese ink sticks), and tea oil.

Livestock raised in Kiangsi include water buffalo, pigs, chickens, and ducks. Inland fishing is a major industry on P'o-yang Lake. In addition, fisheries are found along the numerous rivers and in the almost countless village ponds.

*Industry.* Although Kiangsi was long known for commerce and handicrafts, modern industry had only a limited base by 1949. Subsequently, however, the province made immense progress in establishing both heavy and light industries. Coal and tungsten are the most important minerals. The area around P'ing-hsiang in the west is the coking coal capital of south central China; another major coal-mining centre is Feng-ch'eng, south of Nan-ch'ang. The region surrounding Ta-yü, on the Kwangtung border, is the centre of tungsten mining, and extensive deposits have been discovered at the extreme southern tip of the province. The ore mined in southern Kiangsi contains 60 percent tungsten; the remaining 40 percent permits the production of sizable amounts of tin, bismuth, and molybdenum. Tantalum, lead, zinc, iron, manganese, copper, and salt are also mined.

Nan-ch'ang is the largest industrial centre; it has plants for a wide variety of heavy and light industrial products. Chiu-chiang has an oil refinery and a petrochemical industry, but it is principally a centre for textile mills and textile machinery. Kan-chou is also a major industrial centre. Food processing is an important enterprise in many localities throughout the province. The development of modern industry, however, has not affected the handicrafts for which Kiangsi has been famous throughout history. The ramie cloth produced in the province continues to be the nation's preferred choice for summer wear. Other important local products are the typical Kiangsi varieties of paper—*lien-shih* paper for printing (made of bamboo), *piao-hsin* paper for wrapping (also of bamboo), and *mao-pien* paper for scribing (made of rice and mulberry straw). Hsü-wan, in the southwest, is a major centre of the engraving and printing industry.

The porcelain industry

The porcelain industry, however, is the foremost activity of the province. During the reign of the Sung emperor Chen-tsung (997/998–1022/23), the town of Fou-liang, in northeastern Kiangsi, was by Imperial decree made a centre for fine porcelain. From that time on, Fou-liang was known as Ching-te-chen, after the Imperial patron's year title Ching-te. For 10 centuries it has supplied the Chinese people with porcelain ware of all descriptions—ranging from items of daily use to artistic works of rare beauty made for the enjoyment of emperors and collectors. The beautiful translucence and hardness of the porcelains from Ching-te-chen are attributable to kaolin (china clay) and petuntse (white briquette), both of which are found in the Ch'ang Valley and along the east shore of P'o-yang Lake. Most of the population of Ching-te-chen is still employed in one way or another in the making of porcelain. The bulk of the output is for domestic trade, although some items are shipped abroad. The government has been making an effort to revive and preserve the secret formulas of the Ming and Ch'ing potters, but the tendency seems to be away from handicrafts and toward mechanization.

*Transportation.* Kiangsi has an abundance of inland waterways. Most of the rivers flow diagonally, from east and west toward the centre, emptying into the Kan River and P'o-yang Lake; many are navigable. On many shallow streams, as well as on the headwaters of the Kan, navigation is by junk. Thus, there are adequate transportation facilities for all counties of the province; Nan-ch'ang and Chiu-chiang are the main centres for transshipment and distribution. Goods for export are carried by large steamships on the Yangtze.

The first major railroad in Kiangsi, built on the eve of World War I, runs north–south, linking Chiu-chiang with Nan-ch'ang. Another, the Chekiang–Kiangsi railroad, runs east–west, from the Chekiang border, westward to the Hunan border. This line forms part of a national trunk line that extends westward through Hunan into Kweichow to connect with the rail network of the southwest. Another line runs southeastward to Amoy.

Highways

Kiangsi's highways were well developed in the Nationalist period. Many new roads have since been added. The focal centres for the highway system—Nan-ch'ang, Lin-ch'uan, Shang-jao, Chi-an, and Kan-chou—are the hubs of regional road networks and the termini of interprovincial highways.

**Administration and social conditions.** *Government.* From 1950 to 1954 Kiangsi was part of the Central South greater administrative region. In 1954 Kiangsi Province became directly subject to the central government. Kiangsi's administrative divisions are arranged in a hierarchy of levels. Immediately below the province (*sheng*) there are five prefectures (*ti-ch'ü*) and six prefecture-level municipalities (*shih*). Below that level are counties (*hsien*) and county-level municipalities (*shih*). The lowest political units are the townships.

*Education.* During the 1950s Kiangsi served as a laboratory for a number of revolutionary educational experiments. Perhaps the most significant innovation in higher education was the Kiangsi Labour University, founded in 1958. It has its main campus in Nan-ch'ang but operates a network of branch campuses, in addition to affiliated technical schools, throughout the province. Aiming at the development of productive work through the dissemination of advanced education, the branch campuses have pioneered a multiplicity of development projects, including building roads in mountainous areas, founding new villages, reclaiming land, building factories, and promoting afforestation.

Centres of higher learning include the Kiangsi branch of the Academia Sinica (Chinese Academy of Sciences), the Kiangsi Library, the Kiangsi Provincial Museum, the Kiangsi Agricultural Institute, and the Kiangsi Medical College, all located in Nan-ch'ang. Popular education has also made advances, and some three-fifths of the population has at least a primary-level education. The adult literacy rate is at the national average.

*Health and welfare.* Before 1949 the greatest scourge was the prevalence of malaria. This debilitating disease annually took a heavy toll of lives. Since 1949 draining the swamps and pools of stagnant water, the breeding grounds of the disease-carrying *Anopheles* mosquito, and measures taken for epidemic prevention have reduced malaria to a minimum. Another menace to health peculiar to the P'o-yang Lake region was liver fluke (a kind of flatworm). Many thousands of lives were previously lost every year from this parasite, but this disease, too, is rapidly becoming a danger of the past, following mass control of the fluke embryo in the lake and surrounding waters.

Curative medicine

In curative medicine, many improvements have been made. Clinics providing free medical care have been made widely available, while modern hospitals have been established in all cities and counties.

An adequate social welfare program is available. For industrial workers there are measures for accident prevention, as well as insurance programs that provide for hospital treatment, sick leave, disability compensation, maternity leave, and old-age and death benefits. Extra benefits are available based upon cooperation with government policies, such as birth control. In Nan-ch'ang and other industrial towns and in the countryside, the government has constructed new housing and expanded recreational facilities.

**Cultural life.** For nearly 2,000 years the people of Kiangsi lived under the pervading influence of Confucian culture. With village life rooted in intensive agriculture and government in the hands of the landlord-scholar-officials, the dynamics of society were regulated by Confucian ethics. Such a culture gave the province many famous people. Besides T'ao Yüan-ming (a great Chin poet of the reclusive life), Chu Hsi (the Sung dynasty Neo-Confucian philosopher), and Wang Yang-ming (the Ming philosopher), all of whom either taught or lived there, Kiangsi produced a full quota of statesmen during both the Sung and the Ming dynasties.

Yet, despite the dominance of Confucian learning and culture, peasant rebellions were a strong tradition in the province. The first major revolutionary base of the Chinese Communist Party was at Jui-chin, in southeastern Kiangsi, and an uprising in 1927 at Nan-ch'ang serves as the founding date of the Red Army.

HISTORY

From 770 to 476 BC, during the Ch'un-ch'iu (Spring and Autumn period) of the Chou dynasty, Kiangsi was a part of the kingdom of Ch'u. During the Chan-kuo (Warring States period; 475–221 BC) the territory east of P'o-yang Lake was annexed by the kingdom of Wu. When a unified empire was established under the Han dynasty (206 BC–AD 220), Kiangsi became the western portion of the large province of Yang-chou and grew rapidly in population and culture.

From 220 to 589, the period of the Six Dynasties, large numbers of families from North China, fleeing the Tatar invaders, settled in Kiangsi. Initially, there were clashes between the northern newcomers and the original inhabitants. In time mutual accommodation prevailed, and the province benefited immensely from the introduction of northern arts, culture, and administrative skills. It was during this period that the Kan Valley became the main highway of the empire. Under the T'ang dynasty (618–907) the growth of commerce and population in Kiangsi was even greater than in earlier times. This was caused first by the opening of the Grand Canal, linking Lo-yang with the Lower Yangtze, and second by a new influx of people from North China. Equally noteworthy was the spread of Buddhism in this period.

Period of Confucian domination

In the Sung dynasty (960–1279) Kiangsi became a model of the Confucian state, governed by scholar-officials. The Pai Lu Tung ("White Deer Grotto") Academy, near Lu-shan, where Chu Hsi taught, became a renowned centre of Confucian learning. From 1069 to 1076 Wang An-shih, a native of Lin-ch'uan, southeast of Nan-ch'ang, was prime minister; Wang introduced reforms to curb the rich and help the poor, only to be overthrown by the conservative champions of the traditional order. In the late Sung period and throughout the era of the Mongol conquest, Kiangsi's cultural and political vigour declined. Such was the obscurantism of the government that it sanctioned a Taoist "papacy" at Mount Lung-hu, near Kuei-hsi, which lasted into the mid-20th century.

In the early years of the Ming dynasty (1368–1644) Kiangsi produced a number of great statesmen, but after a time the government's despotic tax program alienated the people. From the early 16th century onward, peasant brigands living in the hills fought the government. The widespread unrest was ended after the Ch'ing dynasty (1644–1911/12) reunified the country. During this period of prolonged peace Kiangsi again became one of the richest regions of China, but its days of prosperity ended in the mid-19th century, when the Yangtze Valley was devastated by the great Taiping Rebellion against the ruling Ch'ing dynasty and when treaties with the Western powers diverted trade to coastal regions.

In the first half of the 20th century Kiangsi became a focal point for revolution and war. After the 1911–12 revolution the province fell victim to warlord rule, until Chiang Kai-shek brought it under Nationalist control in 1926. Chiang's break with the Communists, however, made Kiangsi a bone of contention between the two sides. An uprising was staged in Nan-ch'ang by the Communists in 1927, followed by the establishment of peasant bases in the southern counties under the Communist leaders Mao Zedong (Mao Tse-tung) and Zhu De (Chu Teh). Such was the growth of their strength that, in 1931, Jui-chin was declared the capital of the Chinese Soviet Republic. In the continuing struggle the Communist guerrillas withstood Chiang's "annihilation campaigns," but his use of an economic blockade forced the Communists to flee Kiangsi on their Long March (1934–35) to northwestern China. Chiang then briefly regained control of southern Kiangsi, and Nationalist government reforms were undertaken on an experimental basis in 1934–37. From 1938 to 1945 Kiangsi was under Japanese occupation. The Communists carried on guerrilla activities inside Kiangsi throughout the period.

After the Japanese withdrawal Communist guerrillas dominated the countryside, while the Nationalist government took precarious control of the cities. In 1949 Communist forces crossed the Yangtze from the north and took possession of the province. Kiangsi then entered an era of stability and progress, and many new economic and social developments were pioneered there. (V.C.F.)

## Kwangtung

Kwangtung (Kuang-tung in Wade–Giles romanization, Guangdong in Pinyin), the southernmost of the mainland provinces of China, also constitutes the region through which South China's trade is primarily channeled. Kwangtung has one of the longest coastlines of any province in China. The province has an area of 76,100 square miles (197,100 square kilometres). Kwangtung is bounded by the Chuang Autonomous Region of Kwangsi to the west, by the provinces of Hunan and Kiangsi to the north and Fukien to the northeast, and by the South China Sea to the south. Two foreign holdings remain on the coast of Kwangtung—the British crown colony of Hong Kong and the Portuguese territory of Macao. The capital is Canton (Pinyin, Guangzhou).

Historically Kwangtung and Kwangsi often were jointly governed. Kwangtung was first administered as a separate entity in AD 997; it was from this time that the term Kwangtung (Chinese, "Eastern Expanses") began to be used. Kwangtung has its own physical and cultural identity. Its topography separates it somewhat from the rest of China, and this factor, together with its long coastline, its contact with other countries through its overseas emigrants, and its early exposure to Western influence through the port of Canton, has resulted in the emergence of a degree of self-sufficiency and separatism. Canton dominates the province to an unusual extent.

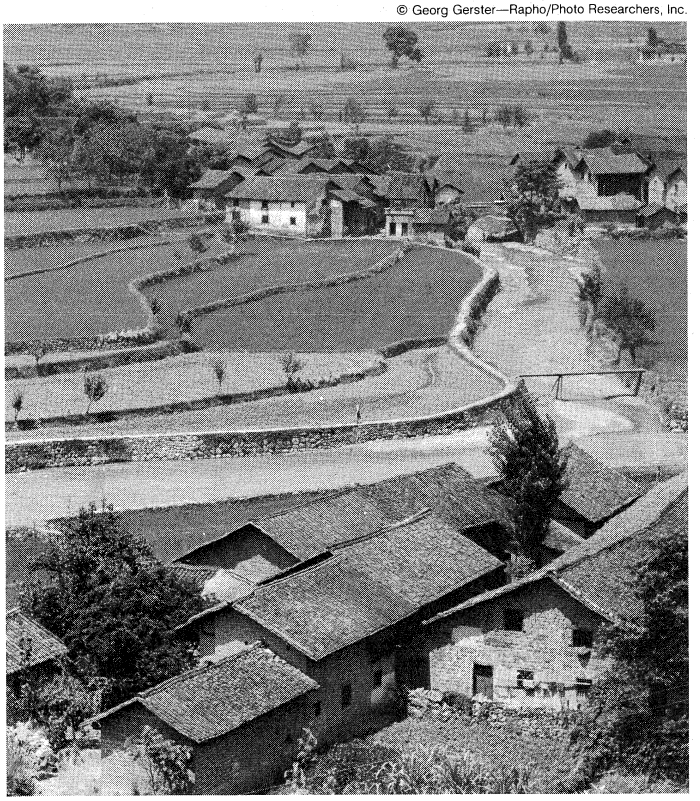
PHYSICAL AND HUMAN GEOGRAPHY

**The land.** *Relief.* The surface configuration in Kwangtung is diverse, being composed primarily of rounded hills, cut by streams and rivers, and scattered and ribbon-like alluvial valleys. Together with the Kwangsi region, Kwangtung is clearly separated from the Yangtze Basin by the Nan Mountains, the southernmost of the three major Chinese mountain ranges running from east to west. The greater part of eastern Kwangtung consists of the southerly extension of the Southern Uplands, which stretch down from Fukien and Chekiang provinces. A series of longitudinal valleys running from northeast to southwest extends as far as the vicinity of Canton. Smooth, low hills cover about 70 percent of the province. Most peaks range in elevation from 1,500 to 2,500 feet (450 to 750 metres), with a few reaching 5,500 feet. Level land of any size is primarily found in the alluvial deltas, formed where rivers empty into the South China Sea.

Pearl River Delta

*Drainage.* Of great extent and importance, the Pearl River Delta, measuring about 3,500 square miles, is marked by hilly outliers and by a labyrinth of canalized channels and distributaries totaling some 1,500 miles in length. The delta marks the convergence of the three major rivers of the Pearl River system—the Hsi (West), Pei (North), and Tung (East) rivers. The Pearl River is the name given to the lower course of the Hsi beyond the confluence. Entirely rain-fed, these rivers, which are subject to extreme seasonal fluctuations, collect so much water that, anomalously, the Pearl system discharges annually six and a half times as much water as the Huang Ho (Yellow River), although its basin area is only about half as large. Altogether, Kwangtung has some 1,300 large and small rivers. The Han is the most important river outside the Pearl system. Other important rivers and lowlands are located in the southwest.

© Georg Gerster—Rapho/Photo Researchers, Inc.



Waterway serving a farming village in the Pearl River Delta region of Kwangtung Province.

*Soils.* In general, soils are poor, as high temperatures and plentiful rainfall result in podzolization (bleaching) and leaching. Almost all of western Kwangtung is covered with mature red soils, whereas the rest of the province is covered with a mixture of old and young red soils that usually have been subjected to a high degree of podzolization. In the wettest and hottest parts of Kwangtung, lateritic (heavily leached, iron-bearing) soils are common; like the red soils, they do not resist erosion and require substantial fertilization in their cultivation. Yellow soils are found in the wettest and coolest parts of Kwangtung, occurring in small pockets of flatland with imperfect drainage.

Of more limited distribution but of greater economic significance are the alluviums deposited in the river valleys and deltas. As a result of the cultivation of rice, the alluviums have developed special morphological characteristics, the most striking of which is the formation of iron hardpans (hard impervious layers composed chiefly of clay) in the zone of the fluctuating water table.

*Climate.* Since much of Kwangtung lies south of the Tropic of Cancer, it is the only Chinese province with tropical and subtropical climates. The average July temperature in the Hsi Valley, which is from 82° to 86° F (28° to 30° C), is little different from temperatures in the lower Yangtze and on the Huang Ho, but the average January temperature is considerably higher, ranging from 55° to 61° F (13° to 16° C). Except at higher elevations, frost is rare, so that almost the entire province lies within the area where two crops of rice can be grown. True winter does not occur in the province, but the hot summer varies in length from about 10 months in the south to six months in the north.

The rainfall regime shows a pronounced summer maximum, with the rainy season lasting from mid-April, when Kwangtung starts to be dominated by moisture-laden tropical air masses from the Equator and the Indian Ocean, until mid-October. More than half of the total precipitation falls between June and August. The months between July and September form the main typhoon season, which ordinarily is accompanied by heavy rains and widespread destruction. The driest period is from December to February. Kwangtung's annual rainfall is approximately 60 to 80 inches (1,500 to 2,000 millimetres), decreasing with distance from the coast to the northwest but increasing with altitude and exposure to the prevailing summer monsoon winds.

The typhoon season

*Plant life.* Abundant moisture, moderate to high temperatures, and variegated physiography support luxuriant and highly diversified plant growth. Broad-leaved evergreen forests, intermixed with coniferous and deciduous trees, originally covered much of the land, while a more tropical type of vegetation predominates on the south coast. With the exception of the more remote mountainous areas, much of this natural vegetation cover has been stripped by fire and by the use of trees and shrubs for fuel. This circumstance, together with millennia of uninterrupted cultivation, has resulted in much of the natural vegetation now taking the form of secondary forests of hardwoods and horsetail pine. On the more severely eroded hills, coarse grasses and ferns have taken hold. Bamboo groves, varying greatly in height and extent, are widespread, particularly in humid river valleys. The most productive and least disturbed forests cover the mountainous areas. Certain trees, notably camphor, have been revered and protected for centuries and are found around grave plots and cultivated fields. Since 1949 massive afforestation programs have been undertaken. In the highlands, where coniferous and deciduous species thrive together, the broad-leaved evergreen forests are characterized by tropical oaks, tan oaks (oaks that yield tannin), and chestnut oaks (or chinquapins). The more significant coniferous species of economic value include horsetail pine, Chinese fir, and Chinese hemlock. Some of the species of cypress and pine are little known outside China. Truly tropical monsoon rain forests are common in the south.

*Animal life.* Among the mammals found in Kwangtung are many tropical bats, and squirrels, mice, and rats of many species are abundant. Insectivores are generally more diverse than in other regions of China, and carnivores are exemplified by civet cats and small-clawed otters. Types of birds vary according to habitat. In the tropical forest, wildfowl, peacocks, and silver pheasants are common. Reptiles are more restricted in distribution. Kwangtung has a number of pit vipers, including the huge and deadly Chinese vipers and bamboo vipers, as well as nonpoisonous pythons, which are up to 20 feet long. Insects of every description—crickets, butterflies, dragonflies, grasshoppers, cicadas, and beetles—are found in profusion. Amphibians include ground burrowers and many types of frogs and toads.

Tigers, rhinoceroses, panthers, wolves, bears, and foxes once roamed the hills of Kwangtung, but their numbers have been decimated by forest fires and persistent deforestation; they are now considered to be extinct in the area. In the tropical monsoon forest, however, a great number of animals, many of which live in the trees, still remain.

*Settlement patterns.* Most of the people of the province live in villages, which remain the basic functional units in the countryside. Population distribution bears almost a one-to-one correlation with agricultural productivity. The greatest numbers of villages are in the fertile river deltas and along the waterways. To an even greater extent, towns and cities are located in the deltas and coastal areas

Urbanization

and along major communication lines. The most highly urbanized area within the province is the Pearl River Delta, where almost two-thirds of the population lives in urban areas. Kwangtung is a relatively highly urbanized province for China, with about one-fifth of its population being classified as urban. The urban hierarchy is headed by Canton. It is far greater in size than the second largest city, Swatow (Shan-t'ou). Other important municipalities include Shao-kuan, Chiang-men, and Chan-chiang. Canton and Chan-chiang have been designated "open" coastal cities and have become central to the planning of the province's economic future.

**The people.** Kwangtung is populated largely by the Han (Chinese), the other ethnic minorities totaling only a tiny portion of its population. The Yao are the largest ethnic minority in Kwangtung and are concentrated principally near its northwestern border in autonomous counties. A heavily Sinicized group, the Chuang-chia, live in Lien-shan, and the She live in the northeast around Ch'ao-an. The Ching were transferred to Kwangsi in 1965, when the multinational Tung-hsing Autonomous County in extreme southwestern Kwangtung changed its provincial jurisdiction. The so-called Tan, or Tanka, the Boat People, are not officially designated as a national minority. Whereas some scholars believe they are descendants of aborigines, others regard them as simply a people who live on boats and speak Cantonese. They generally live along the rivers in the Pearl Basin as well as along the coast.

The relative ethnic homogeneity prevailing in Kwangtung stands in contrast to the great diversity of dialects. By far the most important dialect is Cantonese, spoken in central and western Kwangtung. There is considerable variety among the Cantonese speakers, but the form spoken in Canton is generally regarded as the standard. Hakka is another important dialect; it predominates in the north and northeast of the province. Offshoots of Hakka are common in central Kwangtung. A third major dialect, Min-nan (or south Fukien dialect), is spoken mostly along an eastern coastal area centred on Swatow.

In addition to these Han dialects, there are the languages and dialects of the ethnic minorities. New scripts have been created for a number of these languages. They not only are taught in minority-area schools but also are used in conjunction with Chinese in official communications in minority communities.

**The economy.** *Agriculture.* The economic foundation of Kwangtung is primarily agriculture. Rice is the leading crop. Since less than one-fifth of the land is under cultivation, agriculture is of necessity extremely intensive; but the limited extent of sown land available is partly offset by repeated use of it. Progress in irrigation and flood control has made water control possible for almost all of the cultivated area, producing good rice yields. Farming and irrigation have become increasingly mechanized, with more reliance placed on the use of chemical fertilizers.

Two crops of rice a year can be grown on most cultivated land, and in the Pearl River Delta three crops are not unusual. Thus, although average yields per harvest are below the national average, annual yields exceed the average. Although food-grain crops occupy almost all of the total cultivated area, the industrial and fruit crops grown on the remaining land are of national importance. Kwangtung annually produces much of China's total output of sugarcane. In tropical Kwangtung, a number of industrial crops are successfully raised, including rubber, sisal, palm oil, hemp, coffee, and black pepper. Other traditional agricultural products include sweet potatoes, peanuts (groundnuts), and tea. No less than 300 types of fruits are grown, among the more representative of which are citrus, litchi, pineapples, and bananas.

Fishing

Kwangtung, with its long coastline, produces about one-fifth of China's fish. Fish production accounts for as much as one-third of the income of some localities. More than 400 species of saltwater fish, including yellow croaker, white herring, mackerel, golden thread, and pomfret, are caught from numerous fishing ports.

*Industry.* In the first half of the 20th century Kwangtung experienced modern growth as Canton developed into an industrial, commercial, and transportation centre. But because of the paucity of its iron deposits, Kwangtung received only scant attention during the First Five-Year Plan (1953–57). The discovery of other mineral deposits, however, prompted the development of some heavier industries, including metal and petrochemical processing, the manufacture of machinery, and shipbuilding and ship repairing. A large proportion of these industries is still concentrated in Canton.

Coal reserves and manganese deposits are located on the Lei-chou Peninsula; quantities of oil-shale deposits have also been discovered there. Tungsten, which is associated with bismuth, molybdenum, and tin deposits, is mined near the Kiangsi border, where uranium is also found. The province also produces some lead and antimony.

Light industry has always been of significance in the province. Apart from handicrafts, light industry—especially food processing and the manufacture of textiles—accounts for about two-thirds of industrial production. Almost all of the major light industries are located in the Pearl River Delta. The largest and most widespread industry is rice milling, which takes place in nearly every county and municipality. Kwangtung's light industrial production has grown dramatically, partly because of the province's level of exports. Kwangtung has been given special authority to develop trade and investment ties with other countries; three of China's first four special economic zones were established in the province.

*Transportation.* Economically and culturally, the different regions of Kwangtung are linked by the waterways of the Pearl River system. In addition, a number of coastal and international shipping routes are variously linked to more than 100 large and small ports. The leading ports, including Huang-p'u (Canton's seaport), Chan-chiang, and Swatow, are of national significance. Water transport accounts for more than two-fifths of Kwangtung's total traffic tonnage. The waterways are maintained by continually dredging, widening, and clearing channels.

Water transport

Connections with other provinces depend principally on land transportation. Kwangtung has developed the best highway network in China, running primarily along river valleys. Interprovincial links, both for highways and railroads, usually run north–south. The vital Canton–Hank'ou railroad was double-tracked in the early 1960s. The low priority placed on east–west transport is indicated by the absence of a railroad running parallel to the Hsi River and by the fact that the Canton–Chan-chiang line was only opened in 1963.

Kwangtung provides a crucial link in China's domestic and international civil aviation routes. Air services connect the province to numerous international cities. To cope with the increasing traffic, Canton's Pai-yün airport has been enlarged and modernized.

**Administration and social conditions.** *Government.* The administrative system in Kwangtung has undergone many changes since 1949. Autonomous administrative units were established in the early 1950s for areas with large ethnic minority populations. The status of Canton was changed in 1954 from a centrally administered municipality (*t'eh-pieh-shih*) to a prefecture-level municipality (*shih*) under the jurisdiction of the provincial government. In addition to the municipality of Canton, Kwangtung is subdivided into seven other prefecture-level municipalities and three prefectures (*ti-ch'ü*). Kwangtung is further divided into counties (*hsien*). Rural administration was reorganized in 1958 when communization replaced the administrative villages, market towns, and municipal districts. In 1980–81 the government implemented a policy of greater decentralized economic management, and the communes lost their administrative role.

*Education.* Education, health, and other social conditions in Kwangtung have generally been improved since 1949. There are now many more kindergartens and nurseries for preschool education, secondary schools, and post-secondary schools and universities. Repeated campaigns have succeeded in reducing illiteracy throughout the province. Special attention has been given to the education of the ethnic minorities. New schools, including a national minority college, have been established in minority communities.

Hospitals, clinics, and medical services

*Health and welfare.* In general, hospitals, clinics, and many health stations, including maternity centres, are available at the local level. Better equipped and better staffed hospitals are maintained at the county and provincial levels. Medical education has been greatly expanded and includes a college devoted to Chinese medicine (acupuncture and herbal medicine). Many short-term medical-training classes are organized for health workers assigned to rural areas. The development of medical services, coupled with the general improvement in sanitation and health education, has succeeded in eliminating many previously common diseases such as malaria, schistosomiasis, and filariasis.

**Cultural life.** Kwangtung has long been noted for the distinctive cultural traits of its people, as evidenced by the variety of dialects spoken. Kwangtung is famous for its two types of local opera: the Yüeh Opera and the Ch'ao Opera, which are popular among the Cantonese and Fukienese communities, respectively. Kwangtung also has some characteristic puppet plays. The hand puppets of Canton are distinguished by their size—they are between three and four feet high—and by the beautiful carving of their wooden heads. Many places in Kwangtung have distinctive forms of folk art; examples are the woodcuts of Ch'ao-an and the stone engravings of Shun-te.

Cantonese food is widely recognized as among the best in China. Living in a coastal province, the people are particularly fond of seafood. Especially in winter, the "big-headed fish" (tench) is often served raw in a fish salad—a departure from habitual Chinese culinary practice. Some other food habits, such as the eating of newborn rats, monkey's brain, and fried snake, are regarded as revolting by most Chinese in other provinces. Chinese who have returned from Southeast Asia have popularized the chewing of betel nut wrapped in *Celosia* (cockscomb) leaves.

Religions

Ancestor worship, folk religions, and all the institutional religions of Taoism, Buddhism, Christianity, and Islām coexist in the province, as they do in most places in China. Among these religions, ancestor worship has the most pervasive influence. Although some folk religions are national in outlook, others are of a more regional or local character, such as the worship of the goddess of fishing and navigation, T'ien-hou Sheng-mu. With the possible exception of Muslims and Christians, people in Kwangtung are polytheistic, visiting temples or priests of different faiths as occasions demand.

Kwangtung is a province where lineage—an important social institution in China—has been emphasized. The importance of ancestry is often reflected in the settlement pattern of lineage groups. The inhabitants of many villages belong exclusively to one or two lineages. In such villages, community and lineage organizations are virtually identical. Conflicts between lineages were once common and often took the form of community strife, with bitter vendettas sometimes lasting for long periods of time.

With the founding of the new regime in 1949, systematic efforts were made to change these cultural patterns in accordance with governmental ideology and policy, although in the early 1980s limited religious practice was again allowed. On the other hand, many aspects of traditional culture, especially the folk arts and the theatre, were revived and extolled.

HISTORY

Physically separated from the early centres of Chinese civilization in North China, Kwangtung was originally occupied by non-Han ethnic groups. It was first incorporated into the Chinese empire in 222 BC, when Shih huang-ti, first emperor of the Ch'in dynasty, conquered the area along the Hsi and Pei river valleys down to the Pearl River Delta. In 111 BC Chinese domination was extended to the whole of what is now Kwangtung, including Hai-nan, by Wu-ti of the Han dynasty. The conquest, however, was not followed by successful colonization, and Kwangtung remained part of the empire only politically.

During the five centuries of the Sui, T'ang, and Pei (Northern) Sung dynasties, from AD 581 to 1127, the military and agricultural colonization of Kwangtung gradually took place. This colonization, combined with increasing overseas trade channeled through Canton, led to an increase of migration into Kwangtung and to the emergence of Canton as a metropolis with a population of hundreds of thousands. At the end of the period, however, Kwangtung was still occupied predominantly by its original ethnic population. The region was viewed as a semicivilized frontier, and disgraced officials often were exiled there.

The southward thrust of the Han was greatly intensified from 1126, when the Juchen of the Chin dynasty captured the Pei Sung capital at what is now K'ai-feng, forcing the Sung to migrate south. Another major population movement followed a century and a half later as China fell to the Mongols. These migrations marked the beginning of effective Han occupation and the rapid cultural development of Kwangtung. Especially after the 16th century the growth of population was so fast that, by the late 17th century, Kwangtung had become an area from which emigration took place. Migrants from Kwangtung moved first to Kwangsi, Szechwan, and Taiwan and then in the mid-19th century began to pour into Southeast Asia and North America, and some were also taken as indentured labourers to British, French, and Dutch colonies.

Occupation by the Han

Since the mid-19th century, Kwangtung has produced a number of prominent political and military, as well as intellectual, leaders. Many of the leaders of political movements during this period—such as Hung Hsiu-ch'uan, leader of the Taiping Rebellion (1850–64); K'ang Yu-wei and Liang Ch'i-ch'ao of the Reform Movement (1898); and Sun Yat-sen, who led the republican revolution of 1911–12—had associations with Kwangtung.

In the 1920s Chiang Kai-shek made Canton the base from which his program to reunify China under Nationalist rule was launched. Foreign privileges in the city were reduced, and modernization of the economy was undertaken. The almost simultaneous rise of the Communist movement and the advent of Japanese aggression in the 1930s, however, thwarted the plans of Chiang and the Nationalists. From 1939 to 1945 the Japanese occupied Kwangtung Province. After World War II the conflict between the Communists and the Nationalists erupted into full-scale civil war and continued until the Communist victory in late 1949. (Y.-m.Y./C.-t.C./V.C.F.)

## Taiwan

For coverage of Taiwan, see the article TAIWAN.

# SOUTHWEST CHINA

## Kweichow

Located in the southwestern part of China, the province of Kweichow (Kuei-chou in Wade–Giles romanization, Guizhou in Pinyin) is bounded on the north by Szechwan Province, on the east by Hunan Province, on the south by the Chuang Autonomous Region of Kwangsi, and on the west by Yunnan Province. Kweichow measures more than 350 miles (560 kilometres) from east to west and about 320 miles from north to south. It has an area of 68,000 square miles (176,100 square kilometres). The provincial capital is centrally located at Kuei-yang.

Kweichow has the frontier character of other southwestern plateau lands: rough topography, difficult communication and consequent isolation, and many ethnic minority groups. It has long been considered one of China's poorest and most disadvantaged provinces, as characterized by the folk poem: "The sky is not three days clear; the land is not level for three *li* (2,115 feet, or 645 metres); the people don't have three cents."

PHYSICAL AND HUMAN GEOGRAPHY

**The land.** *Relief.* Kweichow Province is part of an old eroded plateau, variously known as the Kweichow Plateau

or Yunkwei Plateau, which connects with plateau areas in Yunnan Province. Situated between the Plateau of Tibet and the hilly regions of Hunan and Kwangsi, the plateau forms part of a continuously ascending profile of the southwest, its altitude increasing from about 2,300 feet in eastern Kweichow to about 6,600 feet in the west. The Szechwan Basin to the north and the Kwangsi Basin to the south are both the results of faulting. The entire terrain of Kweichow thus slopes at a steep angle from the centre toward the north, east, and south. In areas adjacent to Szechwan and Hunan in the north and east, the elevation is about 2,300 feet, while the province's southern slopes descend 1,600–2,000 feet into Kwangsi. Accordingly, rivers in the province flow in three directions, north, east, and south.

The plateau, which is mostly of limestone and basalt, has undergone complicated and extensive folding, faulting, and stream erosion and consequently has abrupt relief, an example of which is the famous Huang-kuo-shu Waterfall near An-shun in the southwest. Incised valleys, steep gorges, and cliffs are very common. In the limestone areas the landscape is karstic (characterized by precipitous slopes, abrupt, protuberant mountains, caverns, and subterranean streams). Only the anticline (upfold of stratified rock) and syncline (downfold of stratified rock) of the plateau in central Kweichow are broad and relatively flat.

Karstic features of the landscape

*Drainage and soils.* Most of the rivers in Kweichow are the upper streams of large rivers, such as the Yangtze and the Hsi. The abrupt change of gradient, the great fluctuation in the flow volumes, and the many rapids and reefs make them unsuitable for navigation, though they have enormous hydroelectric power potential.

Because of the high humidity, a yellow soil with a yellowish-brown subsoil originated from sandstone, shale, and clay constitutes the largest area in the province. In the limestone area in the south there are broad areas of red soil. In the west the red soils are originated from basalt and sandstone and developed under a relatively drier climate.

*Climate.* Kweichow enjoys a mild climate with warm summers and mild winters. Kuei-yang has a mean July temperature of about 76° F (24° C), lower than that of all other cities to the east on the same latitude. This is due to its high altitude and the cloudiness of the summer months. In winter, cold air from Siberia cannot easily reach Kweichow because of the barrier effect of the Tsin-ling Mountains to the north of the Szechwan Basin. In spite of its high altitude, Kweichow thus has few snowy days and even fewer freezing days. The mean January temperature at Kuei-yang is about 41° F (5° C).

Rainfall is fairly uniform and plentiful, with an annual average of 31–51 inches (787–1,295 millimetres), decreasing toward the north and west. The southern and eastern parts of Kweichow are open to the influence of the moist maritime air mass in summer. For the same reason, there is a summer maximum in rainfall, averaging 45 percent of the annual total. About 25 percent falls in spring and 23 percent in autumn. Typically, the province has high relative humidity, lengthy cloudy and rainy days, and little sunshine. The capital, Kuei-yang, has more than 260 rainy and cloudy days in an average year. Most of the precipitation results from frontal activity, though some is a result of convection or condensation.

*Plant and animal life.* Because of the steep gradient and the exposure of limestone, wasteland accounts for nearly half of the total area. Yet part of the province's natural wealth lies in its forests. The plateau surface is mostly dry and barren, but the peripheral valleys have rich and valuable woodlands. About one-tenth of the land area is under natural forest. There are four main forested areas: the drainage areas of the Ch'ing-shui River in the east, the Jung River in the southeast, the Nan-p'an and Pei-p'an rivers in the southwest, and the Ta-lou Mountains in the north.

Forest resources

The forests of the northern valleys, still among the most important in China, consist chiefly of conifers and other trees, such as the tung tree, lacquer tree, camellia, birch, maple, pine, and fir. Forests in the southeast produce camphor, banyan tree, and other broad-leaved varieties. Trees of the southern subtropical valleys typically include willow, cedar, bamboo, and various species of pine and fir. Oak, Yunnan pine, *Hua-shan* pine, and camphor are grown in the west near Yunnan. Cedar, cypress, poplar, and palm trees are also found in the province.

In addition to domesticated animals, such as buffalo, horses, donkeys, asses, and pigs, the province's fauna includes leopards, otters, foxes, badgers, tigers, and squirrels. In most of the larger rivers carp and savoury fish are abundant.

**The people.** About three-fourths of the province's population is Han (Chinese). Non-Han tribesmen account for the remainder. An ethnic frontier, Kweichow has a large number of minority peoples. At least 30 different groups have been identified. Of these the most important are the Miao, the Puyi, the Yi, the Tung, the Shui, and the Chuang. All of the minority groups intermingle with Han people. Only at the low *hsiang,* or village level, can one find any exclusive racial grouping. Generally, very few minority people live in northern Kweichow, particularly in areas north of the Wu River. The Miao are mainly found in southeastern Kweichow, especially in the drainage area of Ch'ing-shui River and in the Miao Mountains. Most of the Puyi live in south central and southwestern Kweichow in the P'an River drainage area, including the suburbs of Kuei-yang. The Tung are found mainly in the southeastern areas adjacent to Hunan Province and the Chuang Autonomous Region of Kwangsi. The Shui concentrate in southern Kweichow, around San-tu and Li-po, while the Yi, who once were rulers of this frontier region, are scattered in western Kweichow. The Hui (Chinese Muslims) in Kweichow migrated there from Yunnan in the late Ch'ing dynasty after the defeat of a local rebellion. They are found chiefly in towns and cities along the main lines of communication in western and southern Kweichow, especially in Wei-ning.

Main ethnic groups

Most of the population is rural, and agriculture is the chief occupation. Rice cultivators dominate the peripheral valleys of the plateau. On the plateau itself, the Miao practice a more primitive form of agriculture, growing subsistence upland crops. Most of the Puyi live on level lands in the valleys and cultivate rice. While the Tung are experienced lowland rice cultivators, they are also skillful in forestry and in growing upland crops. The Shui, living together in large families and tribes, are rice cultivators as well. In addition to growing upland crops, the Yi undertake animal husbandry.

There are few cities in Kweichow. Kuei-yang is the most important, although larger and more populous is Liu-p'an-shui, a municipality created by combining the Liu-chih, Pan-hsien, and Shui-ch'eng special districts in Kweichow's coal-rich west. Most of the other cities are the seats of government and are the economic and communications centres for the various regions of the province.

Chinese is the common language of the Han and the Hui in Kweichow, Mandarin being spoken almost exclusively by the former group. Among other minority peoples, only the Miao, Chuang, Yi, and Puyi have their own languages. The languages of the Puyi, Shui, Tung, and Chuang are Tai languages. Those of the Miao and the Yao belong to the Miao-Yao group, and that of the Yi to the Tibeto-Burman group.

**Economy.** *Resources.* Kweichow has rich mineral resources. Its most widespread metallic mineral is mercury, reserves of which are large; there are also small deposits of manganese, zinc, lead, antimony, aluminum, copper, iron, and gold. Its most nonmetallic minerals include coal, petroleum, oil shale, phosphate, gypsum, arsenic, limestone, and fluorite. Extractive industries are consequently very important in Kweichow.

*Agriculture.* Most of the cultivated area of Kweichow is under grain crops, the most important of which is rice. Corn (maize) is grown chiefly in eastern Kweichow. Other food crops include wheat, barley, sweet potatoes, potatoes, oats, and broad beans. Increasingly more of the cultivated area is under industrial crops, of which the most important is rapeseed, followed by tobacco, peanuts (groundnuts), sugarcane, jute, tea, sugar beets, hemp, and sesame. Kweichow is also known for its production of *mao-tai* liquor, made from wheat and kaoliang.

Main food crops

Timber and other forestry products are plentiful. Kweichow ranks among the leading provinces in the production of raw lacquer and tung oil. Other important forestry products include camellia oil, cypress oil, gallnut extract, lichens, and various medicinal herbs.

*Industry.* With a sufficient local supply of raw material and fuel and an expanding market in southwestern China, an iron and steel industry has been developed in Kweichow; machinery manufacturing has also been established, primarily for the production of mining machinery, agricultural and irrigation equipment, steel-rolling machines, and steel-smelting and other smelting equipment. With its abundant water resources, the province has developed numerous small hydroelectric facilities. The local supply of phosphate and other raw materials has given rise to a chemical industry that produces chemical fertilizers, soda acid, and other chemicals; petrochemicals are also made. The textile industry has grown significantly. Based on good timber resources, paper mills are found in several cities.

Kuei-yang is the most important industrial centre of the province, producing a wide variety of heavy and light industrial goods. Other major centres include Tsun-i, which has a considerable amount of heavy industry and is also the focus of silk textile production, and the municipality of Liu-p'an-shui, with its coal-based extractive and other heavy industries.

*Transportation.* River transportation is of little importance in Kweichow due to the presence of reefs and rapids, although the Wu River is a prosperous waterway. Several of the province's other rivers are partially navigable. Kweichow's highways are relatively well-developed. Since 1958 all counties have been connected by roads, the majority of which have all-weather surfaces. Prior to the completion of the railway from Kweichow to Kwangsi in 1959, the Kweichow–Kwangsi highway was the principal freight and passenger route to Kwangtung and Kwangsi and to central and east China. Since then, railway lines have been extended to Yunnan, Szechwan, and Hunan.

**Administration and social conditions.** *Government.* Administratively, Kweichow is divided into two prefecture-level municipalities (*shih*), four prefectures (*t'i-ch'ü*), and three autonomous prefectures (*tzu-chih-chou*). These are further divided into counties (*hsien*), autonomous counties (*tzu-chih-hsien*), and county-level municipalities (*shih*).

Kweichow was in turmoil during the most violent phase (1967–69) of the Cultural Revolution. Although Revolutionary Committees were established in most major cities, local military units maintained order and stability. In May 1971 the Kweichow Provincial Party Committee was reestablished, and in 1980–81 the Revolutionary Committees were abolished and replaced with People's Governments and People's Congresses.

*Education.* Kweichow has one of the highest rates of illiteracy in China, especially among its many minority peoples. Nonetheless, there are a number of institutions of higher learning in the province, including Kweichow University, Kuei-yang Medical College, Kweichow Agricultural Institute (founded in 1939), the Kuei-yang Nationalities Institute (for training members of ethnic minority groups), and Kweichow Chinese Medical College. The province also has some 200 natural science institutes.

*Health and welfare.* Since 1949 great strides have been made in public health, although Kweichow lags behind most of the rest of China in such areas as life expectancy and the eradication of endemic diseases. Several hundred health stations, mother-and-child-care centres, and maternity centres have been established. Health-work teams have also been established, and larger numbers of medical personnel have been trained and organized.

**Cultural life.** The minority peoples in Kweichow are among the most artistic and musical in China. The Han also have a long and mixed cultural background. Various types of folk dramas with varying degrees of elaboration, some of which are combined with folk dances, are popular among different nationalities in different areas. Some of the Han folk dramas, *hua teng* ("flower lantern") in northern Kweichow and *ti-hsi* ("floor plays") in southern Kweichow, are also popular among the minority groups. Buffalo fighting is part of the festival activity over the New Year, especially among the Miao, Yao, and Chuang peoples. The Miao often sing of their revolutionary history and heroes, and both the Miao and the Tung folk songs are well known. Embroidery and paper cutting are both important forms of folk art among all minority peoples. The Puyi and Ch'i-lao are particularly known for their batik, the Miao and Puyi for their intricate, coloured cross-stitch work, and the Miao for their heavy silver ornaments.

Folk dramas and dances



(Left) Young Miao women embroidering and (right) Yi woman wearing embroidered clothing. Embroidery is one of the folk arts of minority peoples in Kweichow Province.
(Left) Zhou Haorong—Xinhua News Agency, (right) © Leon Lee

HISTORY

Although the area has been known to the Chinese since time immemorial, Kweichow came under large-scale Chinese influence only in the modern era, particularly during the Ming dynasty (1368–1644), when it was made a province. The colonization policy of the Ming and Ch'ing dynasties encouraged a large number of Chinese immigrants from Hunan, Kiangsi, and Szechwan to move into the eastern, northern, and central parts of Kweichow.

Kweichow's history of rebellion

During the Ch'ing dynasty (1644–1911/12), when the government decided to replace local chiefs by officials appointed by the central government, struggles broke out between the minorities, especially the Miao, and the Han. Rebellions and suppressions were so common that there was a saying, "a riot every 30 years and a major rebellion every 60 years." In 1726 at the Battle of Mount Lei-kung, more than 10,000 Miao were beheaded and more than 400,000 starved to death. The Pan-chiang Riot of 1797 was said to have been started by the Puyi people, and thousands of them were either burned to death or beheaded. The most important popular revolt against the central government was one led by Chang Hsiu-mei, a Miao, in 1854. He and his followers united with the Taiping revolutionaries, and the joint army with a centralized command that was organized soon controlled eastern and southern Kweichow and won numerous victories under the Miao leaders Yen Ta-wu and Pa Ta-tu. When the Miao were eventually defeated in 1871, however, countless numbers of them were massacred. The most recent revolt, known as the Ch'ien Tung (eastern Kweichow) Incident, occurred between 1941 and 1944 as a result of exploitation and suppression by the warlord Wu T'ing-chang. Bitter struggles between the Miao and Wu's armies went on until 1944. (C.-K.L./R.L.Su.)

## Szechwan

Szechwan (Ssu-ch'uan in Wade–Giles romanization, Sichuan in Pinyin), the second largest province of China, is located in the Upper Yangtze Valley in the southwest part of the country. It covers an area of 220,100 square

miles (570,000 square kilometres) and is bordered by the provinces of Kansu and Shensi to the north, Hupeh and Hunan to the east, Kweichow and Yunnan to the south, the Tibet Autonomous Region to the west, and Tsinghai to the northwest. It is the most populous province in China. The name Szechwan means "Four Streams" and refers to the four main tributaries of the Yangtze River, which flows through the province. The capital, Ch'eng-tu, is located in the centre of the province.

From economic, political, geographical, and historical points of view, the heart and nerve centre of Szechwan is in the eastern basin area, commonly known as the Szechwan, or Red, Basin. Its mild and humid climate, fertile soil, and abundant mineral and forestry resources make it one of the most prosperous and economically self-sufficient regions of China. The area has been seen by some as China in a microcosm and is often viewed as a country within a country. The Chinese call the basin Tien Fu Chih Kuo, which literally means "Heaven on Earth."

## PHYSICAL AND HUMAN GEOGRAPHY

**The land.** *Relief.* Szechwan is bordered on all sides by lofty highlands. To the north, the Tsinling Mountains extend from east to west and attain an elevation between 11,000 and 13,000 feet (3,300 and 4,000 metres) above sea level. The limestone Ta-pa Mountains rise to approximately 9,000 feet on the northeast, while the Ta-lou Mountains, a lower and less continuous range, with an average elevation of 5,000 to 7,000 feet, border the south. To the west, the Ta-hsüeh Mountains of the Tibetan borderland—rise to an average elevation of 14,500 feet; to the east, the rugged Wu Mountains, rising to about 6,500 feet, contain the spectacular Yangtze Gorges.

Szechwan Basin

In general, the relief of the eastern region of the province is in sharp contrast to that of the west. The extensive Szechwan Basin and its peripheral highlands predominate in the east; the land slopes toward the centre of the basin from all directions. This basin was a gulf of the China Sea in the later Paleozoic Era from 570,000,000 to 225,000,000 years ago; most of it is underlain by soft sandstones and shales that range in colour from red to purple.

Within the basin, the surface is extremely uneven and gives a general appearance of badland topography. Numerous low, isolated, and rolling hills are interspersed with well-defined high ridges, floodplains, valley flats, and small, local basins. The most impressive portion of the basin's surface is the Ch'eng-tu Plain—the only large continuous tract of relatively flat land in the province.

The landforms of western Szechwan include a plateau in the north and mountains in the south. The northern area is part of the edge of the Plateau of Tibet, which consists of highlands above 12,000 feet and higher mountain ranges. There is also an extensive plateau and some swampland. To the south the transverse mountain belt of eastern Tibet and western Yunnan Province rises to an average of between 9,000 and 10,000 feet. Trending from north to south is a series of parallel lofty ranges with narrow divides and canyons more than a mile deep. Mount Kung-k'a (Minya Konka) in the Ta-hsüeh range, is the highest peak in the province, rising to a height of 24,790 feet.

*Drainage.* Seen from the air, the principal drainage pattern of the eastern section of the province has the appearance of a leaf with a network of veins. The Yangtze—flowing from west to east—is conspicuous as its midrib, and the main north and south tributaries appear as its branch veins. Especially important are the Chia-ling and Min river systems in the north and the Wu river system in the southeast. The distribution of these veins is primarily concentrated in the upper, or northern, half of the leaf.

The four main tributaries of the Yangtze, to which the name Szechwan refers, are the Min, T'o, Chia-ling, and Fou rivers, which flow from north to south. Most of the major streams flow to the south, cutting steep gorges in the west or widening their valley floors in the soft sediments of the Szechwan Basin; they then empty into the Yangtze before it slices its precipitous gorge through the Wu River below Wan-hsien. Within the basin most of the rivers are navigable and are a common means of transportation.

*Soils.* There are six major soil regions—three in the east and three in the west. In the east, they include the highly fertile, purple-brown forest soils for which the Red Basin is named. This group of soils rapidly absorbs and loses water, so that it erodes easily. The other eastern soils consist of the noncalcareous alluvium and rice-paddy soils of the Ch'eng-tu Plain and other river valleys and the yellow earths of the highlands and ridges. The alluvial soils are the most important group agriculturally, as they are very fertile and are formed mainly from the rich black soils washed down from the Tibetan borderlands. The yellow earths are usually gray-brown in colour, are generally less fertile, and are agriculturally unimportant. The three major groups of soils in the west are the degenerated chernozem (dark-coloured soils containing deep, rich humus) soils of the Sung-p'an grassland, the alluvial soils of the numerous valleys, and the podzolized (leached), gray-brown soils of the mountain slopes.

The problem of soil creep

In Szechwan a form of soil erosion known as soil creep has developed. On hillsides where the surface slopes are composed of smooth sandstones, the covering soil gradually slides downward under the influence of gravity. In many places the thin surface soils have been completely removed, leaving only bare rocks. When the surface rock is composed of comparatively rougher shales, the soil is less easily moved.

*Climate.* The eastern basin area and the lower western valleys are sheltered from cold polar air masses by the surrounding mountains. The climate is therefore milder than would be expected and is similar to that of the Yangtze Delta region. There are more than 300 frost-free days in the eastern basin, and the growing season lasts nearly all year round. In the west, the sheltering effect of the mountains is evident from the contrast between the perennially snow-capped peaks and the mild weather prevailing in the valleys beneath them. During the summer, in the month of July, the mean temperature is about 84° F (29° C) at Chungking in the southeast and less than 68° F (20° C) in most parts of the west. During the winter the mean temperature in the west decreases northward from 54° F (12° C) in Hsi-ch'ang to 18° F (−8° C) in Ch'ien-ning.

The eastern rainy season begins in April and reaches its peak during July and August. Annual rainfall and precipitation measures about 40 inches (1,020 millimetres) annually. The east is noted for its frequent fogs, its many cloudy days, the relative absence of wind, and the high relative humidity. The extent to which the region is overcast is reflected in the saying, "Szechwan dogs bark when they see the Sun." Precipitation is lower in the west than in the east. The average total of about 20 inches falls mainly during the summer and early autumn, and there is heavy snowfall in the mountains during the winter.

Vegetation regions

*Plant and animal life.* There are four major vegetation regions—the pine-cypress-banyan-bamboo association of the basin area, the dense mixed association of coniferous and deciduous trees in the eastern highlands, the grasslands of the northwest, and the dense coniferous forests of the western highlands. Its great altitudinal differences, low latitudinal position, diversified topography, and high rainfall make the area what has been called a paradise for botanists. Extensive forests grow on the upper slopes, and rich rhododendrons are found at higher elevations; arid vegetation prevails on many canyon floors.

One of the outstanding features of vegetation of Szechwan Province is its division into vertically differentiated zones. Cypress, palm, pine, bamboo, tung, and citrus fruit trees grow below 2,000 feet, while between 2,000 and 5,000 feet there are evergreen forests and oaks. From 5,000 to 8,000 feet the vegetation is characterized by dense groves of mixed coniferous trees. Between 8,500 and 11,500 feet there is a subalpine zone of coniferous forest, while above 11,500 feet there are alpine zones of scrub and meadow up to the snow line, which occurs at 16,000 feet. One of the unique vegetational features is the presence of the dawn redwood, or *Metasequoia*—a tree previously believed to be extinct.

Two of the most interesting indigenous animal species are the *hsiung-mao* (lesser panda, or bear cat) and the *ling yang* (a special species of antelope). Both inhabit the

highlands of western Szechwan, and both have become endangered because of overcutting of their diet.

*Settlement patterns.* As one of the most densely populated provinces of China, Szechwan may be compared to the Yangtze Delta and the North China Plain. The population, however, is unevenly distributed, with most of the population concentrated in the eastern part of the province. The majority of the population is rural. There are comparatively few large villages and nucleated hamlets, except for the provincial and prefectural capitals. In the hilly regions, farmsteads are scattered through generally small and irregular terraced fields. In the Ch'eng-tu Plain the larger field units are commonly square or oblong in shape, and the farmsteads are surrounded by groves of banyan, cypress, mimosa, palm, or bamboo.

Most urban settlements give the appearance of being compactly built. Generally, the houses have only one story. There are no yards or sidewalks in front of the houses, which abut streets that are narrow and often are paved with limestone slabs. One of the outstanding features of urban settlement is the concentration of cities on river terraces, notably along the Yangtze River. Because water transportation is vital, large cities are always found wherever two major streams converge. Chungking, located at the confluence of the Yangtze and the Chia-ling rivers, is the largest city of the province and the most important trading, transportation, industrial, and cultural centre of southwestern China. Other such cities include Lu-chou on the Yangtze and the T'o rivers and Lo-shan on the Ta-tu and the Min. The principal disadvantage of these urban sites is that their areas are limited by their locations, so that development is hindered; the hazards of flooding are always a problem. Ch'eng-tu, the provincial capital and Szechwan's second largest city, is located in the centre of the Ch'eng-tu Plain.

The relation between cities and rivers

**The people.** Szechwan Province has one of the most diversified ranges of ethnic groups in all China, including Han (Chinese), Yi, Tibetans, Miao, T'u-chia, Hui (Chinese Muslims), and Ch'iang peoples. Most of the Han—who comprise the major part of the population—live in the basin region of the east. The Yi reside in the Liang-shan-i-tsu Autonomous Prefecture of the southwest, while the Tibetans are distributed in the plateau region of the west. The Miao live in the mountains of the southwest near Kweichow and Yunnan provinces. The Hui are concentrated in the Sung-p'an grassland of the northwest and are also scattered in a number of districts in the east. The Ch'iang are concentrated in the Mao-wen area on both banks of the Min River.

Minority peoples

The majority of the non-Han ethnic groups are fiercely independent and have maintained their traditional way of life. In most cases, they practice a mixture of agriculture, animal husbandry, and hunting. Among the Han there has been an influx of people from various neighbouring provinces, particularly from Hupeh and Shensi. This immigration was especially intensified in the early part of the 18th century, as a result of the massacre of the people of Szechwan by a local warlord. The immigrants brought with them agricultural techniques that are reflected in the heterogeneity of present cultivation patterns.

There are three major linguistic groups: the Han, who speak Southern Mandarin; the Tibeto-Burman group, including the Tibetans and the Yi; and the Hui, who speak Southern Mandarin but use Turkish or Arabic in their religious services. The Han practice a mixture of Confucianism, Buddhism, and Taoism. They do not maintain rigid boundaries in religious belief. The Tibetans follow their own form of Buddhism. Many people in the northwest profess Islām, while some hill tribes of the southwest may be classified as animistic. All the religious communities suffered increasingly severe proscriptions culminating in the Cultural Revolution, but a limited toleration of religion has since been instituted as government policy.

**The economy.** Szechwan Province has a varied economy based on agriculture, forestry, mineral deposits, and diversified branches of industry.

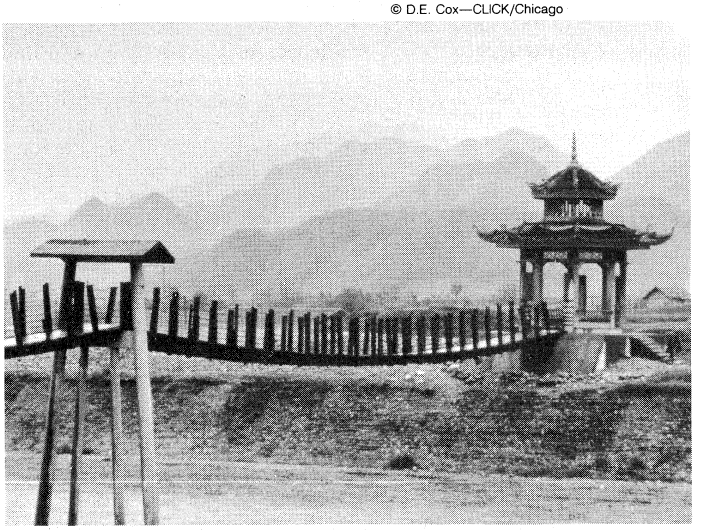
*Resources.* Mineral deposits are abundant and varied. They include both metallic and nonmetallic deposits, such as iron, copper, aluminum, platinum, nickel, cobalt, lead and zinc, salt, coal, petroleum, antimony, phosphorus, asbestos, and marble. The production of brine salt is the most extensive mining activity. Petroleum and natural gas are often located together and are widely spread throughout the province, especially in the Tzu-kung area. Natural gas deposits are also being exploited in the Pa-hsien area and elsewhere. Natural gas has been used for centuries in the production of brine salt. Most coalfields are located in the eastern and southern mountain areas; those of the Hua-ying Mountains are the richest. The most important iron deposits are along the southern and western plateau areas; those of the western sector are of high-quality titaniferous magnetite associated with vanadium. Some placer gold is panned along the Chin-sha ("Gold Sand") River. Other valuable minerals include tin and sulfur.

*Agriculture.* Most of the population of the province earn their livelihood from agriculture, and most of the provincial exports are agricultural products. Cultivation is characterized by the diversity of crops, intensive land use, the extensive practice of terracing, irrigation, the cultivation of *tsai-sheng-tao* (or "rebirth" rice), and the special methods of soil culture, fertilization, composting, and crop rotation.

Agricultural practices

The basin area of eastern Szechwan is extensively terraced and is often known as a "land of 1,000,000 steps." The terraces are of varying dimensions, commonly long narrow strips of land that frequently have rather steep slopes. They are easy to construct because the bedrock is soft and weathers easily. Even 45-degree slopes have tiny steps of terraced land. Irrigation is widely practiced in the terraced fields, and numerous methods and devices are employed. Among the most spectacular is the Tu Chiang-yen system of the Ch'eng-tu Plain, which captures the torrential flow of the Min River and guides it through an artificial multiplication of channels into numerous distributaries along the gently graded plain. Annual dredging keeps the river level constant. The system is not only the oldest but also the most successful and easily maintained irrigation system in China. It has freed the plain from the hazard of floods and droughts and ensured the agricultural prosperity of the basin. A special landscape feature of the eastern basin is the *tung-shui-t'ien* (literally "winter water-storage field") system, in which large tracts of terraced fields are left fallow during the winter season and are used for the storage of water that is needed in the paddy fields in the spring; from the air they resemble a mosaic of broken mirrors.

© D.E. Cox—CLICK/Chicago



Footbridge over a channel in the Tu Chiang-yen irrigation system on the Ch'eng-tu Plain, Szechwan Province.

Tu Chiang-yen irrigation system

Crops range from those of subtropical climates to those of the cool temperate zone. Although Szechwan is generally classified as a rice region, it is also a leading producer of such crops as corn (maize), sweet potatoes, wheat, rapeseed, kaoliang (a variety of grain sorghum), barley, soybeans, millet, and hemp and other fibre crops. Tropical

fruits—such as litchi and citrus—grow together with the apples and pears of cool temperate climates. Other principal cash crops include sugarcane, peanuts (groundnuts), cotton, tobacco, silkworm cocoons, and tea.

Szechwan leads the nation in the total number of its cattle and pigs. It is the only region in China in which both water buffalo of South China and oxen of North China are found together. Pig bristles from Szechwan have been an important item of foreign trade for years. About half of the inhabitants of the west are pastoral. Their animals include cattle, sheep, horses, donkeys, and yaks.

*Forestry.* Szechwan is second only to China's Northeast as a lumber region. Valuable forests are located in the east, on the peripheral highlands that surround the basin area, and on the numerous hills within the basin. Western Szechwan still has much of its original forest cover. The most important products from the forests are tung oil, white wax, and various kinds of herbs.

*Industry.* There has been considerable industrial development since the 1950s, and Szechwan has become the most industrialized province of southwestern China. The most important industries include iron and copper smelting, the production of machinery and electric power, coal mining, petroleum refining, and the manufacture and processing of chemicals, textiles, and food. Chungking is the principal industrial centre, although industry in the other large cities of the province is also important. Szechwan is also known for its cottage industries; it has a long history of silk production. Other products include handwoven cloth, embroidery, porcelain, carved stone, and silver and copper items.

*Transportation.* Among the problems facing the province, none is more important and more acute than that of transportation. For centuries, travel into or out of the province has been extremely difficult; the main entrances to it were the dangerous Yangtze Gorges in the east, a treacherous plank road across the mountains in the north, and the deep canyons and swift currents of the Ta-tu and Chin-sha rivers in the west. Since the 1950s great efforts have been made to improve transportation. Dangerous shoals along the Yangtze Gorges routes have been removed by blasting, a railway has been built across the northern mountains, and steel bridges have been constructed over rivers in the west.

Water routes are the most important means of transportation. Of the approximately 300 streams in the province, the Yangtze River is the most significant, traversing the entire width of the basin from the southwest to the northeast. It is the spinal cord of the river transportation system. In the west, water transport is difficult and limited except for the lower reaches of the An-ning and Ta-tu rivers.

Railways are important for the transport of bulky products. Since the 1950s railway construction has included the Ch'eng-tu–Pao-chi railroad—the first to cross the Tsinling range—which connects with the principal east–west Lunghai Railway and thus links Szechwan to both northwest and coastal China; and the Ch'eng-tu–Chungking railroad, which links the Ch'eng-tu Plain with the Yangtze River.

The thoroughly dissected terrain and the easily weathered rock structures of the province have made the construction and maintenance of highways costly and hazardous because of the constant threat of landslides, the presence of numerous steep slopes and hairpin turns, and the necessity of constructing many solid embankments. Ch'eng-tu and Chungking form the two principal highway centres.

Principal highways

Major highway routes connect with bordering provinces in the north, Hupeh in the east, Kweichow and Yunnan in the south, and Tibet in the west. Within the province, the most important highways are the Ch'eng-tu–Chungking route and the Ch'eng-tu–Wan-hsien route.

The province's first commercial air service began in 1937. Since then, commercial flying has grown steadily. Ch'eng-tu is the principal air transportation centre, and Chungking is also a major hub.

**Administration and social conditions.** *Government.* In 1955, former Sikang Province at the edge of the Plateau of Tibet was incorporated into Szechwan Province, doubling its area. The province is divided into six prefecture-level municipalities (*shih*), nine prefectures (*ti-ch'ü*), and three autonomous prefectures (*tzu-chih-chou*). The province is further divided into counties (*hsien*), autonomous counties (*tzu-chih-hsien*), county-level municipalities (*shih*), and the so-called industrial-agricultural districts (*kung-nung-ch'ü*). They are the most important administrative units because it is through them that the government exercises control.

Administrative subdivisions

The autonomous prefectures are the A-pa Tibetan Autonomous Prefecture, with its headquarters at Ma-erh-k'ang; the Kan-tzu Tibetan Autonomous Prefecture, with its capital at K'ang-ting; and the Liang-shan-i-tsu Autonomous Prefecture with its capital at Hsi-ch'ang. As a rule, the autonomous prefectures represent little more than a symbolic cultural indulgence of local minorities. The actual control of the units is exercised by the central government at Ch'eng-tu. The ethnic groups, however, enjoy their own mode of life and preserve their language and cultural traditions with a minimum of interference by the Han-controlled provincial government.

Szechwan Province was a leader in the economic reform movement that began in the late 1970s, introducing innovative policies such as the one that linked farmers' incomes to actual output. Three counties in the province became the first areas to dissolve communes, a practice that soon spread nationwide.

*Education.* Szechwan has many institutions of higher education, nine of which are "key" schools for training China's most talented students. These are Szechwan University, Ch'eng-tu University of Science and Technology, Ch'eng-tu Institute of Telecommunication Engineering, and Szechwan Medical College (all in Ch'eng-tu); Chungking University, Chungking Agricultural Engineering Institute, Southwest Agriculture Institute, and the Southwest Institute of Political Science and Law (all in Chungking); and Southwest China Chiao-t'ung University (in O-mei).

*Health and welfare.* The warm and wet climate of most of the province makes respiratory ailments a major health problem. Because of the severe pressure of the people on the land, the farmers of Szechwan must work extremely hard to eke out a living. The farmers of the Ch'eng-tu Plain are the most prosperous and have the highest standard of living. Rural life is harder in the hills surrounding the basin, and the standard of living is considerably lower in the west, where pastoral activities predominate. In the western mountains, many of the people migrate seasonally from the lowlands to the highlands in search of pasturage.

**Cultural life.** Ch'eng-tu has always played a vital role in the cultural and intellectual life of Szechwan. The city is a haven for intellectuals and scholars, and—with its heavy traffic, rich nightlife, and luxurious surroundings—is sometimes called the "Little Paris" of China.

Architecture

The unique form of architecture of the eastern basin is characterized by projecting eaves, gracefully curved roofs, and rich, elaborate roof ornaments. Because there is little wind and practically no snow in the basin, these fragile and extraordinarily beautiful structures and decorations can safely be constructed. The frequent misty rains make it necessary to project the roof eaves over the walls to protect them from the rain.

## HISTORY

Apart from the Upper Huang Ho (Yellow River) Valley provinces, Szechwan was the first area of China to be settled by the Han. The first organized Han migration took place in the 5th century BC. Szechwan was known as the Shu Pa territory during the Chou dynasty (1111–255 BC). During the Ch'in dynasty (221–206 BC) the territory was incorporated within the Ch'in Empire and began to assume considerable importance in China's national life. It was at this time that the Tu Chiang-yen irrigation system was built. In the time of the San-kuo (Three Kingdoms; AD 220–280) the Szechwan region constituted the Shu kingdom. From the end of this period until the 10th century, Szechwan was known by various names and administered through various political subdivisions. During the Sung dynasty (AD 960–1279) it was known as Szechwan Lu (Szechwan Province). Szechwan was established as a province during the Ch'ing, or Manchu, dynasty (1644–1911/12).

During the early years of the Chinese republic (1911–

30) Szechwan suffered seriously from the feudal warlord system; at one time it was divided into as many as 17 independent military units, and not until 1935 was it unified under the Nationalist government. During the Sino-Japanese War of 1937–45 there was a great influx of people and new ideas from coastal China, which resulted in extensive economic development. Many factories and trading posts were moved from the coastal area into Szechwan, and a number of industrial centres were established, especially in Chungking and at Ch'eng-tu.

Because of its geographic isolation, inaccessibility, extensive area, large population, and virtual economic self-sufficiency, Szechwan has served periodically as a bastion in its own right. The area is easily defensible, and geography has encouraged political separatism. During the war with Japan, the province was the seat of the Nationalist government from 1938 to 1945; the Japanese were never able to penetrate the area. (C.Hu./Y.-G.G.H./R.L.Su./Ed.)

## Yunnan

Yunnan (Yün-nan in Wade–Giles romanization, Yunnan in Pinyin), the fourth largest province of China, is a mountain and plateau region on China's southwestern frontier. It is bounded by the Tibet Autonomous Region on the northwest, Szechwan on the north, and the Chuang Autonomous Region of Kwangsi and Kweichow Province on the east. To the south and southeast it adjoins Laos and Vietnam, and to the west it borders Burma for 600 miles (950 kilometres). The area of Yunnan is 152,100 square miles (394,000 square kilometres). The provincial capital is K'un-ming.

The name Yunnan has been in use since the region was made a province under the Yüan (Mongol) dynasty (1206–1368). Literally meaning "Cloudy South," it denotes the location as south of the Yün-ling ("Cloudy Mountain") Range. Although richly endowed with natural resources, Yunnan remained an underdeveloped region until recent times; and for centuries the ethnic, religious, and political separatism of the province posed obstacles to the efforts of a central government to control it. Although the province remains relatively underdeveloped and isolated, its economic, political, and cultural integration into the Chinese nation essentially is complete.

### PHYSICAL AND HUMAN GEOGRAPHY

**The land.** *Relief and drainage.* Yunnan's topography is determined by a series of high mountain chains that, starting close together from the Tibetan border, branch out southeastward across the province in fanlike fashion. Running roughly northwest to southeast, these high ranges are, from west to east, the Kao-li-kung, the Nu, and the Yün-ling. Branching farther out from the Yün-ling are some secondary ranges—the Wu-liang and the Ai-lao in the central south area, the Liu-chao in the southeast, and the Wu-meng in the northeast.

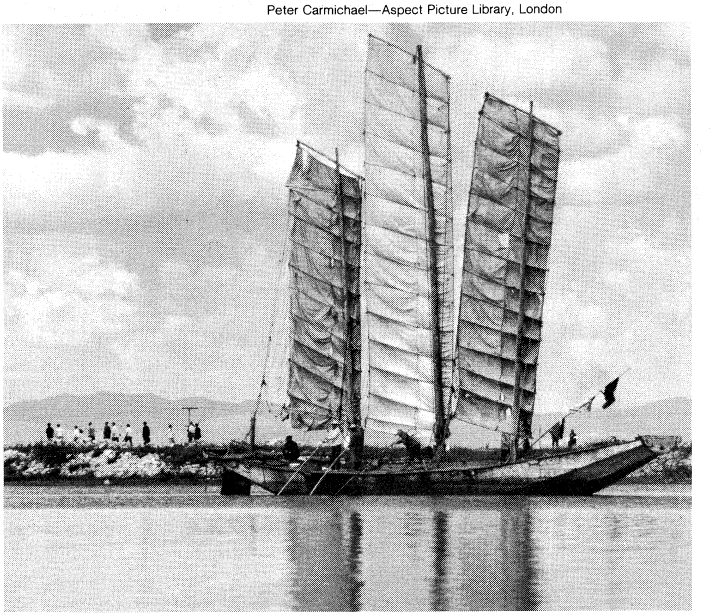
The canyon and plateau regions

The province consists of two distinct regions separated by the Ai-lao Mountains—the canyon region to the west of it and the plateau region to the east. In the canyon region the great mountains descend from an altitude of more than 18,000 feet (5,500 metres) above sea level in the north to 6,000 feet in the south. Flowing through the deep, V-shaped valleys between these mountains are the major rivers of the province: the Salween (Nu), flowing between the Kao-li-kung and Nu mountains; the Mekong (Lan-ts'ang), between the Nu and Yün-ling ranges; and the Black (Li-hsien), between the Wu-liang and Ai-lao mountains. The towering height of the mountains in the north is such that valley floors lie at heights averaging 4,000 to 5,000 feet below the mountaintops. The river currents, too swift for navigation, represent an enormous potential for hydroelectric power. In the southern part of the canyon region the mountains are much lower and the valleys more open, with many upland plains and fertile irrigated fields.

The eastern plateau region, stretching from the Ai-lao Mountains to the Kweichow–Kwangsi border, is separated from Szechwan by the Chin-sha (Yangtze) River. Streams on the western fringe of the plateau drain into the Red (Yüan) River, which flows along the eastern slope of the Ai-lao Mountains to enter the Gulf of Tonkin via Vietnam. The water of the central and eastern parts of the plateau drains into the Nan-p'an River, which is a headstream of the Hsi River of Kwangsi and Kwangtung. In the north and northeast of the plateau P'u-tu Lake and the Niu-lan and Heng rivers drain northward at right angles into the Chin-sha. The elevation of the entire plateau varies from 7,000 feet at its western end to 4,500 feet on the Kweichow border, where intermontane basins provide large stretches of level country suited for agriculture. Yunnan has more lakes than most Chinese provinces, many of them formed when grabens (large areas that dropped along fault lines) filled with water. Tien Lake in K'un-ming and Erh Lake in Ta-li are among the lakes of great beauty.

*Climate.* Despite its tropical latitude, the eastern plateau region is noted for its moderate temperatures. Because of the high elevation, summers are cool and winters are mild. The June–July temperature averages 71° F (22° C), and the December–January temperature averages 48° F (9° C). The farmers enjoy a growing season of 10 months. In the western canyon region the high mountains and deep valleys create vertical climatic zones. Sultry heat with high humidity dominates the bottom of the valleys, a temperate zone prevails at 6,000 to 11,000 feet, and freezing winds envelop the top of the mountains. Yunnan, under the influence of rain-bearing monsoon winds blowing from both the Pacific and Indian oceans, enjoys good rainfall. The bulk of rain falls during the months from May through October, the rain being heavier in the western canyon region than in the eastern plateau region. Average annual rainfall in K'un-ming is about 49 inches (1,245 millimetres), whereas in T'eng-ch'ung it is about 58 inches.

Peter Carmichael—Aspect Picture Library, London



Junk on Tien Lake at K'un-ming, Yunnan Province.

*Plant and animal life.* In the western canyon region, pine and other coniferous trees thrive up to 6,000 feet above sea level. Here are located some of China's largest timber reserves, covering nearly one-quarter of the province. From 6,000 to 11,000 feet above sea level, shrubs of the azalea and rhododendron family carpet the hills. The gorgeous colours of azaleas, camellias, roses, and fairy primroses make the mountain meadow country a gigantic botanical garden. At 11,000 to 15,000 feet above sea level, fir, bamboo, dwarf juniper, and flowering herbs grow on mountain podzolic (bleached) soil.

A wide variety of domesticated animals are kept in both the western canyon region and the eastern plateau region. In the tropical forests of the south, monkeys, bears, elephants, and porcupines are found in large numbers.

*Settlement patterns.* Approximately one-tenth of Yunnan's population is urban; the rest is rural. Following the development of industries in recent years, urban growth

has been marked by the emergence of medium-sized cities rather than giant metropolitan complexes. The only exception to this is K'un-ming, the provincial capital, which experienced significant growth due to rapid industrial development. Other important cities include Ko-chiu, the "tin capital"; Tung-ch'uan, centre of the copper industry; Hsia-kuan, the junction of highways to Tibet and Burma; and Lu-hsi, near the border with Burma.

K'un-ming, the capital

**The people.** Yunnan's population is noted for the great complexity of its ethnolinguistic groups. Out of the total population, the Han (Chinese) form the bulk of both the city and the agricultural population on the plains and valleys devoted to rice cultivation. Descendants of the conquering armies and immigrants who arrived through the centuries, they have both pushed back the non-Han peoples and intermarried with them. There is a large number of Hui (Chinese Muslims), the descendants of the Muslim immigrants sent in to help rule the province after the 13th century. The non-Han population of Yunnan remains substantial; it comprises more than 20 recognized nationalities and numerous other minority peoples, accounting for nearly one-third of Yunnan's population. In distribution, these groups are highly intermixed; not one county is inhabited by a single nationality.

The Yi are the largest minority group in the province. Once rulers of large parts of Yunnan, the Yi are a hill people with subsistence agriculture and proud warrior traditions. Linguistically, they belong to the Tibeto-Burman group. Second largest in population are the Pai in northwestern Yunnan. Long Sinicized, the Pai are rice cultivators who are among the original inhabitants of the region. Other peoples in the Tibeto-Burman linguistic family are the Hani, Lisu, and Lahu of the Yi subgroup; the Nasi (Na-hsi), who are a branch of the Hsi-fan subgroup; the Tibetans, who inhabit the far northwest corner of the province and practice Tibetan Buddhism; and the Ching-p'o, who speak the same language as the Kachin of upper Burma.

Speakers of Tibeto-Burman languages

A second major linguistic family represented in Yunnan is the Tai group. Most of the Tai peoples inhabit the semitropical lowlands, raise paddy rice, and practice Buddhism; they are ethnically related to the Shan tribes of Burma and the Thai (Siamese) of Thailand. Another important linguistic group is the Mon-Khmer, represented by the Wa, former headhunters who inhabit several counties along the border with Burma. The smaller Pu-lang and Peng-lung tribes also speak a Mon-Khmer language. The Miao and Yao peoples of southeastern Yunnan make up a separate linguistic group; they are hill dwellers whose traditional slash-and-burn method of clearing land for cultivation has been replaced by more sedentary farming practices. Descended from the aborigines of neighbouring Kweichow, the Miao until recently had no written language. Finally, a significant number of Chuang inhabit the southeastern part of Yunnan, adjacent to Kwangsi.

**The economy.** *Resources.* The province has one of the world's largest tin deposits, and the leading industry is tin mining. It is mined in the southeastern part of the province. In prewar years, China exported a major portion of the tin mined there, but now despite increased production, most goes to satisfy an increased domestic demand. Tin resources have been augmented by the discovery of additional deposits. Yunnan is also a large producer of copper, which is mined chiefly in the Hui-tse region. The copper industry in Hui-tse, which supplied most of the metal for minting coins in the Ch'ing (Manchu) dynasty (1644–1911/12), has been in the process of modernization and expansion. This has led to the creation of a special municipality at Tung-ch'uan, just south of Hui-tse. Tung-ch'uan is also the centre of lead and zinc mining.

Tin and copper production

Yunnan has moderate deposits of coal and iron, but thus far, no oil has been discovered. Other mineral products include antimony, tungsten, mercury, phosphorus, silver, placer gold, cinnabar (the ore of mercury), and manganese. Tungsten and phosphorus are mined near K'un-ming. Gypsum, sulfur, fluorite, arsenic, alum, and asbestos also exist in large quantities. Deposits of bauxite provide the basis for an aluminum industry. Marble quarried at Ta-li is eagerly sought, both as building material and as material for interior decoration. The saltpetre extracted from the rock salt mined at K'un-ming is used to make fertilizers, explosives, and food preservatives.

*Agriculture.* Red soil of various ages covers both the eastern and western regions of Yunnan. Although only about 6 percent of Yunnan's land is arable, the wide climatic variations assure the province a variety of crops. Rice is by far the basic food grain raised in Yunnan. In the upland plains, in the open valleys, and on the terraced hillsides, rice is the principal summer crop, with corn (maize) an important secondary crop. Other summer crops in the rice regions include sweet potatoes, vegetables, sugarcane, and tea. Winter crops in the rice regions include wheat, barley, beans, peas, and rapeseed. Among the hill peoples, corn, barley, and wheat are raised in summer in drier fields. Peaches, persimmons, walnuts, and chestnuts are also produced locally. In the extreme south, especially in the low-lying valleys, such produce as bananas, coconuts, and coffee is grown. Yunnan is one of China's major producers of tobacco; other industrial crops include cotton and hemp. The western canyon region holds enormous timber reserves and produces some tung oil. Livestock raised in Yunnan include water buffalo, ponies, mules, cattle, sheep, goats, and pigs. Ham from the city of Hsüan-wei is celebrated as a gourmet's delight.

Summer and winter crops

*Industry.* Manufacturing industries using traditional techniques produce paper, sugar, leather, hemp, native cloth, woolen yarn, and rugs. These traditional industries are vital to the province's economy because peasant purchasing power has increased faster than modern industry has been developed to provide items for purchase. Profiting from the government's policy of locating new industries in interior provinces with large natural resources, Yunnan has experienced great industrial growth. The K'un-ming region is a giant industrial complex, consisting of steelworks, iron- and copper-smelting facilities, and plants for manufacturing fertilizers, trucks, industrial chemicals, optical instruments, textiles, and processed foods.

Industrial growth

*Transportation.* Because of its rugged and broken terrain, Yunnan has long suffered from poor communications. Until World War II the only rail link with the outside world was the French-built railroad from K'un-ming to Hanoi and Haiphong in Vietnam. Since the mid-1950s, railroads have been built to link K'un-ming with both Kweichow and Szechwan and thus to other parts of China.

It is in the development of highways that Yunnan has made the fastest progress, opening links with neighbouring provinces and achieving a balanced network within the province. K'un-ming, Hsia-kuan, and P'u-erh (to the southwest) form the triangular axis of Yunnan's road system, from which radiate numerous highways. The most famous of these routes is the Burma Road, running from Hsia-kuan to Lashio in Burma. The vigorous road-development program has produced significant effects. Travel and trade with Kweichow, Kwangsi, and Szechwan have increased, and the close links with Tibet and Sinkiang have proved their strategic value. But most important has been the momentum for development and modernization in the remote regions inhabited by non-Han peoples. The slow transport of goods on men's backs or by pack animal is relied upon in the more isolated areas; but truck transport reaches most villages, making available large quantities of modern tools, fertilizers, and daily necessities to the farmers, while making it possible to ship farm products to near or distant markets where they can be sold to the best advantage of the producers.

The effects of road building

Most of the rivers in Yunnan are unnavigable, except for short distances or in broken stretches. Steam launches ply between towns on the shores of Erh Lake, but they cannot sail beyond there to connect with other waterways. Aviation, however, has added an important dimension to transportation in Yunnan. K'un-ming is the hub of both domestic and international services of the Civil Aviation Administration of China (CAAC), China's national airline.

**Administration and social conditions.** *Government.* Like those of the other provinces, Yunnan's administrative divisions are hierarchically organized. Immediately below the province, there are two prefecture-level municipalities

(*shih*), seven prefectures (*ti-ch'ü*), and eight autonomous prefectures (*tzu-chih-chou*), designated for various minority nationalities. At the next lower level there are municipalities (*shih*), counties (*hsien*), and autonomous counties (*tzu-chih-hsien*). The lowest political units are the villages and towns. At all levels, People's Congresses are the organs of government authority. The executive functions of government are performed by the People's Councils elected at each level by their respective People's Congresses. The Chinese Communist Party (CCP), with its own congresses and committees at every level, exercises a controlling power over this system of government.

The large number of autonomous districts and autonomous counties reflects the government's intent to end the traditional antagonism between Han and non-Han peoples. The policy seeks to preserve the language, customs, and cultural traditions of the minority peoples and to afford equal opportunity to all racial groups. Adequate representation of the nationalities on every level of government is a prerequisite. At the same time, the government is determined that the minorities should undergo the same socialist transformation as the Han majority.

*Education.* Besides the regular school system, spare-time schools of all types bring adult education to farms, factories, offices, and other places. Evening classes or off-work study sessions enable working people to go to school without leaving their jobs. The movement to upgrade the education of adults complements the campaign against illiteracy. A basic Chinese vocabulary in simplified strokes is taught to millions of illiterate people in short but intensive courses. For the minority peoples an effort is made to spread the knowledge of their own written language if one exists. Nonetheless, Yunnan's illiteracy rate is second only to that of Tibet, largely due to inadequate education among ethnic minorities.

The literacy campaign

In higher education, Yunnan has one "key school"—Yunnan University in K'un-ming. There is also a growing number of technical schools, among which the most prominent are the Yunnan College of Forestry, Yunnan Agricultural University, Yunnan Academy of Agricultural Science, K'un-ming Medical Colleges No. 1 and No. 2, Yunnan College of Traditional Chinese Medicine, and K'un-ming Institute of Technology. Other notable establishments of learning are the K'un-ming branch of the Academia Sinica (Chinese Academy of Sciences), the Feng-huang-shan Astronomical Observatory, and the Yunnan Provincial Library.

*Health and welfare.* Since 1949 Yunnan has made giant strides in sanitation and public hygiene. Such previously widespread diseases as trachoma (a contagious eye disease), smallpox, malaria, measles, snail fever, and bubonic plague have been brought under control. Iodization of water has put an end to the high incidence of goitre. In curative medicine, improvements have been slow. Clinics providing free medical care are available in all the counties, but modern hospitals are found only in the major cities. There is an expanding demand for larger medical staff, more hospital beds, and more modern equipment. The government has been promoting a massive movement to expand cooperative medical service and to collect local medicinal herbs.

Medical progress

A minimum of social welfare is available to the people. Local work-unit welfare funds provide care for the sick, the disabled, the aged, and victims of drought or flood. For industrial workers there are measures for accident prevention and insurance programs that provide for hospital treatment, sick leave, disability compensation, maternity leave, old-age benefits, and death benefits. The government has been steadily improving the housing situation and expanding recreational facilities, including hot springs, swimming pools, cinemas, and theatres.

**Cultural life.** Yunnan's cultural life is one of striking contrasts. Archaeologists have discovered sepulchral mounds containing magnificent bronzes at Chin-ning, south of K'un-ming, dating to the Han dynasty (206 BC–AD 220). At Chao-t'ung, in the northeastern part of the province, frescoes belonging to the Tung (Eastern) Chin dynasty (AD 317–420) also have been uncovered. Other historical landmarks of Chinese culture in subsequent ages abound. Yet the roots of the tribal life of the non-Han peoples remained untouched until the mid-20th century. Although the CCP abolished some minority practices, such as Yi slaveholding and Wa headhunting, the post-Mao Zedong policy for the liberalization of nationalities permitted many local customs and festivals to flourish again. In contrast to the period of the Cultural Revolution (1966–76), when minority culture and religious practices were repressed, Yunnan has come to tolerate and even celebrate its cultural diversity.

## HISTORY

In classical antiquity, Yunnan was inhabited by aboriginal tribes that were beyond the reach of Chinese civilization though they acknowledged Chinese suzerainty under the Ch'in (221–206 BC) and Han (206 BC–AD 220) dynasties. Governmental power rested with tribal chiefs, and Chinese settlers penetrated only the eastern parts of the province. Under the T'ang dynasty (AD 618–907) a Tai kingdom, known as Nanchao, flourished in the Ta-li region. First sanctioned as a bulwark against Tibetan incursions, Nanchao eventually threatened Chinese power, which declined during the period of the Wu-tai (Five Dynasties; 907–960) and the Sung dynasty (960–1279).

The Nanchao kingdom

This state of affairs came to an end during the Yüan dynasty (1206–1368). The Mongols destroyed Nanchao in 1253, and, having named the area Yunnan, they made it a province of the Yüan Empire. Marco Polo visited the region in the latter part of the 13th century. To resettle the region, which had been depopulated by warfare, the governor brought in large numbers of Hui (Chinese Muslims) from northwestern China. Thus, the Mongol conquest drew Yunnan into the orbit of Chinese affairs but failed to reduce local interracial tension between Han and non-Han minorities.

Ming dynasty rulers (1368–1644), seeking to tighten their control over the province, used military units to promote the migration of the Chinese people from the Yangtze Valley to Yunnan. The province was governed through a system of hereditary *t'u-ssu;* that is, local leaders serving as agents of the Chinese magistrates. This policy of indirect rule was continued under the Ch'ing dynasty (1644–1911/12) and the republic (1911–49), when efforts to bring the province more thoroughly under the control of the central government were undertaken, with varying degrees of success.

Regional separatism coupled with ethnic and religious differences made Yunnan a frequent scene of strife. In 1674–78, Wu San-kuei, originally sent by the Ch'ing government to crush opposition in Yunnan, used the province as a base for rebellion against the Ch'ing government. In 1855–73 Muslims, led by Tu Wen-hsiu (alias Sultan Sulaymān), who obtained arms from the British authorities in Burma, staged the Panthay Rebellion, which was crushed with great cruelty by the Chinese Imperial troops, aided by arms from the French authorities in Tonkin. In 1915 Ts'ai Ao, onetime governor of the province, launched his drive in Yunnan to defeat the monarchist movement of Yüan Shih-k'ai, the president of the republic, who attempted to make himself emperor of China. Then, spanning the decades between World Wars I and II, the warlords T'ang Chi-yao and Lung Yün ruled the province as a satrapy, keeping it beyond the control of the central government, fostering cultivation of the opium poppy, and inflicting great suffering on the people by the collection of high taxes.

During the 19th century Yunnan fell victim to British and French imperialism. Already established in Vietnam, France regarded Yunnan as its sphere of influence and built the Hanoi–K'un-ming railway at the turn of the century to exploit the resources of the province. In 1910 the British, then established in Burma, induced the *t'u-ssu* of P'ien-ma (Hpimau) to defect from the central Chinese government and occupied his territory in northwestern Yunnan. Britain also forced China to give up a tract of territory in what is now the Kachin State of Burma (1926–27), as well as the territory in the Wa states (1940).

The era of Anglo-French imperialism

The war against Japan (1937–45) brought progress and modernization to Yunnan, as the Nationalist govern-

ment developed the province into a war base against the Japanese. Factories, universities, and government agencies were transplanted there from the coastal regions, and fresh manpower, capital, and ideas poured into the province. Industries were established, and efforts were made by the government to develop the resources of the region. The Burma Road made Yunnan the corridor through which supplies flowed to Allied war bases in all parts of China, and K'un-ming became a major U.S. Air Force base. A major advance by the Japanese Army along the upper Salween River in 1944 was halted at the city of T'eng-ch'ung, indicating the vital role that Yunnan played in the nation's defense. A decade of war forced Yunnan out of its stagnation, while its strategic location made it possible to instill the ideal of national unification in place of separatism; and the process of modernization was accelerated after the establishment of the People's Republic in 1949.

(P.-c.K./R.L.Su./Ed.)

# WESTERN CHINA

## Tibet Autonomous Region

Tibet, one of the autonomous regions (*tzu-chih-ch'ü*) of China, is often called "the roof of the world." It occupies about 474,300 square miles (1,228,400 square kilometres) of the plateaus and mountains of Central Asia, including Mount Everest (Chu-mu-lang-ma Feng). It is bordered by the Chinese provinces of Tsinghai to the northeast, Szechwan to the east, and Yunnan to the southeast; Burma, India, Bhutan, and Nepal to the south; the disputed territory of Jammu and Kashmir to the west; and the Uighur Autonomous Region of Sinkiang to the northwest. Lhasa is the capital city.

The Tibetan name for the region is Bod, and the Wade–Giles transliteration from Chinese is Hsi-tsang (Xizang in Pinyin). The name Tibet is derived from the Mongolian Thubet, the Chinese Tufan, the Tai Thibet, and the Arabic Tubbat.

Before the 1950s Tibet was a unique entity that sought isolation from the rest of the world. It constituted a cultural and religious whole, marked by the Tibetan language and Tibetan Buddhism. Little effort was made to facilitate communication with other countries, and economic development was minimal. After its incorporation into China, fitful efforts at development took place in Tibet, disrupted by ethnic tension between the Han (Chinese) and Tibetans and Tibetan resistance to the imposition of Marxist values. In the 1980s a policy of conciliation appeared to have placed Han-Tibetan relations on a more balanced footing, and a new "open door" policy stimulated tourism and economic development. (T.W.D.S./V.C.F.)

### PHYSICAL AND HUMAN GEOGRAPHY

Rugged terrain

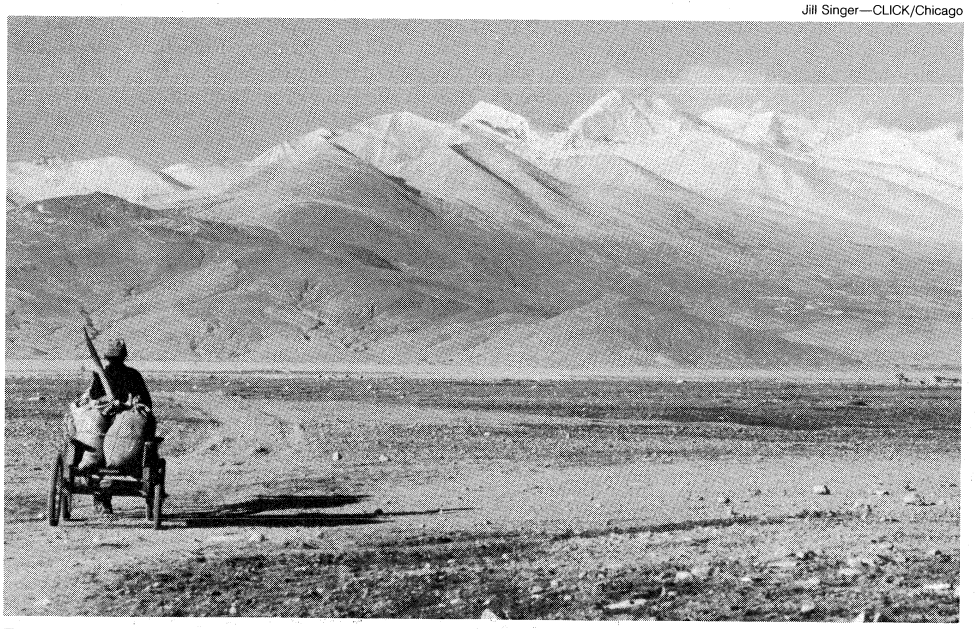
**The land.** *Relief.* Tibet is on a high plateau—the Plateau of Tibet—surrounded by mountain masses. The relatively level northern part of the plateau is called the Ch'iang-t'ang; it extends more than 800 miles (1,300 kilometres) from west to east at an average elevation of 15,000 feet (4,500 metres) above sea level. The Ch'iang-t'ang is dotted with brackish lakes, remnants of the Tethys Sea, which covered Asia Minor during the Eocene Epoch, from 54,000,000 to 38,000,000 years ago. There are, however, no river systems there. In the east the Ch'iang-t'ang begins to descend in elevation. The mountain ranges in southeastern Tibet cut across the land from north to south, creating meridional barriers to travel and communication. In central and western Tibet the ranges run from northwest to southeast, with deep or shallow valleys forming innumerable furrows.

The Ch'iang-t'ang is bordered on the north by the Kunlun Mountains, with the highest peak, Mu-tzu-t'a-ko (on the Tibet–Nepal border), reaching 25,338 feet (7,723 metres). The western and southern border of the Plateau of Tibet is formed by the Himalayan mass; the highest peak is Mount Everest, which rises to 29,028 feet (8,848 metres) on the Tibet–Nepal border. North of Ma-fa-mu-ts'o (Mapam Lake; conventional Manasarowar) and stretching eastward is the Kailas (Kang-ti-ssu) Range, with clusters of peaks, several exceeding 20,000 feet. This range is separated from the Himalayas by the Brahmaputra River, which flows across southern Tibet and cuts south through the mountains to India. (T.V.W.)

Source of rivers

*Drainage and soils.* Tibet is a prime source of water supply for Central Asia. The Indus River, known in Tibet as the Shih-ch'üan Ho (in Tibetan, Sênggê Zangbo: "Out of the Lion's Mouth"), has its source in western Tibet near Mount Kailas, a mountain sacred to Buddhists and Hindus; it then flows westward across Kashmir to Pakistan. Three other rivers also begin in the west. The Hsiang-ch'üan River (Tibetan Langqên Kanbab: "Out of the Elephant's Mouth") flows west to become the Sutlej in western India; the K'ung-ch'üeh River flows into the Kauriālā to eventually join the Ganges; and the Ma-ch'üan River (Tibetan Damqog Kanbab: "Out of the Horse's Mouth")

Jill Singer—CLICK/Chicago



The Plateau of Tibet looking toward the Himalayas and (right of centre) Mount Everest.

flows east and, after joining the Lhasa River south of Lhasa, forms the Brahmaputra.

The Salween (Nu) River has its source in central Tibet, from where it flows through eastern Tibet and Yunnan and then enters Burma. The Mekong River begins in northern Tibet as two rivers—the Ang-ch'ü and the Cha-ch'ü—which flow through eastern Tibet, merge at Ch'ang-tu (Chamudo), and enter Laos and Thailand. The T'ung-t'ien River rises in the northeast and is the source of the Yangtze River; the Huang Ho (Yellow River) rises in northeastern Tibet, where it is called the Ma-ch'ü.

Among the province's lakes, the three largest are located in central Tibet, northwest of Lhasa: T'ang-ku-la-yu-mu-ts'o (in Tibetan, Tangra Yum Tso), Na-mu-ts'o (Nam Tso), and Ch'i-lin-ts'o (Ziling Tso). South of Lhasa lie two large lakes, Yang-cho-yung (Yamdrok Tso) and P'u-mo-ts'o (Pomo Tso). In western Tibet two adjoining lakes are located near the Nepal border, Ma-fa-mu-ts'o, sacred to both Buddhists and Hindus, and La-ang-ts'o (Langak Tso).

Soils are alluvial and are often composed of sand that is blown by the wind to form a layer above gravels and shingles. Colour varies from light brown to gray, according to the humus content, which is generally poor.

*Climate.* Although Tibetans refer to their country as Gangs-ljongs or Kha-ba-can ("Land of Snows"), the climate is generally dry, and most of Tibet receives only 18 inches (460 millimetres) of rain and snow annually. The Himalayas act as a barrier to the monsoon (rain-bearing) winds from the south, and precipitation decreases from south to north. The perpetual snow line lies at about 16,000 feet in the Himalayas but rises to about 20,000 feet in the northern mountains. Humidity is low, and fog is practically nonexistent.

Temperature ranges

Temperatures in the higher altitudes are cold, but the lower valleys and the southeast are mild and pleasant. Seasonal variation is minimal, and the greatest temperature differences occur during a 24-hour period. Lhasa, which lies at an elevation of 11,830 feet, has a maximum daily temperature of 85° F (30° C) and a minimum of −2° F (−19° C). The bitterly cold temperatures of the early morning and night are aggravated by the gale winds that blow throughout most of the year. Because of the cool dry air, grain can be safely stored for 50 to 60 years, dried raw meat and butter can be preserved for more than one year, and epidemics are rare.

*Plant and animal life.* The windswept Ch'iang-t'ang is devoid of trees and larger forms of vegetation. Its arid climate supports little except grasses. The varied plant life of Tibet is found in the river valleys and in the lower, wetter regions of the south and southeast. Plant life includes willows, poplars, several types of conifers, teak, rhododendrons, oaks, birches, elms, bamboo, sugarcane, babul trees, thorn trees, tea bushes, *gro-ba* (a small white tree that grows mainly in hilly regions), *'om-bu* (a bushlike tree with red flowers that grows near water), *khres-pa* (a strong durable forest tree used to make food containers), *glang-ma* (a willow tree used for basketry), and *rtsi-shing* (the seeds of which are used for making varnish). Fruit-bearing trees and certain roots are used for food, as are the leaves of the *lca-wa, khumag,* and *sre-ral,* all of which grow in the low, wet regions. Both wild and domestic flowers flourish in Tibet. Among the wildflowers are the blue poppy, lotus, wild pansy, oleander, orchid, *tsi-tog* (a light pink flower that grows at high altitudes), *shang-dril* (a bell-shaped flower, either white, yellow, or maroon, that also grows at high altitudes), and *ogchu* (a red flower that grows in sandy regions).

Forest animals

Animal life in the forest regions includes tiger, leopard, bear, wild boar, wild goat, stone marten (a kind of cat), langur (a long-tailed monkey), lynx, jackal, wild buffalo, *pha-ra* (a small member of the jackal family), and *gsa'* (a spotted cat that is smaller than a leopard).

In the high grasslands and dry bush areas there are brown bears, wild and bighorn sheep, mountain antelope, musk deer, wild asses, wild yaks, snakes, scorpions, lizards, and *dre-tse* (members of the wolf family). Water life includes various types of fish, frog, crab, otter, and turtle.

Undisturbed by aircraft or hunters, the bird reigns supreme in the Tibetan sky. Among the many kinds to be seen are the jungle fowl, ptarmigan, spotted tinamou, mynah, hawk, and hoopoe. Other kinds include the gull, crane, sheldrake, cinnamon teal, *sing-bya* (a tiny, owllike bird), *khra* (a crow-sized, hawklike bird), *bya-long* (a bird about the size of a duck), and *skya-ka* (a black and white, crow-sized bird). The *rmos-'debs*—a small, gray bird that inhabits agricultural regions—gives a call that signals the opening of the planting season.

*Settlement patterns.* Tibet was traditionally divided into three regions, or *Chol-kha-gsum* (*Chol-kha* means "region"; *gsum* means "three"). The Dbus-Gtsang region stretches from Mnga'-ris skor-gsum at the border of Jammu and Kashmir to Sog-la skya-bo near the town of Sog. The Khams, or Mdo-stod, region consists of the territory between Sog-la skya-bo and the upper bend of the Huang Ho, now located in Tsinghai Province. The A-mdo, or Mdo-smad, region reaches from the Huang Ho to Mchod-rten dkar-po in Kansu Province, comprising most of present-day Tsinghai. Tibetans say that the best religion comes from Dbus-Gtsang, the best men from Khams, and the best horses from A-mdo.

Within the three *Chol-kha-gsum* approximately one-third of the area is uninhabitable, about one-fifth is roamed by nomads, and the rest is occupied by seminomads and agriculturalists, with a small percentage claimed by trappers in the forest belt.

The principal agricultural region is the 1,000-mile-long great valley of southern Tibet, stretching from the upper Indus Valley in the west to the valley of the upper Brahmaputra. Most of the agriculture, animal husbandry, and industry of Tibet is concentrated in this valley, which includes the main cities of Lhasa, Jih-k'a-tse, and Chiang-tzu.

**The people.** The population of the region is almost entirely Tibetan, with Han (Chinese), Hui (Chinese Muslims), Hu, Monba, and other minority nationalities. Thus, the majority of the people of Tibet have the same ethnic origin, have traditionally practiced the same religion, and speak the same language.

Languages

The Tibetan and Burmese languages are related, although they are mutually unintelligible in their modern forms. Spoken Tibetan has developed a pattern of regional dialects and subdialects, which can be mutually understood. The dialect of Lhasa is used as a lingua franca. There are two social levels of speech—*zhe-sa* (honorific) and *phal-skad* (ordinary); their use depends upon the relative social status between the speaker and the listener. Chinese has been imposed on the Tibetans since the 1960s.

Tibetan is written in a script derived from that of Indian Gupta in about AD 600. It has a syllabary of 30 consonants and five vowels; six additional symbols are used in writing Sanskrit words. The script itself has four variations—*dbu-can* (primarily for Buddhist textbooks), *dbu-med* and *'Khyug-yig* (for general use), and *'bru-tsha* (for decorative writing).

Bon is considered to be the first known religion in Tibet, although there is some argument as to the time of its establishment. It is a form of shamanism, encompassing a belief in gods, demons, and ancestral spirits who are responsive to priests, or shamans. With the rise of Buddhism, Bon adopted certain Buddhist rituals and concepts; the Buddhists also adopted certain features of Bon, so that the two religions have many points of resemblance.

Tibetan Buddhism

Although Chinese Buddhism was introduced in ancient times, the mainstream of Buddhist teachings came to Tibet from India. The first Buddhist scripture may have arrived in the 3rd century AD, but active promulgation did not begin until the 8th century. In later centuries numerous Buddhist sects were formed, including the Dge-lugs-pa sect, which emphasizes monastic discipline; in the 17th century this sect, known also as the Yellow Hats sect, gained political supremacy that lasted until 1959.

In recent times the overwhelming majority of Tibetans have traditionally been Buddhists. Before the Chinese occupation, prayer flags flew from every home and adorned the mountain slopes. Monasteries were established throughout the country, and the Dalai Lama (the spiritual head of Tibetan Buddhism) was the supreme political head of the nation. A minority, however, were adherents of Islām,

Hinduism, Bon, or Christianity. Until a moderation of policy in the 1980s the Chinese attempted to eliminate the influence of religion in Tibetan life. The Dalai Lama was forced into exile in 1959, temples were closed, religious artifacts and scriptures were destroyed, and prayer flags were temporarily taken down. (T.W.D.S./V.C.F.)

**The economy.** *Resources.* Although Tibet is rich in mineral resources, its economy has remained underdeveloped. Surveys of the Kailas and Ma-fa-mu-ts'o districts in western Tibet conducted in the 1930s and '40s discovered extensive goldfields and large deposits of borax, as well as reserves of radium, iron, titanium, lead, and arsenic. Subsequent investigative teams dispatched in the 1950s by the Academia Sinica (Chinese Academy of Sciences) reported the existence of a huge variety of minerals and ores. The most significant of these include a belt of iron-ore deposits located on the western bank of the Mekong River stretching for almost 25 miles south of Ch'ang-tu; graphite obtained from Ning-chin and coal reported to be plentiful around Ch'ang-tu; deposits of iron ore in concentrated seams of high quality and extractable depth found in the T'ang-ku-la Mountains on the Tibet–Tsinghai border; and oil-bearing formations, a reserve of oil shales, and lead, zinc, and manganese.

Minerals

The most valuable woodland is the Khams district, though extensive forest-clad mountains are also found in the Sutlej Valley in the southwest and in the Ch'u-mu-pi Valley in the far south. In the late 1950s some 30 kinds of trees, including those of economic value such as varnish trees, spruce, and fir, were discovered; and the estimated total of forest timber resources in the Khams area alone was placed at more than 3,510,000,000 cubic feet (100,000,000 cubic metres).

The swift-flowing rivers and mountain streams have enormous hydroelectric power potential, totaling about one-third of all China's potential hydroelectric resources. Especially promising are the Brahmaputra, Lhasa, and Ni-yang-ch'ü rivers. The coal deposits and forests represent possible sources of thermal power production, and there are vast opportunities for geothermal, solar, and eolian power production. (T.V.W./V.C.F./Ed.)

*Agriculture and forestry.* The staple crops are barley, wheat, and pulses; other important crops include millet, buckwheat, *rgya-bra* (a grain similar to buckwheat), beans, hemp, and mustard. Butter from the yak (large, long-haired ox) or the *mdzo-mo* (a crossbreed of the yak and the cow) is the main dairy product. The diet is supplemented by a variety of garden vegetables. Some rice is raised in the southeast. The only imported foods are tea, sugar, and rice. Most farmers keep domestic animals such as yaks, horses, mules, donkeys, and goats, and meat is obtained from cows, sheep, pigs, and chickens.

Because of the inaccessibility of Tibet's forests, forestry is developing. The forest dwellers derive their main source of income from the production of such wood products as planks, beams, printing blocks, and kitchen utensils.

Handicraft industries

*Industry.* Before the 1950s Tibet had no modern industries. There were small handicraft centres that were owned either individually or collectively and that produced scroll paintings, metal images, wooden block prints, and religious images. For these crafts, the *lag-shes-pa,* or craftsmen, had to be well versed in literature and mathematics. There were also carpet weavers, tanners, potters, gold- and silversmiths, carpenters, tailors, and incense-stick makers—all of whom learned their trade through apprenticeship. Because the government rewarded outstanding artists and craftsmen with official titles, estates, and money, the arts and crafts of Tibet were well preserved.

The initial steps toward industrial development came in 1952, when an iron- and woodwork factory was opened in Lhasa. This was followed by an automobile repair shop in 1957 and a tannery in 1958.

Under the Chinese government the small hydroelectric power station at Lhasa was repaired and reinforced with three generators. A new thermal station was installed in Jih-k'a-tse. Hydrographic stations in Lhasa and elsewhere were established to determine the hydroelectric potential of the Brahmaputra, Lhasa, and Ni-yang-ch'ü rivers. An experimental geothermal power station began generating electricity in the early 1980s, with the transmitting line terminating in Lhasa. In the 1980s emphasis was placed on agricultural-processing industries and tourism.

*Finance.* There were no banks before 1951. Small loans to be paid with interest could be obtained from local merchants, and the Tibetan government loaned public funds at interest as a means of collecting revenue. The Chinese have established branches of the People's Bank of China and have also extended agricultural and commercial credit and introduced Chinese currency. (T.W.D.S./V.C.F.)

*Transportation.* Before 1951 traveling in Tibet was done either on foot or on the backs of animals. Coracles (small boats made of wicker and hides) were used to cross the larger rivers. The Tibetan government obstructed the development of modern transportation to make access to the country difficult for outsiders. For trading, the Tibetans relied on the centuries-old caravan routes leading to Lhasa, of which the most important were from Tsinghai (via Na-ch'ü) and Szechwan (via Ch'ang-tu), India (via Kalīmpong and Ya-tung); Nepal (via Skyid-grong and Nya-lam rdzong); and Jammu and Kashmir (via Leh and Ka-erh).

Under the Communist Chinese, a network of roads was constructed, notably the Tsinghai and Szechwan highways. Additional trunk roads have been constructed that connect Tibet to Sinkiang, Yunnan, and Nepal.

Communications

The first air link between Tibet and Peking was inaugurated in 1956. The first telegraph line was strung between Kalīmpong (India) and Chiang-tzu by the British in 1904. In the 1920s another line connecting Chiang-tzu with Lhasa was erected, this being the only telegraph system in use until the Chinese took over in 1951. Postal and telecommunication stations, including mobile units, serve remote border areas and geological, hydrological, and construction teams. (T.V.W./V.C.F.)

**Administration and social conditions.** *Government.* Before the Chinese occupation, Tibet had a theocratic government of which the Dalai Lama was the supreme religious and temporal head. After 1951 the Chinese relied on military control and a gradual establishment of regional autonomy, which was granted in 1965.

Since 1965, as part of the separation of religion and civil administration, Tibet has been an autonomous region (*tzu-chih-ch'ü*) of China. The region is divided into the municipality (*shih*) of Lhasa, directly under the jurisdiction of the regional government, and seven prefectures (*ti-ch'ü*), which are subdivided into counties (*hsien*).

The army consists of regular Chinese troops under a Chinese military commander, who is stationed at Lhasa. There are military cantonments in major towns along the borders with India, Nepal, and Bhutan. Tibetans have been forcibly recruited into regular, security, and militia regiments.

*Education.* There were a few secular schools in Tibet before the Chinese established control. The monasteries were the main seats of learning, and some of the larger ones were similar in operation to theological universities. Secular facilities were established in the 1950s, including government-run primary schools, community primary schools, and secondary technical and tertiary schools including Tibet University. The state has also opened a 10-year doctoral degree program in Buddhism at the new Tibet Buddhist College.

*Health and welfare.* Under the health program of the Tibetan government, medical advice and medicine were provided free to expectant mothers. In addition to free vaccinations, sacred pendants known as *rims-srungs* were distributed annually to prevent epidemics. The construction and maintenance of proper drainage systems, wells, and canals—and security facilities to guard against pollution of water sources—were undertaken through the health program. Care of the kinless aged and handicapped persons was also undertaken. The Chinese have built modern hospitals, improved the drainage system, and placed mobile health units at key locations. (T.W.D.S./V.C.F.)

**Cultural life.** *The arts.* Tibet is most renowned for its religious scroll paintings (tankas, or *thang-ka*), metal images, and wooden block prints. There are three categories of images—representing the peaceful, moderate, and angry deities—and three schools of painting—the *Sman-*

*thang, Gong-dkar Mkhan-bris,* and *Kar-ma sgar-bris*—which are differentiated by colour tones and depicted facial expressions.

The rich and ancient culture is based on religion. The *Gar* and the *'cham* are stylistic dances performed by monks; they reenact the behaviour, attitudes, and gestures of the deities. Ancient legendary tales, historic events, classical solo songs, and musical debates are elaborately staged in the open air in the form of operas, operettas, and dramas. The folk songs and dances of local regions abound with colour, joy, and simplicity: the *bro* of the Khams region, the *sgor-gzhas* of the *dbus-gtsang* peasants, and the Kadra of the A-mdo area are spectacles that are performed in groups; on festive occasions they continue for several days. These cheerful performances tell of the people's loves and celebrate their faith in their religion, the beauty of their country, and the brave deeds of their ancestors.

Tibetan marriage practices

*Customs.* Traditional marriage ceremonies involve consultations with both a lama and an astrologer in order to predict the compatibility of a couple. The signing of a marriage contract is followed by an official ceremony at the home of the bridegroom. Appearance in a temple or before a civil authority is not required. After a couple is officially wedded, prayer flags are hoisted from the bride's side of the family upon the rooftop of the bridegroom's house to symbolize the equality of the bride in her new home. Although polygamy was practiced on a limited scale, monogamy is the predominant form of marriage.

When a death occurs, the family members make charitable contributions in the hope of ensuring a better reincarnation for the deceased. In the case of the death of an important religious figure, his corpse is preserved in a tomb. Otherwise, tradition calls for the corpse to be fed to the vultures, as a symbol of charity. The customs of burial and cremation exist but are seldom practiced.

A white scarf (*kha-btags*) is offered during greetings, visits to shrines, marriage and death ceremonies, and other occasions. The tradition was derived from the ancient custom of offering clothes to adorn the statues of deities. Gradually, it evolved into a form of greeting, and the white scarf offering, symbolizing purity, became customary. Another tradition is the hoisting of prayer flags on rooftops, tents, hilltops, and almost anywhere a Tibetan can be found. These flags signify fortune and good luck.

*Food and drink.* The staple Tibetan food is barley flour (*rtsam-pa*), which is consumed daily. Other major foods include wheat flour, yak meat, mutton, and pork. Dairy products such as butter, milk, and cheese are also popular. The people in the higher altitudes generally consume more meat than those of the lower regions, where a variety of vegetables is available. Rice is generally restricted in consumption to the well-to-do families, southern border farmers, and monks.

Two beverages—tea and barley beer (*chang*)—are particularly noteworthy. Brick tea from China and local Tibetan tea leaves are boiled in soda water. The tea is then strained and poured into a churn, and salt and butter are added before the mixture is churned. The resulting tea is light reddish white and has a thick buttery surface. *Chang,* which is mildly intoxicating, is thick and white and has a sweet and pungent taste.

*Festivals.* Festivals are both national and local in character. The many local celebrations are varied; national festivals, though fewer, are marked with a spirit of unity and lavishness.

The first day of the first month of the Tibetan calendar (February or March of the Gregorian calendar) is marked by New Year celebrations all over Tibet. Monasteries, temples, *stūpa*s (outdoor shrines), and home chapels are visited at dawn, and offerings are made before statues and relics of deities and saints. A special fried cookie known as *kha-zas* is prepared in every home. Either a real or an artificial head of a horned sheep adorns the offerings. A colourful container filled with barley flour and wheat grain and another container of *chang* are presented to all visitors, who take a pinch of the contents and make an offering to the deities by throwing it in the air.

Religious celebrations

The New Year celebrations are almost immediately followed by the Smom-lam ("prayer") festival, which begins three days after the New Year and is celebrated for 15 days. The festival marks the victory of Buddha over his six religious opponents through debates and the performance of miracles. During this festival, special prayers are offered daily. Prayers, fasting, and charitable donations mark *sa-ga zla-ba,* the celebration of the anniversary of Buddha's birth, enlightenment, and death—three events that all occurred on the 15th day of the fourth month of the Tibetan calendar.

The death of Tsong-kha-pa, founder of the Dge-lugs-pa sect, is celebrated on the 25th day of the 10th month by the burning of butter lamps on the roofs and windowsills of every house. This festival is known as *lnga-mchod.* The *dgu-gtor* festival, or festival of the banishment of evil spirits, takes place on the 29th day of the last month of the Tibetan year. At night a bowl of flour soup and a bunch of burning straws are taken into every room of every house, and the evil spirits are called out. Outside, on a distant path, the soup and straws are thrown and left to burn.

*Superstitions.* Superstition is prominent in Tibet. A traveler who encounters either a funeral procession, the source of running water, or a passerby carrying a pitcher of water is considered to have good fortune awaiting him. If a vulture or an owl perches on a rooftop, it is believed that death or misfortune will soon befall the household. If snow falls during a marriage procession, it is believed that the newlyweds will face many misfortunes or difficulties. A snowfall during a funeral, however, symbolizes an impediment to death in the family for a long period of time. (T.W.D.S.)

## HISTORY

According to legend the Tibetan people originated from the union of a monkey and a female demon. The Chinese T'ang annals (10th century) place the Tibetans' origin among the nomadic, pastoral Ch'iang tribes recorded about 200 BC as inhabiting the great steppe northwest of China. That region, where diverse racial elements met and mingled for centuries, may be accepted as the original homeland of the Tibetans, but until at least the 7th century AD they continued to mix, by conquest or alliance, with other peoples. From that heritage two strains in particular stand out—the brachycephalic, or round-headed, peoples and the dolichocephalic, or long-headed, peoples. The former, which predominate in the cultivated valleys, may have derived from the Huang Ho basin and be akin to the early Chinese and Burmese; the latter, found mainly among the nomads of the north and in the noble families of Lhasa, seem to have affinities with the Turkic peoples, whose primitive wandering grounds were farther to the north. In addition, there are Dardic and Indian strains in the west, and along the eastern Himalayan border there are connections with a complex of tribal peoples known to the Tibetans as Mon.

Impact of Tibet on Asian history

From the 7th to the 9th century the Tibetan kingdom was a power to be reckoned with in Central Asia. When that kingdom disintegrated, Tibetans figured there from the 10th to the 13th century only casually as traders and raiders. The patronage of Tibetan Buddhism by the Yüan, or Mongol, dynasty of China made it a potential spiritual focus for the disunited tribes of Mongolia. This religious significance became of practical importance only in the 18th century when the Oyrat, who professed Tibetan Buddhism, threatened the authority of the Ch'ing dynasty throughout Mongolia. In the 19th century Tibet was a buffer between Russian imperial expansion and India's frontier defense policy.

**Early history to the 9th century.** Credible history begins late in the 6th century, when three discontented vassals of one of the princes among whom Tibet was then divided conspired to support the neighbouring lord of Yar-lung, whose title was Spu-rgyal btsan-po. *Btsan-po* ("mighty") became the designation of all kings of Tibet (*rgyal* means "king"; and *spu,* the meaning of which is uncertain, may refer to a sacral quality of the princes of Yar-lung as divine manifestations). Their new master, Gnam-ri srong-brtsan, was transformed from a princeling in a small valley into the ruler of a vigorously expanding military empire.

Gnam-ri srong-brtsan (*c.* AD 570–*c.* 619) imposed his au-

Srong-brtsan's reign

thority over several Ch'iang tribes on the Chinese border and became known to the Sui dynasty (581–618) as the commander of 100,000 warriors. But it was his son, Srong-brtsan-sgam-po (*c.* 608–650), who brought Tibet forcibly to the notice of T'ai-tsung (reigned 626–649), of the T'ang dynasty. To pacify him, T'ang T'ai-tsung granted him a princess as his bride. Srong-brtsan-sgam-po is famed as the first *chos-rgyal* ("religious king") and for his all-important influence on Tibetan culture, the introduction of writing for which he borrowed a script from India, enabling the texts of the new religion to be translated. He extended his empire over Nepal, western Tibet, the T'u-yü-hun, and other tribes on China's border; and he invaded north India.

In 670, 20 years after Srong-brtsan-sgam-po's death, peace with China was broken and for two centuries Tibetan armies in Tsinghai and Sinkiang kept the frontier in a state of war. In alliance with the western Turks, the Tibetans challenged Chinese control of the trade routes through Central Asia.

The reign of Khri-srong-lde-brtsan (755–797) marked the peak of Tibetan military success, including the exaction of tribute from China and capture of its capital, Ch'ang-an, in 763. But it was as the second religious king and champion of Buddhism that Khri-srong-lde-brtsan was immortalized by posterity. In 763, when he was 21, he invited Buddhist teachers from India and China to Tibet, and *c.* 779 he established the great temple of Bsam-yas, where Tibetans were trained as monks.

Establishment of Buddhism in Tibet

Buddhism foreshadowed the end of "Spu-rgyal's Tibet." The kings did not fully appreciate that its spiritual authority endangered their own supernatural prestige or that its philosophy was irreconcilable with belief in personal survival. They patronized Buddhist foundations but retained their claims as divine manifestations.

**Disunity, 9th to 14th century.** In the 9th century, Buddhist tradition records a contested succession, but there are many inconsistencies; contemporary Chinese histories indicate that Tibetan unity and strength were destroyed by rivalry between generals commanding the frontier armies. Early in the 9th century a scion of the old royal family migrated to western Tibet and founded successor kingdoms there, and by 889 Tibet was a mere congeries of separate lordships.

Tibetan generals and chieftains on the eastern border established themselves in separate territories. The acknowledged successors of the religious kings prospered in their migration to the west and maintained contact with Indian Buddhist universities through Tibetan scholars, notably the famous translator Rin-chen bzang-po (died 1055). In central Tibet, Buddhism suffered an eclipse. A missionary journey by the renowned Indian pandit Atīśa in 1042 rekindled the faith through central Tibet, and from then onward Buddhism increasingly spread its influence over every aspect of Tibetan life.

Inspired by Atīśa and by other pandits whom they visited in India, Tibetan religious men formed small communities and expounded different aspects of doctrine. Atīśa's own teaching became the basis of the austere Bka'-gdams-pa sect. The Tibetan scholar Dkon-mchog rgyal-po established the monastery of Sa-skya (1073), and a series of lamas (Tibetan priests) founded several monasteries of what is generally called the Bka'-brgyud-pa sect.

Hermits such as Mi-la ras-pa (1040–1123) shunned material things; but the systematized sects became prosperous through the support of local lords, often kinsmen of the founding lama, and, except for the Bka'-gdams-pa, each developed its own system of keeping the hierarchical succession within a noble family. In some sects the principle of succession through reincarnation was evolved. Although lamas of different schools studied amicably together, their supporters inevitably indulged in worldly competition. This tendency was intensified by the intervention of a new Asian power, the Mongols.

Relations with the Mongols

Although it has been widely stated that the Tibetans submitted *c.* 1207 to Genghis Khan to avert an invasion, evidence indicates that the first military contact with the Mongols came in 1240, when they marched on central Tibet and attacked the monastery of Ra-sgreng and others. In 1247, Köden, younger brother of the khan Güyük, symbolically invested the Sa-skya lama with temporal authority over Tibet. Kublai Khan appointed the lama 'Phags-pa as his "Imperial preceptor" (*ti-shih*), and the politicoreligious relationship between Tibet and the Mongol Empire is stated as a personal bond between the emperor as patron and the lama as priest (*yon-mchod*).

A series of Sa-skya lamas, living at the Mongol court, thus became viceroys of Tibet on behalf of the Mongol emperors. The Mongols prescribed a reorganization of the many small estates into 13 myriarchies (administrative districts each comprising, theoretically, 10,000 families). The ideal was a single authority; but other monasteries, especially 'Bri-gung and Phag-mo-gru of the Bka'-brgyud-pa sect, whose supporters controlled several myriarchies, actively contested Sa-skya's supremacy.

The collapse of the Yüan dynasty in 1368 also brought down Sa-skya after 80 years of power. Consequently, when the native Chinese Ming dynasty evicted the Mongols, Tibet regained its independence; for more than 100 years the Phag-mo-gru-pa line governed in its own right.

The arts

A proliferation of scholars, preachers, mystics, hermits, and eccentrics, as well as monastic administrators and warriors, accompanied the subsequent revival of Buddhism. Literary activity was intense. Sanskrit works were translated with the help of visiting Indian pandits; the earliest codifiers, classifiers, biographers, and historians appeared. In an outburst of monastic building, the characteristic Tibetan style acquired greater extent, mass, and dignity. Chinese workmen were imported for decorative work. Temple walls were covered with fine frescoes; huge carved and painted wooden pillars were hung with silk and with painted banners (tankas). Chapels abounded in images of gold, gilded copper, or painted and gilded clay; some were decorated with stucco scenes in high relief; in others the remains of deceased lamas were enshrined in silver or gilded *stūpa*s. Under Nepalese influence, images were cast and ritual vessels and musical instruments made in a style blending exuberant power and sophisticated craftsmanship; woodcarvers produced beautiful shrines and book covers, and from India came palm-leaf books, ancient images, and bell-metal *stūpa*s of all sizes.

**Tibet, 14th to 19th century.** *The Dge-lugs-pa (Yellow Hat sect).* For 70 peaceful years Byang-chub rgyal-mtshan (died 1364) and his two successors ruled a domain wider than that of the Sa-skya-pa. Thereafter, although the Phag-mo-gru Gong-ma (as the ruler was called) remained nominally supreme, violent dissension erupted again. In 1435 the lay princes of Rin-spungs, ministers of Gong-ma and patrons of the increasingly influential Karma-pa sect, rebelled and by 1481 had seized control of the Phag-mo-gru court.

Already a new political factor had appeared in the Dge-lugs-pa sect. Its founder was a saintly scholar, Blo-bzang grags-pa (died 1419), known, from his birthplace near Koko Nor, as Tsong-kha-pa. After studying with leading teachers of the day, he formulated his own doctrine, emphasizing the moral and philosophical ideas of Atīśa rather than the magic and mysticism of Sa-skya—though he did not discard the latter entirely. In 1409 he founded his own monastery at Dga'-ldan, devoted to the restoration of strict monastic discipline. Tsong-kha-pa's disciplinary reform appealed to people weary of rivalry and strife between wealthy monasteries. Tsong-kha-pa probably did not imagine that his disciples would form a new sect and join in that rivalry, but, after his death, devoted and ambitious followers built around his teaching and prestige what became the Dge-lugs-pa, or Yellow Hat sect, which was gradually drawn into the political arena.

In 1578 the Dge-lugs-pa took a step destined to bring foreign interference once more into Tibetan affairs. The third Dge-lugs-pa hierarch, Bsod-nams-rgya-mtsho, was invited to visit the powerful Tümed Mongol leader Altan Khan, with whom he revived the patron-priest relationship that had existed between Kublai Khan and 'Phags-pa. From this time dates the title of Dalai ("Oceanwide") Lama, conferred by Altan and applied retrospectively to the two previous hierarchs. The holder is regarded as the embodiment of a spiritual emanation of the *bodhisattva*—Avalokiteśvara, the mythic monkey demon and progeni-

Rise of the Dalai Lamas

tor of the Tibetans. The succession is maintained by the discovery of a child, born soon after the death of a Dalai Lama, into whom the spirit of the deceased is believed to have entered. Until 1642 the Dalai Lamas were principal abbots of the Dge-lugs-pa, and in that year they acquired temporal and spiritual rule of Tibet. With Altan's help virtually all the Mongols became Dge-lugs-pa adherents, and on Bsod-nams-rgya-mtsho's death they acquired a proprietary interest in the order and some claims on Tibet itself when the fourth Dalai Lama was conveniently discovered in the Tümed royal family.

To support their protégé the Mongols sent armed bands into Tibet. Their opponents were the Red Hat Lama, head of a Karma-pa subsect, and his patron the Gtsang king. That phase of rivalry ended inconclusively with the early death of the fourth Dalai Lama and the decline of Tümed Mongol authority in Mongolia. The next came when Güüshi Khan, leader of the Khoshut tribe, which had displaced the Tümed, appeared as champion of the Dge-lugs-pa. In 1640 he invaded Tibet, defeating the Gtsang king and his Karma-pa supporters.

*The unification of Tibet.* In 1642 with exemplary devotion, Güüshi enthroned the Dalai Lama as ruler of Tibet, appointing Bsod-nams chos-'phel as minister for administrative affairs and himself taking the title of king and the role of military protector. These three forceful personalities methodically and efficiently consolidated the religious and temporal authority of the Dge-lugs-pa. Lhasa, long the spiritual heart of Tibet, now became the political capital as well. Dge-lugs-pa supremacy was imposed on all other orders, with special severity toward the Karma-pa. A reorganized district administration reduced the power of the lay nobility.

The grandeur and prestige of the regime were enhanced by reviving ceremonies attributed to the religious kings, by enlarging the nearby monasteries of 'Bras-spungs, Sera, and Dga'-Idan, and by building the superb Potala palace, completed by another great figure, Sangs-rgyas rgya-mtsho, who in 1679 succeeded as minister regent just before the death of his patron the fifth Dalai Lama. By then a soundly based and unified government had been established over a wider extent than any for eight centuries.

Relations with the Manchu

The installations of the fifth Dalai Lama at Lhasa (1642) and the Ch'ing, or Manchu, dynasty in China (1644) were almost synchronous. Good relations with Tibet were important to the Manchu because of the Dalai Lama's prestige among the Mongols, from whom a new threat was taking shape in the ambitions of the powerful Oyrat of western Mongolia.

Elsewhere Lhasa's expanding authority brought disagreements with Bhutan, which held its own against Tibetan incursions in 1646 and 1657, and with Ladākh, where a campaign ended in 1684 in Tibetan withdrawal to an accepted frontier when the Ladākhī king appealed for help to the Muslim governor of Kashmir.

*Tibet under Manchu overlordship.* The Dalai Lama's death in 1682 and the discovery of his five-year-old reincarnation in 1688 were concealed by Sangs-rgyas rgya-mtsho, who was intent on continuing the administration without disturbance. He informed the Manchu only in 1696. Emperor K'ang-hsi (reigned 1661–1722) was incensed at the deception. In 1703 he discovered an ally in Tibet and an antagonist to Sangs-rgyas rgya-mtsho when Lha-bzang Khan, fourth successor of Güüshi, sought to assert rights as king that had atrophied under his immediate predecessors.

The behaviour of the sixth Dalai Lama, Tshangs-dbyangs-rgya-mtsho, who preferred poetry and libertine amusements to religion, gave Lha-bzang his opportunity. In 1705 with the Emperor's approval, he attacked and killed Sangs-rgyas rgya-mtsho and deposed Tshangs-dbyangs-rgya-mtsho as a spurious reincarnation. The Tibetans angrily rejected him and soon recognized in east Tibet the infant reincarnation of the dead Tshangs-dbyangs-rgya-mtsho.

In 1717 the Oyrat, nominally Dge-lugs-pa supporters, took advantage of Tibetan discontents to intervene in a sudden raid, defeating and killing Lha-bzang. Fear of hostile Mongol domination of Tibet compelled the Emperor to send troops against the Oyrat. After an initial reverse, his armies drove them out in 1720 and were welcomed at Lhasa as deliverers, all the more because they brought with them the new Dalai Lama, Bskal-bzang-rgya-mtsho. For the next 200 years there was no fighting between Tibetans and Chinese; but after evicting the Oyrat the Emperor decided to safeguard Manchu interests by appointing representatives—generally known as Ambans— at Lhasa, with a small garrison in support.

Ambans

The Tibetans, interpreting this as another patron-priest relationship, accepted the situation, which, generally left them to manage their own affairs. It was only in recurring crises that Manchu participation became, briefly, energetic. Imperial troops quelled a civil war in Tibet in 1728, restored order after the political leader was assassinated in 1750, and drove out the Gurkhas, who had invaded from Nepal in 1792. As Manchu energy declined, the Tibetans became increasingly independent, though still recognizing the formal suzerainty of the emperor, behind which it sometimes suited them to shelter. At no time did the Ambans have administrative power, and after 1792, when Tibet was involved in wars with Ladākh (1842) and Nepal (1858), the Manchu were unable to help or protect them.

*Administration and culture under the Manchu.* No Dalai Lama until the 13th approached the personal authority of the "Great Fifth." The seventh incarnation was overshadowed by Pho-lha, a lay nobleman appointed ruler by the Manchu; the eighth was diffident and retiring. But after the Pho-lha family's regime, Dge-lugs-pa churchmen resumed power and held onto it through a series of monk regents for about 145 years.

Chinese contacts affected Tibetan culture less than might be expected. They helped to shape the administrative machinery, army, and mail service, which were based on existing institutions and run by Tibetans. Chinese customs influenced dress, food, and manners; china and chopsticks were widely used by the upper classes. The arts of painting, wood carving, and casting figures continued on traditional lines, with much technical skill but few signs of innovation. An important effect of Manchu supremacy was the exclusion of foreigners after 1792. That ended the hopes of Christian missionaries and the diplomatic visits from British India, which had been started in 1774. Tibet was now closed, and mutual ignorance enshrouded future exchanges with its British neighbours in India.

**Tibet in the 20th century.** In the mid-19th century the Tibetans repeatedly rebuffed overtures from the British, who at first saw Tibet as a trade route to China and later as countenancing Russian advances that might endanger India. Eventually, in 1903, after failure to get China to control its unruly vassal, a political mission was dispatched from India to secure understandings on frontier and trade relations. Tibetan resistance was overcome by force, the Dalai Lama fled to China, and the rough wooing ended in a treaty at Lhasa in 1904 between Britain and Tibet without Chinese adherence. In 1906, however, the Chinese achieved a treaty with Britain, without Tibetan participation, that recognized their suzerainty over Tibet. Success emboldened the Chinese to seek direct control of Tibet by using force against the Tibetans for the first time in 10 centuries. In 1910 the Dalai Lama again was forced to flee, this time to India.

That dying burst by the Manchu dynasty converted Tibetan indifference into enmity, and, after the Chinese Revolution in 1911–12, the Tibetans expelled all the Chinese and declared their independence of the new republic. Tibet functioned as an independent government until 1951 and defended its frontier against China in occasional fighting as late as 1931. In 1949, however, the "liberation" of Tibet was heralded, and in October 1950 the Chinese invaded eastern Tibet, overwhelming the poorly equipped Tibetan troops. An appeal by the Dalai Lama to the United Nations was denied, and support from India and Britain was not forthcoming. A Tibetan delegation summoned to China in 1951 had to sign a treaty dictated by the conquerors. It professed to guarantee Tibetan autonomy and religion but also allowed the establishment at Lhasa of Chinese civil and military headquarters.

Chinese occupation

Smoldering resentment at the strain on the country's

resources from the influx of Chinese soldiery and civilians was inflamed in 1956 by reports of savage fighting and oppression in districts east of the upper Yangtze, outside the administration of Lhasa but bound to it by race, language, and religion. Refugees from the fighting in the east carried guerrilla warfare against the Chinese into central Tibet, creating tensions that exploded in a popular rising at Lhasa in March 1959. The Dalai Lama, most of his ministers, and many followers escaped across the Himalayas, and after days of fighting and bloodshed the rising was suppressed.

The events of 1959 intensified China's disagreements with India, which had given asylum to the Dalai Lama, and in 1962 Chinese forces proved the efficiency of the new communications by invading northeast Assam.

In 1966 and 1967 the Chinese position was shaken by Red Guard excesses and internecine fighting when the Cultural Revolution reached Lhasa. Military control was restored by 1969; and in 1971 a new local government committee was announced. Between 1963 and 1971 no foreign visitor was allowed to enter Tibet, and only a few Nepalese traders remained. In the late 1970s the repressive posture of the government began to ease. A new policy of religious toleration led to the revival of many sacred local practices, a return of local religious leaders to political prominence, and the beginning of diplomatic initiatives to repair ties with the Dalai Lama. In 1984 further economic and religious freedoms were restored, and economic development was encouraged. An increase in tourism and transborder trade pointed Tibet toward the road to peaceful modernization. (H.E.R./V.C.F.)

## Tsinghai

Tsinghai (Ch'ing-hai in Wade–Giles romanization, Qinghai in Pinyin), a province of northwestern China in the Tibetan Highlands, has an average elevation of 13,000 feet (4,000 metres). It is bounded on the north and east by Kansu, on the southeast by Szechwan, on the south and west by the Tibet Autonomous Region, and on the west and north by the Uighur Autonomous Region of Sinkiang. The province, a historic home of nomadic herdsmen, is noted for its horse breeding; it has earned new prominence as a source of both petroleum and coal.

The province derives its name from a large lake, Ch'ing-hai ("Blue") Lake, which is conventionally known as Koko Nor, in the northeast. Tsinghai has an area of about 284,600 square miles (737,000 square kilometres). It is the fourth largest political unit in China in area, though it is sparsely populated. The capital is Hsi-ning, which is 120 miles west of Lan-chou, Kansu Province.

### PHYSICAL AND HUMAN GEOGRAPHY

**The land.** Most of the province consists of mountains and high plateaus. In the north are the Ch'i-lien Mountains, which form the divide between the interior and exterior drainage systems of China. Through the central part of the plateau extend the Pa-yen-k'a-la Mountains (a spur of the Kunlun Mountains), which serve as the watershed of the headwaters of both the Yangtze and Huang (Yellow) rivers. In the southern part of the plateau the Tsinghai–Tibetan boundary parallels the T'ang-ku-la Mountains. Between these high mountains are broad valleys, rolling hilly areas, and extensive flat tableland.

In the northwestern part of the plateau lies the Tsaidam Basin, an immense, low-lying area between the Pa-yen-k'a-la and the Ch'i-lien ranges; its lowest point is about 8,700 feet above sea level. There are many fertile spots in the piedmont and lakeside areas of the basin. The southeastern part is a broad swamp formed by a number of rivers flowing from the snowcapped T'ang-ku-la Mountains.

Climate

The extensiveness and the complex terrain of the region result in great variations in climate, soil, and vegetation. On the whole, the climate is continental, being influenced by the region's remoteness from the sea and by the mountain ranges in the south and east that bar maritime winds. The average annual precipitation in most places is less than four inches (100 millimetres), most of which occurs during the summer. Winter is dry, cold, and windy; summer is hot. Strong winds from the Mongolian Plateau blanket the region with sand, a serious menace to agriculture. Grass thrives on the vast plateau, however, and the region possesses some of China's best pasturelands for sheep, horses, and yaks. Antelope, wild horses, wolves, foxes, bears, and exotic birds are found there.

**The people.** Most of Tsinghai's population is Han (Chinese), and the rest of the residents are minority nationalities including Tibetans, Mongols, Hui (Chinese Muslims), Salar, and Tu. About half of the population lives in minority nationality autonomous areas.

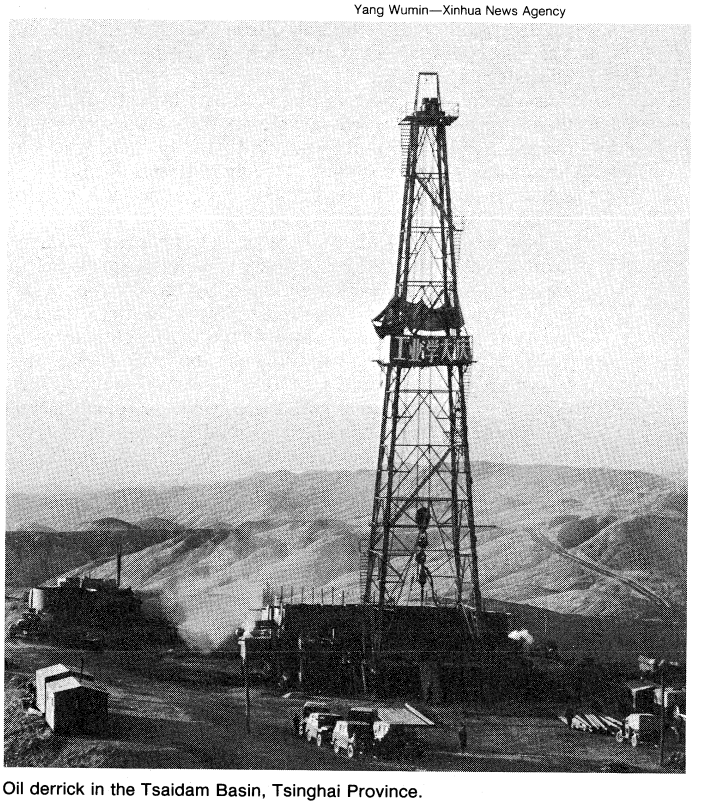
The major population centres are in eastern Tsinghai in the Hsi-ning Valley, which is the main agricultural and industrial centre. A number of cities have grown substantially with development of the province's mineral and oil and natural gas industries. Since the opening of the Tsinghai–Tibet highway, Ko-erh-mu, on the southern edge of the mineral-rich Tsaidam Basin, has become important.

**The economy.** Economically, Tsinghai is divided into two parts by the Ch'ing-hai-nan Mountains. On the eastern side is the Huang Ho drainage, consisting of large tracts of farmland crisscrossed by irrigation canals and dotted with settlements. Spring wheat, barley, and Irish potatoes are produced in much improved yields. Irrigated acreage is low, however, as is the use of chemical fertilizers. On the western side is the plateau basin, where herds of cattle, yaks, horses, and sheep, which represent the province's major source of wealth, graze on vast stretches of grassland. The output of sheep and yak wool is high and of good quality. Vast pastoral land areas have been opened up for cultivation, introducing a mixed farming-livestock economy. The Kunlun and Ch'i-lien ranges are well forested, producing spruce, birch, Chinese pine, and Chinese juniper. In the farming areas there are peach, apricot, pear, apple, and walnut orchards.

Industrial growth

Before 1949 Tsinghai's limited industrial and commercial development was based on food and animal by-products in such centres as Hsi-ning and on a few salt mines in the Tsaidam Basin. Since then, industrial growth has been rapid. Chemical plants, iron and steel factories, and electrical equipment firms have been established in Hsi-ning and other cities. Oil and natural gas reserves are located in the Tsaidam Basin, which contains most of the province's mineral reserves. Tsinghai has become China's

Yang Wumin—Xinhua News Agency



Oil derrick in the Tsaidam Basin, Tsinghai Province.

largest producer of lithium, and the province has reserves of boron, salts, potash, zinc, lead, and magnesium.

Much of the development has been made possible by the opening of new transportation links between Tsinghai and other areas of China. The crucial impetus to growth was the opening in 1959 of the Hsi-ning–Lan-chou line, connecting the province to the national rail network; the line has been extended to Ko-erh-mu and other places in the Tsaidam Basin. The Hsi-ning–Lhasa highway was widened and paved. Truck transportation is important, and main highways lead from Hsi-ning to Lan-chou, Chang-yeh in Kansu, Sinkiang, and Kan-te in Tsinghai. Several highways intersect at the southern margin of the Tsaidam Basin at Ko-erh-mu, making it a communications centre.

**Administration and social conditions.** The provincial capital is Hsi-ning. The province is subdivided into one prefecture (*ti-ch'ü*), six autonomous prefectures (*tzu-chih-chou*), and one municipality (*shih*) under provincial jurisdiction, which are further subdivided into counties (*hsien*) and autonomous counties (*tzu-chih-hsien*). The special status of the Tsaidam Basin was reflected in late 1956 by the establishment of a separate Tsaidam Administrative District, with its headquarters at Ta-ch'ai-tan, a new settlement situated on the northern edge of a salt swamp and at a major road junction. In 1964 the Tsaidam district was reincorporated into an autonomous district designated for the Mongol, Tibetan, and Kazakh minorities.

Education

The educational system of the province includes public and temple schools. For the whole province, there are comprehensive (six-year) elementary schools and junior (four-year) elementary schools for male students only. There are elementary schools for girls. Among the ethnic groups, the Hui have the highest percentage of attendance. Temple education plays an important role in the province. Among the Tibetan Buddhists, a child who becomes a lama begins his studies at the age of 10 and continues for more than 10 years. A Muslim child's studies begin at the age of six and continue for 15 years.

**Cultural life.** Urban cultural institutions such as museums, theatres, universities, and libraries are few. Life is largely rural, strongly influenced by the traditional culture of the several ethnic and nationality groups that make up the population. Among the Mongols and Tibetans, for example, one son from every family is supposed to enter a lamasery. This custom imposes a limitation on population growth. The chief monastery in Tsinghai is about 20 miles from Hsi-ning. It is a centre of Tibetan Buddhism, to which thousands of believers make pilgrimages from the Inner Mongolia Autonomous Region, Tibet, Sinkiang, and Szechwan. (C.-M.H./V.C.F.)

HISTORY

The cultivable land near Koko Nor was settled in prehistoric times and may have been the original home of the tribes who settled in Tibet. The Tsinghai region, called Amdo in Tibetan, was long considered part of Tibet. The Han referred to the people of Koko Nor and beyond as Ch'iang and sought to keep them out of the Han Empire by establishing a military outpost near the lake in AD 4. The post was soon abandoned, however, and the Chinese remained ignorant of the Tsinghai region for centuries.

During the period of political fragmentation following the decline of Han power, a branch of the Hsien-pei tribe established a state based in the Tsinghai region and extending east into present-day Kansu. Called T'u-yü-hun, this state lasted more than three centuries. A Lhasa dynasty assumed control over the region in the 7th century, reaching its peak of power in the 8th century when territory was extended far to the northeast and even reached the T'ang capital of Ch'ang-an (near modern Sian, Shensi Province) for a time.

The T'ang period

Contact was friendly between Lhasa and Ch'ang-an during the T'ang period. Slow caravans of yaks and ponies carried Buddhist monks and pilgrims across the Tsinghai desert, and traders met near Koko Nor to exchange locally bred horses for Chinese tea, which was the chief Tibetan export until the 20th century.

The Tsinghai region was later ruled by Tangut leaders who established a state called Hsi Hsia, based near Koko Nor, in 1038. Genghis Khan began his campaign against this state in 1205 and incorporated it into his expanding Mongol Empire in 1227. After the Mongol conquest of North China, Tsinghai became part of the Yüan Empire based in Peking. The founder of the Dge-lugs-pa (Yellow Hat sect) of Tibetan Buddhism, Tsong-kha-pa, was born near Koko Nor in 1357; his 16th-century successor converted Mongolia to Tibetan Buddhism and was given the title Dalai Lama by the Mongolian Khan.

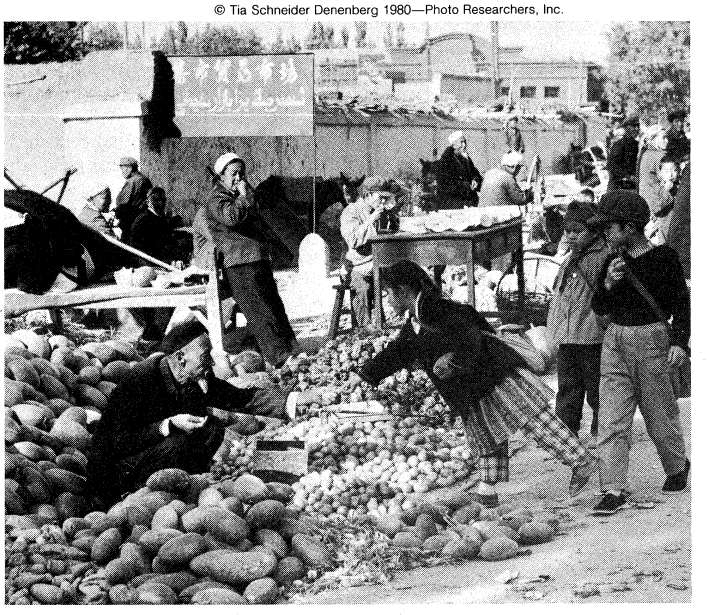
During the Ming period the Tsinghai region remained closely allied with Tibet, despite increased communication with China through trade and tribute missions. In 1642 a Mongolian dynasty was established in Tibet that lasted until 1717, when a local uprising caused the Chinese to directly interfere in the region's affairs. Tsinghai was placed under separate administration in 1724. During the Ch'ing period immigrants from the east settled in Tsinghai, and Chinese political and cultural influence in the region increased. Tsinghai was made a province of China in 1928. (V.C.F.)

## Uighur Autonomous Region of Sinkiang

The Uighur Autonomous Region of Sinkiang (Hsin-chiang in Wade–Giles romanization, Xinjiang in Pinyin) occupies the northwestern corner of China. It is bordered by Mongolia to the northeast, the Soviet Union to the northwest, Afghanistan and the disputed territory of Jammu and Kashmir to the southwest, the Tibet Autonomous Region to the southeast, and the Chinese provinces of Tsinghai and Kansu to the east. China's largest political unit, it covers about 617,800 square miles (1,600,000 square kilometres). The capital city of Wu-lu-mu-ch'i (Urumchi) is a busy commercial and industrial centre.

Known to the Chinese as Hsi-yü (Western Regions) for centuries, the area became Sinkiang ("New Borders") upon its annexation under the Ch'ing (Manchu) dynasty in the 18th century. Westerners long called it Chinese Turkistan to distinguish it from Russian Turkistan. Sinkiang is an area of lonely, rugged mountains and vast desert basins. Its indigenous population of agriculturalists and pastoralists inhabit oases strung out along the mountain foothills or wander the arid plains in search of pasturage. Since the establishment of firm Chinese control in 1949, serious efforts have been made to integrate the regional economy into that of the nation. Despite the great increase in the Han (Chinese) population, the ethnic groups are officially encouraged to develop their own cultures.

© Tia Schneider Denenberg 1980—Photo Researchers, Inc.



Uighur merchant selling melons in T'u-lu-p'an, which is situated on the north side of the Turfan Depression in the Uighur Autonomous Region of Sinkiang. The information on the sign in the background is given in both Chinese and Uighur, the latter using Arabic script.

PHYSICAL AND HUMAN GEOGRAPHY

**The land.** *Relief.* Sinkiang can be divided into five physiographic regions: the Northern Highlands, the Dzungarian Basin, the Tien Shan ("Celestial Mountains"), the Tarim Basin, and the Kunlun Mountains. These regions run roughly from east to west, the high mountains alternating with large, lower basins.

Physiographic regions

In the north the Northern Highlands extend in a semicircle along the Mongolian border. The major range in this area is the Altai Mountains, with average heights of approximately 4,500 feet (1,400 metres) above sea level. The slopes of the Altai Mountains on the Chinese (western) side are relatively gentle, with numerous rolling and dome-shaped hills.

The triangular-shaped Dzungarian Basin, or Dzungaria, with an area of some 270,000 square miles (700,000 square kilometres), is bordered by the Altai Mountains on the northeast, the Tien Shan on the south, and the A-la-t'ao Mountains on the northwest. The basin is open on both the east and west. It contains a ring of oases at the foot of the enclosing mountains and a steppe and desert belt in the centre of the depression.

The Tien Shan occupies nearly one-fourth of the area of Sinkiang. The mountains stretch into the region from the Soviet Union and run eastward from the border for about 1,000 miles (1,600 kilometres). They are highest in the west and taper off slightly to the east. The highest peaks are Mount Han-t'eng-ko-li, which rises to an elevation of 22,949 feet (6,995 metres), and Mount Sheng-li (Russian Pik Pobedy), which attains 24,406 feet (7,439 metres). They are found in a cluster of mountains, from which ridges extend along the boundary between China and the Soviet Union. The Tien Shan is perpetually covered by snow, and numerous long glaciers descend its slopes from extensive snowfields.

The Tarim Basin is surrounded by the Tien Shan to the north, the Pamir Mountains to the west, and the Kunlun Mountains to the south. It occupies about 55 percent of Sinkiang, extending 850 miles from west to east and about 350 miles from north to south. The basin consists of a central desert, alluvial fans at the foot of the mountains, and isolated oases. The desert—the Takla Makan—covers an area of more than 105,000 square miles and is absolutely barren. The core of the basin has an elevation ranging from about 4,000 feet above sea level in the west to about 2,500 feet in the east. The Turfan Depression at the eastern end of the Takla Makan, however, is 505 feet below sea level.

The Kunlun Mountains form the northern rampart of the Plateau of Tibet. With elevations up to 24,000 feet, the central part of the range forms an almost impenetrable barrier to movement from north to south. There are passes on the west and east such as the Karakoram in Jammu and Kashmir and the K'u-erh-kan in Sinkiang. In the east the A-erh-chin Mountains turn northeast and eventually merge with the Tsou-lang-nan Mountains in southern Kansu Province, China.

*Drainage.* The drainage pattern of Sinkiang is unique to China. The only stream whose waters reach the sea is the O-erh-ch'i-ssu, which rises in north central Sinkiang, crosses into the Soviet Union, and joins the Ob, which then empties into the Arctic Ocean. Other streams in Sinkiang issue from the mountains and disappear into inland deserts or salt lakes. The principal river of the region, the Tarim, is fed by largely intermittent streams that rise in the Kunlun Mountains and the Tien Shan. It flows eastward across the Tarim Basin toward the salt-encrusted lake bed of Lop Nor.

*Climate.* Remote from the ocean and enclosed by high mountains, Sinkiang is cut off from marine climatic influences. It therefore has a continental dry climate. The Tien Shan separates the dry south from the slightly less arid north, so the northern slopes of the Tien Shan are more humid than those on the south.

Rainfall is not only scanty, but it fluctuates widely from year to year. Average January temperatures in the Tarim Basin are about 20° F (−7° C), compared with 5° F (−15° C) in many parts of the Dzungarian Basin. In the summer, average temperatures north of the Tien Shan are lower than they are south of the mountains. In the Dzungarian Basin, July averages vary from 70° F (21° C) in the north to 75° F (24° C) in the south. In the Tarim Basin, July temperatures average about 80° F (27° C). The hottest part of Sinkiang is the Turfan Depression, where a maximum of 118° F (48° C) and a July mean of 90° F (32° C) have been recorded.

Temperatures

*Plant and animal life.* Because of the great expanses of desert, the plant life of much of Sinkiang is monotonous. There are pine forests in the Tien Shan and tugrak woods in many places on the edge of the Takla Makan Desert. Apart from these trees, the most common are varieties of poplar and willow. In the Tien Shan and other mountains there is a great assortment of wild plants and flowers, many of which have never been classified.

Animal life is of greater interest, and big-game hunting is an attraction of the Tien Shan. The mountains are inhabited by antelopes, ibex (wild goats), wapiti (elks), various wild sheep, leopards, wolves, bears, lynx, and marmots. There are wild horses in the north, wild camels near Lop Nor, and wild yaks (large, long-haired oxen) and wild asses on the Tibetan frontier. Birdlife is extensive, especially in the Lop Nor district. The few varieties of fish are mostly of the carp family. Snakes are not numerous and appear to be harmless; scorpions and centipedes, however, abound. During the summer, horseflies, mosquitoes, flies, and midges are thick in the woods. A great variety of butterflies are seen in the mountains.

*Settlement patterns.* There are many differences in rural settlement patterns in the north and the south. Oasis agriculture in the Tarim Basin occupies a large part of the population, and only a small percentage are engaged in animal husbandry. North of the Tien Shan the grasslands support many of the inhabitants, who are pastoralists.

There are five major cities in the province. Wu-lu-mu-ch'i is the regional capital. Once an agricultural centre for the Dzungarian Basin, it has undergone considerable industrial development. K'o-la-ma-i (Karamai), also in the Dzungarian Basin, was developed in the late 1950s as a centre of the petroleum industry. Shih-ho-tzu, near the southern edge of the Dzungarian Basin, is a significant agricultural-processing centre. I-ning (Kuldja), located in the Upper I-li Valley near the Soviet border, is an administrative town with a growing food-processing industry. Kashgar, the largest city of the Tarim Basin, is an ancient centre for the manufacture of handicrafts such as textiles, rugs, and tanned leather.

Cities

**The people.** Sinkiang is inhabited by more than 40 different ethnic groups, of which the largest are the Uighur and the Han (Chinese). Other groups include the Mongolians and Khalkha, Hui (Chinese Muslims), Kazakh, Uzbek, Tungusic-speaking Manchu and Sibos, Tadzhik, Tatars, Russians, and Tahurs.

The Han migration altered the pattern of population distribution and ethnic composition of Sinkiang. In 1953 about three-fourths of the population lived south of the mountains in the Tarim Basin. The Han influx was directed mainly to the Dzungarian Basin because of its resource potential. The Kazakh, the third largest minority group in the region, are nomadic herdsmen in the steppes of the Dzungarian Basin; they are especially concentrated in the Upper I-li Valley.

There are two major language groups besides Chinese in the region. The Mongolians speak languages of the Mongolian branch of the Altaic group, and the Uighur, Kazakh, and Uzbek speak the Turkic branch of the Altaic group. The Tadzhik, however, belong to the Iranian branch of the Indo-European language group. Mongolian, Uighur, and Kazakh are written languages in everyday use; Mongolian has its own script, while Uighur and Kazakh are written in the Arabic script.

Language groups

The largest Muslim groups in China are the Uighur and the Hui. The Kazakh and Tadzhik also follow Islām, while the Mongolians are adherents of Buddhism.

**The economy.** Because of the dry climate, most of the cultivated land in Sinkiang depends entirely on irrigation. The various nationalities in the region have had rich experience in water conservancy techniques, of which the wells of the *qanāt* system in the Turfan and Ha-mi de-

pressions are a fine example. Since the 1950s, these have been greatly supplemented with canals and reservoirs, and the amount of arable land has almost tripled.

Principal crops

Sinkiang is self-sufficient in food grains. About half of the total crop area produces winter and spring wheat. Corn (maize), another important crop, is grown more in the south than in the north. Rice, kaoliang (a variety of grain sorghum), and millet are also produced in large quantities. Significant crops of long-staple cotton are produced in the Turfan Depression and the greater Tarim Basin, and cotton has become an important cash crop. Sinkiang is one of China's main fruit-producing regions; its sweet Ha-mi melons, seedless Turfan grapes, and I-li apples are well known. Sugar beets support a small sugar-refining industry. Animal husbandry has been given renewed attention, particularly north of the Tien Shan.

Mineral resources include deposits of lead, zinc, and copper, as well as molybdenum and tungsten (used in strengthening steel), although none of these are of industrial significance. Gold is produced from placer and lode deposits on the southern slopes of the Altai Mountains. Sinkiang's only product of national significance is oil.

Oil reserves

Since the first well was developed at K'o-la-ma-i in 1955, nearly 20 fields have been developed. A major new field was discovered in the area in 1983, after which exploration for oil was begun in the Tarim Basin.

Sinkiang's heavy industry includes an iron and steel works and a cement factory at Wu-lu-mu-ch'i and a farm-tool plant at Kashgar. Industries processing agricultural and animal products have been established near the sources of raw materials and include several textile mills and a beet sugar mill.

A system of roads encircles the Tarim Basin along the foothills of the surrounding mountain ranges, and roads run along the northern foothills of the Tien Shan in the Dzungarian Basin. The two basins are connected by a road that crosses the Tien Shan near Wu-lu-mu-ch'i. There are roads leading to the Soviet Union in the north through passes in the Dzungarian Basin and in the south through the pass at Kashgar, which was the historic gateway of the silk trade between Asia and Europe. The region is also connected by road to the Chinese provinces of Kansu and Tsinghai in the southeast.

A railway crosses Sinkiang from Kansu Province through Ha-mi, Wu-lu-mu-ch'i, and the Dzungarian Gate (a pass through the Pamir Mountains), connecting with the Soviet railway system in the Kazakh S.S.R. The northern and southern sectors of the province have also been linked by a railway constructed across the Tien Shan.

**Administration and social conditions.** *Government.* The administrative structure of Sinkiang reflects the policies of recognition of ethnic minorities and self-administration, in which local leaders are appointed to governmental positions. The Uighur Autonomous Region of Sinkiang is divided on the subprovincial level into three types of administrative units. There are three municipalities (*shih*) under direct regional administration, five autonomous prefectures (*tzu-chih-chou*), and seven prefectures (*ti-ch'ü*). The region is further subdivided into counties (*hsien*) and autonomous counties (*tzu-chih-hsien*).

*Education.* Before World War II the educational system was minimal. Since 1949, educational facilities have been broadened and the literacy rate is better than the national average.

Institutions of higher learning

Institutions of higher learning, concentrated in Wu-lu-mu-ch'i, include the Sinkiang University; the Sinkiang "August First" Institute of Agriculture, which offers a course on water conservation; the Sinkiang Institute of Minorities, which offers courses in art, Chinese language and literature, history, and science; the Sinkiang Medical College; and the Sinkiang Institute of Languages, which offers instruction in several Western languages and literatures. Standard education is supplemented by instruction broadcast over radio and television. The provincial library and museum are also in Wu-lu-mu-ch'i.

**Cultural life.** The indigenous peoples of Sinkiang exhibit their own cultures. The dominant Uighur are sedentary farmers whose social organization is centred upon the village. Many of the important Uighur cultural forms are rooted in Islām. Spoken Uighur predominates despite the popularization of Chinese. Islām itself has revived since the antireligious onslaught of the Cultural Revolution during the 1960s and '70s, and there are now numerous mosques and a new training academy for clergy. The Uighur performing arts tradition emphasizes ancient songs and dances accompanied by traditional instrumental groups. Professional troupes, first organized in the 1950s, are dominated by Uighur balladeers and dancers, although administrative duties are often performed by Han troupe members.

The Kazakh are pastoralists related to the people of the Kazakh S.S.R. They migrate seasonally in search of pasturage and live in dome-shaped, portable tents known as yurts. Livestock includes sheep, goats, and some cattle; horses are kept for prestige. The basic social unit is the extended family; political organization extends through a hierarchy of chiefs. Although there is a concept of national origin, the chiefs are seldom united.

Like the Kazakh, the Mongolians are pastoralists who live in yurts, but their society is more firmly organized. The basic social unit is the nuclear family. There is an established political hierarchy of groups, the smallest of which is a group of several households known as a *bag*. The average person, or free nomad (*arat*), owes allegiance to nobles (*taiji*) and princes (*noyan* or *wang*). National power, however, is fragmented. (C.-M.H./V.C.F.)

## HISTORY

Far to the northwest of the heartland of Chinese civilization, the Sinkiang region was thinly populated by herdsmen and oasis farmers organized into small kingdoms and tribal alliances. Southern Sinkiang came under the loose control of the Han dynasty (206 BC–AD 220), when an extension of the Great Wall was built 300 miles west of the present Kansu–Sinkiang border. The Han capital of Ch'ang-an, near modern Sian in Shensi Province, came into contact with the Roman Empire over a trade route that passed through a series of oasis settlements south of the Tien Shan. Known as the Silk Road, this route carried Chinese silk to the Roman world in exchange for precious metals, glassware, and woolen cloth.

With the decline of Han power in the 3rd century AD, the area passed under the control of local Uighur leaders. The resurgence of Imperial power during the T'ang period (618–907) increased Chinese influence in the region, though many elements of western Asian culture were transmitted along the trade routes. The subsequent decrease of T'ang power resulted in an increase in Arab influence, and Islām gained many converts. Turkic language came to be spoken in the oases, while Mongolian remained the language of the steppes.

Growth of Islām

Sinkiang was again incorporated into the Chinese empire when it was conquered by the Mongol leader Genghis Khan in the 13th century. The Ch'ing, or Manchu, dynasty (1644–1911/12) successfully asserted control over the Sinkiang region, defeating the resistance of stubborn tribes in the north and sending loyal Muslims from Kansu to settle in the oases of northern Sinkiang in the 17th and 18th centuries. In 1884 the Ch'ing government created a new Sinkiang province.

After the revolution of 1911–12 Yang Tseng-hsin, a Han commander of native Turkic troops, seized control of Sinkiang and was later appointed governor by the Peking government. He maintained control until his assassination in 1928, which was followed by a series of rulers and shifting allegiances. After the Communist victory in 1949, the central government implemented moderate policies toward the local minorities, and Sinkiang was established as an autonomous region in 1955. Radical policies established elsewhere in China during the Great Leap Forward (1958–60) and the Cultural Revolution (1966–76) were also implemented in Sinkiang, however. This resulted in a mass exodus of Kazakh people into the Soviet Union in 1962, massive political instability, and heightened ethnic tensions.

After the Cultural Revolution, political and economic policies were moderated, leading to widespread improvement in the livelihood of farmers and pastoralists and to relative stability and economic growth. (V.C.F.)

# NORTHWEST CHINA

## Hui Autonomous Region of Ningsia

The Hui Autonomous Region of Ningsia (Ning-hsia in Wade–Giles romanization, Ningxia in Pinyin), in north central China, is bounded on the east in part by Shensi; on the east, south, and west by Kansu; and on the north by the Inner Mongolia Autonomous Region. Most of the region is desert, but the vast plain of the Huang Ho (Yellow River) in the north has been irrigated for agriculture for centuries. The total area of the autonomous region is about 23,000 square miles (60,000 square kilometres). Its capital is Yin-ch'uan.

Ningsia is nearly coextensive with the ancient kingdom of the Tangut people, known in China as the Hsi Hsia; after its conquest by Genghis Khan, it was named Ningsia ("Peaceful Hsia").

### PHYSICAL AND HUMAN GEOGRAPHY

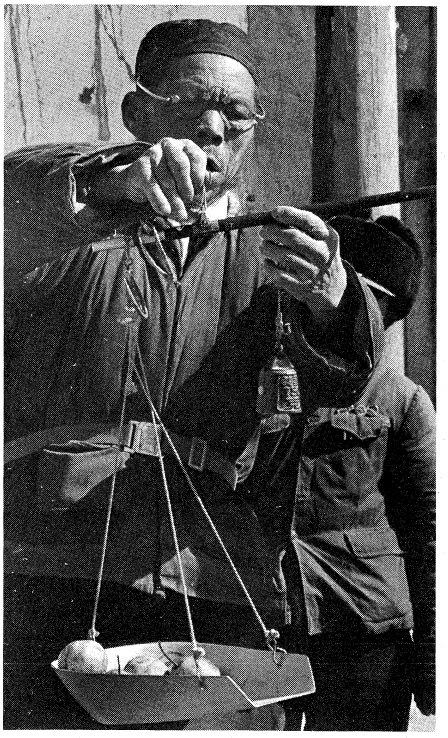
**The land.** Physiographically, the Ningsia region can be divided into two parts. Southern Ningsia is part of the Loess Plateau, with the Liu-p'an Mountains as the main ridge. The region is covered with a thick layer of loess (wind-deposited soil)—which in some places is more than 300 feet deep—and the topography is generally fairly flat. Northern Ningsia is made up for the most part of the Ningsia plain of the Huang Ho. The river enters Ningsia from the Tsinghai plateau in Kansu and flows east, then north into Inner Mongolia. West of the plain are the Holan Mountains. These mountains serve as a shelter against the sandstorms from the T'eng-ko-li Desert, which lies to the west of the mountains.

On an elevation of 3,600–3,900 feet (1,100–1,200 metres) above sea level, the Ningsia plain slopes gradually from south to north. The plain is an arid area, but the Huang Ho provides irrigation. Many canals have been built over the centuries. The network of willow-lined canals and paddy fields gives the landscape a look resembling that of southern China.

Climate

The climate of Ningsia is continental. Temperatures range from an annual average maximum of 80° F (27° C) to an annual average minimum of 7° F (−14° C). Yearly precipitation on the Ningsia plain is only about eight inches (200 millimetres).

© Leon Lee



Hui (Chinese Muslim) tradesman weighing pears in a market in Yin-ch'uan, Hui Autonomous Region of Ningsia.

**The people.** The ethnic composition of Ningsia includes the Han (Chinese), who constitute the vast majority of the population; Hui (Chinese Muslims), the largest minority group; and Tibetans, Mongolians, and Manchu. The people speak Mandarin Chinese, Tibetan, and Mongolian, and the predominant religions are Islām and Buddhism. Islām is well developed and has the most believers, including the Hui.

The region is predominantly rural, with most of the population engaged in pasturing and farming the land. It is one of China's more sparsely settled areas. In the widely scattered cities, residents have traditionally been devoted to handicrafts. Since 1949, however, more workers have begun to be employed in mining and manufacturing.

**The economy.** The Ningsia plain produces abundant wheat and good-quality rice. An intricate system of ancient and new irrigation canals has improved agricultural yields in the region, although the use of chemical fertilizers is well below the national average. Cash crops include sugar beets. In the mixed agricultural and pastoral areas, a good breed of sheep, a domesticated form of the argali of eastern Mongolia, is raised. Its wool is soft, white, and lustrous. The Manchu especially have long been known for breeding and raising pigs. Melons and apricots are also grown in quantity.

Mineral resources of Ningsia are limited to coking coal reserves in the P'ing-lo and Shih-tsui-shan areas, near the Inner Mongolian border. Coal was mined on a small scale in the past but has been expanded since the construction of the Pao-t'ou–Lan-chou railroad in 1958. There also are reserves of gypsum, glass, and limestone.

Well known in ancient times as a border city on the western frontier of China, the capital, Yin-ch'uan, in the centre of the Ningsia plain, was formerly a trading centre for farm and animal products. Medium-sized and small factories, including a farm-tool plant and a woolen-textile mill, have since been built there. The Huang Ho, to the east, provides irrigation and facilities for water transportation.

Industry

Industry has grown steadily. Natural resources and agricultural products such as wool and sugar form the foundation of many enterprises. The region produces consumer goods such as paper, foodstuffs, and wool and cotton fabrics. An extension of the main Peking–Pao-t'ou railway, completed in 1958, links Yin-ch'uan to the two new industrial bases of Pao-t'ou (in Inner Mongolia, to the north) and Lan-chou (in Kansu, to the south). A highway bridge built across the Huang Ho in the 1970s near Yin-ch'uan has further stimulated economic development.

**Administration and social conditions.** The region has two prefecture-level municipalities (*shih*), Yin-ch'uan and Shih-tsui-shan. There are two prefectures (*ti-ch'ü*), Yin-nan and Ku-yüan. The autonomous region is further subdivided into counties (*hsien*).

Advances in education

Ningsia was formerly a backward area in education. In 1935 there were only two high schools, two normal schools, and about 200 elementary schools, attended by very few of the children. Since the Communist government was organized in 1949, there has been much improvement. Illiteracy has been reduced to less than 30 percent. Health care has also been transformed with the building of clinics and hospitals.

**Cultural life.** Traditional Hui cultural life was intimately interrelated with Islām. The Hui woman traditionally kept house; her role was domestic, and she could not undertake outside work. When they went out, Hui women traditionally wore the veil to conceal their faces, and they were forbidden to talk to males. The traditional culture has undergone changes, however. Under the Communist regime, for example, Hui women have had to do farm work in the communes and production work in the factories.

HISTORY

The region south of the Huang Ho was incorporated into the Ch'in Empire in the 3rd century BC, at which time walls were built throughout the area. Irrigation canals on the Ningsia plains of the Huang Ho dating from the Ch'in (221–206 BC), Han (206 BC–AD 25), and T'ang (AD 618–907) dynasties provide further evidence that the area has long been inhabited. In the 11th century the area became part of the kingdom of the Tangut people, Hsi Hsia, in western China. Yin-ch'uan was captured by Genghis Khan early in the 13th century and remained tributary to China.

Islāmization of the Ningsia region

As Mongol power declined and Turkish-speaking Muslims migrated from oasis settlements to the west, Ningsia came increasingly under Islāmic influence. The descendants of Muslim settlers maintained their separateness from Chinese society. In the mid-19th century Ningsia became embroiled in the general Muslim revolt in the northwest, and tension between Han and Hui continued well into the 20th century. After 1911 the region came under the control of Muslim warlords, and Ningsia, as part of the "Muslim" belt, became part of the political base of the Ma clan of Ho-chou. Wooed by the Nationalist Party (Kuomintang)—to which they declared nominal allegiance—the Japanese, and the Russians, the region remained an arena of conflict throughout the period between World Wars I and II.

In 1914 the Ningsia area became a part of the province of Kansu, and in 1928 it was constituted as the province of Ningsia. During the Sino-Japanese War (1937–45) parts of Ningsia were incorporated into the Shen-Kan-Ning border region, where Communist authorities appealed for minority support by proclaiming their cultural and political rights. Although some Hui leaders joined the Communists and rose to positions of influence in the region, most Ningsia Hui supported the Ma clan. In the end, a Communist victory in Ningsia was won by the People's Liberation Army in battle with the armies of the Ma clan.

From 1949 to 1954 the province was subject to the authority of the Northwest Military Administrative Committee. Ningsia was then made directly subordinate to the central government as part of Kansu. At the same time, autonomous Hui regions were established on the east and west bank sections of the Ningsia irrigated plain and in the foothills of the Liu-p'an Mountains. In 1958 these areas were combined to form the Hui Autonomous Region of Ningsia. From 1969 to 1979 Ningsia reacquired the T'eng-ko-li Sha-mo region from Inner Mongolia.

(C.-M.H./V.C.F.)

## Inner Mongolia Autonomous Region

Inner Mongolia (Nei-meng-ku in Wade–Giles romanization, Neimenggu in Pinyin), is one of the five autonomous regions (*tzu-chih-ch'ü*) of China. It is a territory 800 miles (1,300 kilometres) long in North China, with an area of 456,800 square miles (1,183,000 square kilometres). It is bordered to the north by the Mongolian People's Republic; to the east by the Chinese provinces of Kirin and Liaoning; to the south by the Chinese provinces of Hopeh, Shansi, and Shensi and the Hui Autonomous Region of Ningsia; and to the west by the Chinese province of Kansu. Since 1952 its capital has been Hu-ho-hao-t'e (Hohhot).

PHYSICAL AND HUMAN GEOGRAPHY

**The land.** *Relief.* Inner Mongolia is essentially an inland plateau with a flat surface lying at an elevation of about 3,000 feet (900 metres) above sea level and fringed by mountains and valleys. Its southern boundary is formed by a series of high ridges with an average height of between 4,500 and 6,000 feet. To the northwest, the land falls away toward the centre of the Gobi (Desert). An arid zone with low summer rainfall, strong evaporation, almost perpetual sunshine, and constant northwesterly winds, this region provides a natural grassland for pastoral animals. The Huang Ho (Yellow River) provides the major irrigation waterway in the southwestern part of Inner Mongolia. In the centre and the north, rainfall and snow are absorbed by the desert.

*Soils.* Soils are of two types, either a chestnut-brown soil, in which cereals can be raised by dry farming once every two or three years after sufficient moisture has accumulated in the soil, or a rich, black soil in the southern extremity of the Mongolian Plateau. The prairie on both sides of the great bend of the Huang Ho is known as a "granary of the frontier."

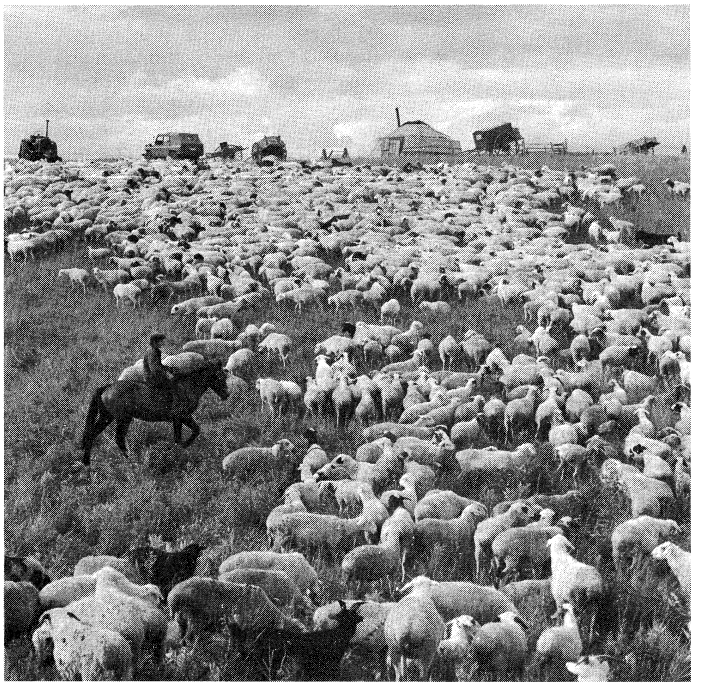
*Climate.* The seasons are marked by sharp fluctuations in the climate. Spring arrives in May and lasts for two months. Summer temperatures are relatively uniform. The July average is about 72° F (22° C) at Hu-ho-hao-t'e in the centre of the region; the yearly variation, however, is about 63° F (35° C). The two hottest months are July and August, when almost 60 percent of the annual rainfall occurs. Winter, which arrives after mid-September and lasts until March, is bitterly cold, with strong, icy winds blowing out of Siberia. Rainfall is meagre. In the Gobi areas the yearly rainfall is below four inches (100 millimetres); the plateau area receives only about 12 inches. The development of farming is handicapped by a frost-free period that lasts from 110 to 160 days; the consequent short growing season demands that the harvest occur within a 100-day period after planting. Raising two crops a year is impossible, and droughts occur almost annually.

*Plant and animal life.* Much of the territory consists of grassland, which provides pasture for sheep, goats, cattle, and the famous Mongolian horses and camels. Sheep and goats (roughly in equal proportions) are by far the most important, the most ubiquitous, and the most numerous of the animals reared on the grasslands. Mongolian horses are hardy and strong and are reputed to be able to travel from 18 to 25 miles a day for days.

*Settlement patterns.* The region remains primarily agricultural and pastoral, with few industrial centres. The three major urban population centres are Pao-t'ou, a large steel-and-iron complex and the terminal of two major railways connecting Peking and Kansu Province; Hu-ho-hao-t'e, capital of the autonomous region and a new cultural and political centre; and Chi-ning, a new commercial centre and the terminal of the Chi-ning–Erh-lien-hao-t'e railway.

**The people.** *Ethnic distribution.* Han (Chinese) constitute the bulk of the population, and the largest minority population is that of the Mongols. Minor groups include O-lun-chun, Solund, Evenk (Tungus), Korean, Daghur (Ta-hu-erh), and Manchu peoples. The population is unevenly distributed, with most people concentrated in the

Bays Galang—Xinhua News Agency



Mongol youth tending a flock of sheep grazing on grassland in the O-wen-k'o autonomous banner in the Hu-lun-pei-erh league, Inner Mongolia. A yurt, the Mongol tent, can be seen in the centre background.

agricultural southern belt, south of the Ta-ch'ing Mountains escarpment of the Mongolian Plateau, near the Huang Ho.

Languages

Because the Han greatly outnumber the Mongols, the most widely used language is Chinese. The Mongolian dialects belong to the eastern branch of Mongolian languages; they are phonetically, morphologically, and syntactically almost the same as the Khalkha Mongol dialect of Outer Mongolia to the north. A writing system of the Mongol language, using the Cyrillic alphabet, was introduced in 1955. The system is identical to the one used in Outer Mongolia. By 1960 all publications, except the classics, were printed in that alphabet.

*Religion.* In addition to ancestor worship, most of the Han in the region follow a religion formed of elements of Confucianism, Buddhism, and Taoism. The Mongols are mostly followers of Tibetan Buddhism, with almost every Mongol family having at least one son in a monastery.

Despite the prevalence of a form of Buddhism marked by ritual and a dominant, hierarchical monasticism, there are some aspects of shamanism. The stronghold of shamanism among the Mongols is the Hu-lun-pei-erh league (*meng*). Chinese, centred on Hu-ho-hao-t'e, are adherents of Islām.

**The economy.** *Agriculture.* Inner Mongolia, with almost one-third of China's grassland, has been traditionally renowned for its livestock. The condition of the livestock industry, despite its prominence, has been far from satisfactory. Various obstacles have slowed efforts to increase the number of livestock. In one year alone, some 58,000 head of cattle were taken by wolves. Many measures, including large-scale wolf hunting, the immunization of cattle, and improved pasturage and animal husbandry, have been taken to improve conditions. Weather stations have also been established to forewarn herders. Crossbreeding with the Soviet Tsgaisky pedigreed sheep by artificial insemination has also brought improvements.

Livestock ndustry

The harsh climate severely restricts intensive agriculture. In some areas, particularly around the great bend of the Huang Ho, oats, spring wheat, kaoliang (a variety of grain sorghum), millet, and other grains are cultivated. In irrigated areas sugar beets and oil-bearing crops such as linseed, rape, and sunflowers are major. Measures to improve agricultural output have been undertaken, such as the increased implementation of water conservation and irrigation programs and the use of chemical fertilizers.

*Industry.* Despite the relative backwardness of its industry, the territory possesses great mineral wealth. Rich iron-ore deposits have been discovered at Pai-yün-o-po, about 75 miles north of Pao-t'ou. Inner Mongolia also has one of the world's largest deposits of rare earth metals. Coal is also mined near Pao-t'ou, and one of the mines—the Jungeer—is the largest open-cut mine in China. The inland drainage of the Mongolian Plateau contained a number of salt lakes; most have dried up, leaving behind deposits of salt and natural alkali (soda). These resources are important for the chemical industry in general and the chemical fertilizer industry in particular.

Industrial development has centred around Pao-t'ou; the city was enlarged to become one of the major iron-and-steel centres of China. Numerous modern plants, including those for ceramics, cement, machine building and repairing, textiles, and chemical fertilizer were built in the area, establishing it as one of the most important industrial bases north of the Huang Ho.

*Transportation.* Most of the rail system is in the south, linking the region to the remainder of China. The Feng-t'ai and Pao-t'ou railway starts from Peking and traverses the southern corner of Inner Mongolia. With the advent of industrial development, two new main lines and several forest railways were constructed. The Chi-ning and Ulan Bator International Railway (completed in 1955) connects China with Mongolia and with the Soviet Union. This route shortened the rail distance between Peking and Moscow by some 700 miles. The most important line constructed since 1949, however, is that from Pao-t'ou to Lan-chou in Kansu Province, which completes the rail link between northern and northwestern China.

The rail ietwork

In addition to the rail network, thousands of miles of highway link most areas. Inland waterway navigation is somewhat limited. Only the upper course of the Huang Ho, from Lan-chou, in Kansu, to Ho-k'ou, in Inner Mongolia, is navigable.

**Administration and social conditions.** *Government.* The administration differs in name and composition from those in other parts of China. The region is divided into eight leagues (*meng*), which are similar to subprovincial units in China proper, and four prefecture-level municipalities (*shih*). The local administrative units are banners (*ch'i*) in the Mongolian areas and counties (*hsien*) in the predominantly Han area. In the Mongol areas the banners are subdivided into administrative villages (*gatsaa*) or *aimak* (units of two or three villages); in the nomadic region the banners are subdivided into *sumun,* which, in turn, are divided into *bag* (groups of nomad farmers), *khoto* (towns), and *ail* (settlements of a few families of nomads).

In accordance with the policy of fostering unity between the nationalities, an effort has been made to set up "democratic coalition governments" in localities where both Mongols and Han are represented in substantial numbers.

*Education.* Education was introduced after 1949, mainly through mobile schools and a "half-study, half-work" scheme in which study time varied according to the requirements of agriculture. More than 60 percent of the population has received at least a primary-level education, and illiteracy has dropped to less than 25 percent. A number of vocational schools, colleges, and universities are also in operation.

*Health and welfare.* Most of the Mongols live in tents, called yurts, that are inadequately ventilated. This, added to a critical shortage of drinking water and the Mongols' reluctance to wash their clothes and bodies, contributes to the spread of epidemic diseases. Syphilis and bubonic plague caused a continuous decline in the Mongolian population in the course of the past century. In 1947, for example, more than 60 percent of the pastoral population suffered from syphilis. Though the infant mortality rate in 1949 was as high as 31 percent, public health has since been greatly improved, and the spread of venereal diseases and plague has been brought under control. Energetic promotion of new midwifery methods significantly reduced the rate of infant mortality, and the population decline has come to a halt.

Conditions of pastoral life

**Cultural life.** Cultural life bears the deep imprint of Tibetan Buddhist influence. In liturgical music, monastery and temple architecture, scriptural learning and commentary, and religious arts, the Mongols accepted the forms of Tibet. Though the specific content and emphasis of Mongol folk legends vary somewhat with the location and with tribal or clan history concerning their origins, most clans have legends of their founders as either a mythical animal or a hero; others preserve legends about historical figures once prominent in the life of their clan. The subjects and themes of Mongol folktales and other forms of vernacular literature tend to be standard among all the tribes. A very large number concern lamas and religious life. Legends and songs as well as riddles and jokes occupy the leisure time of the night camp and its fireside circle, which form a major aspect of Mongolian life.

Mongolian music is not an independent art but serves solely as accompaniment to songs, dances, and rites. Singing is a form of entertainment, communication, historical recollection, group fellowship, and exuberant expression, and it demonstrates the close affiliation of individual Mongols with their culture and traditions. Mongol singing is generally a gregarious activity, mostly taking place around campfires, after the evening meal.

The Mongols observe seasonal celebrations: the New Year, the celebration of the White Month (signifying rebirth) in spring, the Midsummer Festival on the 12th day of the sixth month, the Autumn Festival (Festival of Fire) on the first day of the eighth month of the Chinese lunar calendar, and the Great Sacrificial Feast to the Fire God on the 23rd day in the 12th lunar month.

Besides the temple festivals, there is the Obo (shrine) Festival, held in the fifth month of every year. Toward the end of the ceremonies the festival takes a joyful course without restraint. There are wrestling and archery competitions, and a race is held in which the young men of the

tribes ride their best horses. This is the time for a dashing display of the talent and vigour of the Mongol nomads.

With the increasing Sinicization of the region both in terms of numbers and influence, many Han cultural forms have become prominent. Minority national troupes and a number of regional institutes seek to encourage and preserve the indigenous cultural traditions.

HISTORY

Farming was carried out on the marshes near the present boundary of Inner Mongolia and the provinces to the south in early times. The area was the limit of expansion of intensive agricultural settlement and was thus the scene of frequent confrontations between nomadic steppe dwellers and settled agriculturalists. In 658 BC several states of the North China Plain combined their efforts to build a wall defending what is now Hopeh from nomadic incursions and annexed part of Inner Mongolia to their agricultural territory. This part of Mongolia was inherited by the rulers of the Ch'in dynasty when they unified the Chan-kuo (Warring States) into an empire in the 3rd century BC.

Emperors of the succeeding Han dynasty waged war against the powerful Hsiung-nu, who were based in the valley of the northern bend of the Huang Ho. After pushing the Hsiung-nu north of the river, the Han settled the Ordos Desert region. The decline of the Han dynasty in the 3rd century AD brought a series of nomadic rulers to northern China. Later the T'ang dynasty (618–907) again asserted control over China's northern border, constricting trade and prompting border raids.

Mongolian Empire

The establishment of the Mongolian Empire by Genghis Khan in the 13th century brought prestige and expanded trade to Inner Mongolia. Old raiding patterns returned with the Ming dynasty in China, but peaceful relations with China were reestablished when the Manchu rulers of the Ch'ing dynasty reorganized the tribes into banners and leagues and promoted trade through itinerant Han merchants.

During the 19th century, population pressure to the south brought many Chinese farmers into Mongolia in search of land to cultivate. This caused conflicts with herdsmen that culminated in independence for Outer Mongolia in 1912 and administrative autonomy for Inner Mongolia in 1932. Eastern Inner Mongolia was occupied by the Japanese from 1933 as part of the state of Manchukuo, and Japanese rule extended westward after 1937 during the Sino-Japanese War.

The Inner Mongolia Autonomous Region was founded by the Chinese Communist regime in 1947, more than two years prior to the establishment of its national government at Peking in 1949. It originally consisted of sections taken from western Heilungkiang Province and from former Chahar Province; some of this core territory is in eastern Inner Mongolia as constituted in the early 1970s. In a series of annexations in the 1950s, Inner Mongolia was greatly expanded to the northeast and east, west, and south; from 1956 to 1969 it extended in a great 1,600-mile arc from east of the Greater Khingan Range, then dipped southwest and west to the Pa-tan-chi-lin Desert in north central China proper, now in Kansu Province. During this period more than half of China's frontier with the Mongolian People's Republic was the Inner Mongolian border; at the far northeast, a small section of China's international boundary with the Soviet Union along the Amur River was in Inner Mongolia. In 1969 the Peking government reversed its previous policy by sharply cutting down the area of the autonomous region, transferring territory to the surrounding Han provinces in all directions. Only the international frontier with the Mongolian People's Republic remained unchanged. The areas transferred constituted about two-thirds of the former area of the region and contained almost half of its former population. In 1979 this reorganization was terminated and the territory detached in 1969 was restored to Inner Mongolia. Inner Mongolia traditionally has been an area of mixture and contact between the agrarian Chinese and the pastoral and nomadic Mongolians. The continuous territorial changes that have affected it have therefore signified the contradiction of diverse cultures and conflicting loyalties. Inner Mongolia has thus served as a testing ground for Chinese efforts to integrate Han and Mongols into a single unified political entity. (C.-y.C./V.C.F.)

Region of Sino-Mongolian intermixture

## Kansu

The province of Kansu (Kan-su in Wade–Giles romanization, Gansu in Pinyin), administratively a part of the northwest region, reaches into the geographic centre of China. It is bordered by the Mongolian People's Republic to the north; by the Hui Autonomous Region of Ningsia, the province of Shensi, and the Inner Mongolia Autonomous Region to the east; by the Chinese provinces of Tsinghai and Szechwan to the south; and by the Uighur Autonomous Region of Sinkiang to the west. A vital strategic pivot, linking China proper with the vast territory in the extreme west, the narrow corridor of Kansu has served for several centuries as a passageway between the upper Huang Ho (Yellow River) area and Chinese Turkistan. Kansu covers 175,300 square miles (454,000 square kilometres). The capital of Kansu is Lan-chou, on the south bank of the Huang Ho.

PHYSICAL AND HUMAN GEOGRAPHY

**The land.** *Relief.* Plateaus are the dominant physical features of Kansu. Along the southern border, the lofty Tsou-lang-nan and Ch'i-lien ranges separate Kansu from Tsinghai. These ranges have an average height of 12,900 feet (3,900 metres) above sea level. Near Lan-chou in central Kansu, the Huang Ho Valley opens out, and excellent agricultural land is available. Some 120 miles (190 kilometres) northwest of Lan-chou there is a stretch of interior drainage where the land is relatively flat and where glacier-fed streams, including the Hei River, disappear into the desert; this is the area referred to as the Kansu Corridor. The higher mountains nearby are covered with forests and their lower slopes are green with grasses, but the floor of the corridor itself is monotonously flat and barren yellow earth. Geologically, Tertiary formations (from 2,500,000 to 65,000,000 years old) appear in a number of basins in Kansu, with strata generally composed of red clays, conglomerates, red sandstones, and gypsum.

The topographical features of Kansu are relatively uncomplicated in the west and northwest, in contrast to the southeast, where the land has suffered local dislocations from earthquakes. In the northwest there are very few mountains but rather a hilly terrain that merges into the Gobi (Desert) to the east. The average altitude is about 3,000 feet. The eastern part of Kansu is the principal centre of earthquakes in China. From the 6th century AD to the present, major earthquakes have taken place on an average of once every 65 years, while minor quakes occur at least once every 10 years. One of the greatest disasters of modern times occurred in 1920, when a violent earthquake, centred in eastern Kansu, caused great landslides. The death toll was estimated at 246,000, and many cities and towns totally vanished.

The earthquake zone

*Climate.* The climate in Kansu undergoes sharp fluctuations of temperature in summer (June to August) and winter (December to February), with uneven and unpredictable precipitation throughout the year. In the west the average January temperature in Chiu-ch'üan, for instance, is 18° F (−8° C) and in Tun-huang, 200 miles west of Chiu-ch'üan, is 19° F (−7° C). The temperature in July in Chiu-ch'üan is 70° F (21° C), and in Tun-huang it is 81° F (27° C). Annual temperature variations for most parts of Kansu are more than 54° F (30° C).

Rainfall is meagre throughout most of Kansu. As one goes further inland, the precipitation becomes increasingly less frequent. In the western part of the province annual rainfall ranges from two inches (50 millimetres) at Tun-huang to three inches at Chiu-ch'üan. Irrigation depends mainly on melting snow from the Ch'i-lien Mountains. The southeastern part of the province, something of an exception to the general pattern, receives a relatively abundant rainfall. In P'ing-liang, 170 miles east of Lan-chou, rainfall reaches 20 inches. Summer is usually the period of maximum precipitation.

*Plant and animal life.* Although vegetation is rather

limited in the mountain area, primeval forests still exist in the high mountains of the Liu-p'an range in the eastern part of Kansu. On the floor of the Kansu Corridor, willows and poplars grow along the roads and ditches. Wild animals include marmots, deer, and foxes.

*Settlement patterns.* The Han (Chinese) and the Hui (Chinese Muslims) are essentially agriculturists, although some engage in trade and industry. The Mongols are pastoralists or are seminomadic. Important urban areas are centred on Lan-chou. The largest city in eastern Kansu is P'ing-liang. A major centre in western Kansu is Chiuch'üan, and nearby are the respective oil and mining centres of Yü-men and Chia-yü-kuan. The population is concentrated in the Lan-chou Basin, in the fertile valley plains of the south and central sections where irrigation is possible, and in the dry terrace land of the Liu-p'an Mountains. In western Kansu, population is intensively concentrated in a number of small, isolated oases scattered along the bases of the high snow-capped ranges.

Population distribution

**The people.** The Han constitute the main racial group in Kansu. Other ethnic groups are the Monguors (Mongols), the Turks (Salars and Sarig Uighurs), and the Tibetans. There are Mongols to the west of Lan-chou and Tibetans scattered over an area enclosed by the Chuanglang, Ta-t'ung, and Huang rivers.

The Han majority tends to follow the same traditional religious practices, such as Buddhism, generally observed elsewhere in China. The most important minority group in Kansu is the Hui, living mostly in the north and west; some are of Arab, Turkish, or Mongol origin. A few Muslims are converted Chinese. The Hui include believers in both the Sunnī and Shīʿah traditions. Tibetans and Mongols follow Tibetan Buddhism. Almost every Tibetan family has at least one son in a Buddhist monastery.

Most of the ethnic groups, including the Tibetan minority, speak Chinese as a second language. The Monguors, however, whose language differs completely from either Western or Eastern Mongolian, rarely speak a second language. Hui use both Chinese and Arabic scripts, although Arabic is usually used only for religious purposes.

Hui villages

Village life among the Han inhabitants is generally similar to that elsewhere in North China. In Hui villages, however, the religious-communal life-style is distinctly different. There is a small public building that serves as a mosque, where children gather regularly to receive religious instruction and to learn the alphabet and phonetics. Hui villages are, by comparison, more organized and possess more community spirit than is usual in the Han villages. Hitherto the two peoples have been mutually segregated.

Tibetan villages, in many aspects, are similar to Han villages. Those Tibetans who are sedentary, however, have no clearly defined clan organization, and their family ties are much looser than among the Han.

Village dwellings are generally mud huts. Some people live in caves—which may be elaborate, with fine furnishings, or simply scooped out of the porous yellow earth cliffs. Brick structures predominate in cities and towns. The eating habits of the people are slightly different from those of the Chinese in other parts of the nation. Coarse grains and wheat flour, rather than rice, are consumed.

**The economy.** Traditionally, Kansu has been an area of poverty. The frequency of earthquakes, droughts, and famines has contributed to the economic instability and low agricultural productivity of the region. Endowed with rich mineral resources, however, Kansu is building itself into a vital industrial base to support the exploitation of the province of Tsinghai to the south and the Uighur Autonomous Region of Sinkiang to the far west.

*Resources.* Kansu's minerals of greatest value are the oil reserves of Yü-men, in northwestern Kansu, and coal reserves, the chief mine of which is located about 20 miles south of Lan-chou. There is a large deposit of iron ore in the Tsou-lang-nan Mountains area in western Kansu. Other mineral resources include nickel, copper, lead, zinc, antimony, and rare earth metals. There are also deposits of limestone, gypsum, quartz, and other materials used in construction.

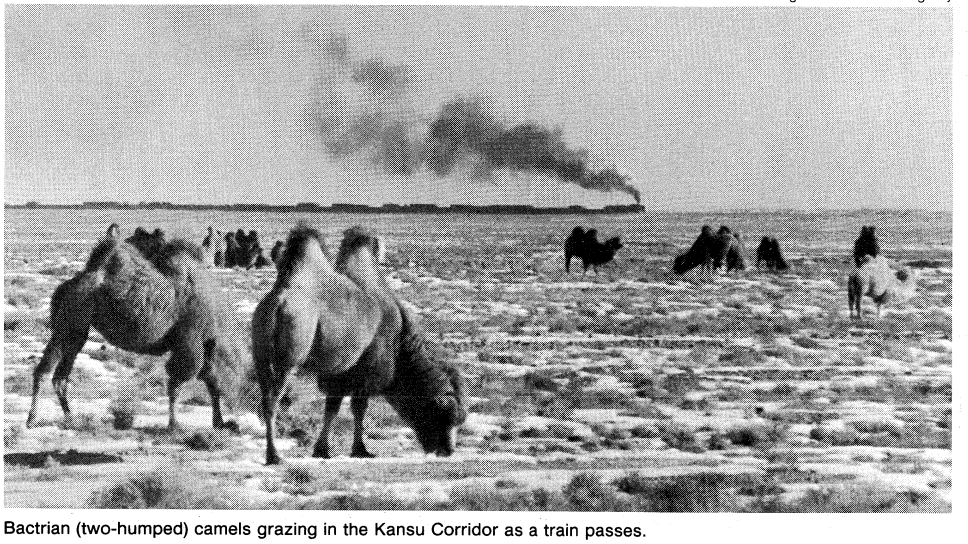
*Agriculture.* Although it is predominantly an agricultural area, and despite the fact that the per capita landholding is much larger than the national mean, output of food grain is insufficient to feed the population. The extent of cultivation in different areas depends on the elevation, the steepness of the slope of the land, and the dryness of the climate. High elevation has a greater precipitation and is therefore more favourable for farming. Terracing is prevalent and is practiced on about one-fifth of all of the cultivated land. Much of the hill land is cultivated by the use of a modified form of contour plowing. Because the slopes of the fields are so steep, however, and the fields so extensive, erosion is a serious problem, and some of the land has been abandoned. Agriculture in this area depends on the improvement of irrigation.

Agricultural resources

Some modernization has taken place since 1949, including increased irrigation and mechanization and the introduction of chemical fertilizers. The fertile Kansu Corridor produces most of the province's food crops, which include wheat, barley, millet, corn (maize), and tubers. The province is also a modest producer of sugar beets, rapeseed, soybeans, and a variety of fruits. Attempts have been made to increase agricultural output by transforming vast areas of wasteland along the Kansu Corridor into cotton fields. More than one-third of this area is suitable for cotton. In addition, wool and tobacco are produced as cash crops. Kansu is famous for its water-pipe tobacco, which is raised near Lan-chou and farther west. Kansu's vast grasslands support large herds of livestock, about half of which are sheep. Bactrian (two-humped) camels are raised in the Kansu Corridor.

*Industry.* Since 1950 strenuous efforts have been made

Li Shengcai—Xinhua News Agency



Bactrian (two-humped) camels grazing in the Kansu Corridor as a train passes.

to develop Kansu into an industrial base for northwestern China, with Lan-chou as its focus. A traditional regional centre located at the crossroads to Central Asia and the old Silk Road, Lan-chou has been a processing centre and entrepôt for centuries. Modern industrial development began only with the arrival of the railroad to Lan-chou in 1952 and its penetration through the Kansu Corridor to Yü-men and beyond in the mid-1950s. During the mid-1950s emphasis was placed on establishing heavy industry in Lan-chou. The city became a major producer of petroleum and now has dozens of other large modern industrial enterprises, including plants that produce petroleum drilling and refining equipment, locomotive equipment, chemical fertilizers, and petrochemicals. Efforts have also been made to build Lan-chou into a base for nuclear industry. Other important industrial installations in Kansu include the oil refinery at Yü-men and the iron-and-steel plant at nearby Chiu-ch'üan.

*Transportation.* The major barrier to development in this area has been the absence of transportation facilities. Before 1952 only the Lung-hai Railway connected Kansu with the coastal area; in that year an extension between Lan-chou and T'ien-shui to the southeast was completed. In addition, a railway extends northwestward from Lan-chou via Yü-men to Wu-lu-mu-ch'i, the capital of Sinkiang. Railways have also been built connecting Lan-chou to the Inner Mongolia Autonomous Region and to the rich mineral area of the Tsaidam Basin in northwest Tsinghai. The highway system has also been greatly expanded. Highways radiate from Lan-chou toward Tsinghai, Sinkiang, Inner Mongolia, Shensi, and Szechwan. Because of considerable silting and the river's seasonal flow, navigation on the Huang Ho is limited to the section between Lan-chou and Chung-wei.

Railway construction

**Administration and social conditions.** *Government.* From 1949 to 1954 Kansu was subject to the authority of the Northwest Military Affairs Commission. After 1954 the province came directly under the jurisdiction of the central government in Peking. Lan-chou, however, remained a military regional headquarters.

The provincial government has its headquarters in Lan-chou. Three municipalities (*shih*) are under the direct supervision of the provincial government—Lan-chou itself; Chia-yü-kuan, at the western terminus of the Great Wall (which runs from northwest to southeast through the province); and Chin-ch'ang in the central sector of Kansu. Intermediate administrative divisions include eight prefectures (*ti-ch'ü*) and two autonomous prefectures (*tzu-chih-chou*)—the Lin-hsia-hui-tsu Autonomous Prefecture, inhabited by Hui, and the Kan-nan-tsang-tsu Autonomous Prefecture, inhabited by Tibetans. On the third level of administration the province is further divided into counties (*hsien*), autonomous counties (*tzu-chih hsien*), and municipalities (*shih*) under county jurisdiction.

*Education.* The educational standard is comparatively lower than elsewhere in North China, and the percentage of people with at least a primary-level education is well below the national average. Since 1950 educational facilities have been greatly expanded, however. Universities and colleges are mostly located in Lan-chou, including Lan-chou University, the Northwest Normal College, and the Northwest Institute for Minorities. Special colleges providing training for railway work, the petroleum industry, animal husbandry, and veterinary medicine are also established in Lan-chou.

*Health and welfare.* By Western standards, the area is backward in health and sanitation. The most common diseases are the fecal-borne intestinal diseases spread through the use of human waste as fertilizer. The shortage of water supplies and the lack of modern doctors, nurses, and pharmacists constitute a serious problem. The state has funded projects to dig wells and channel water in afflicted areas.

Welfare is more concerned with the victims of natural disasters than with the poor in general. Frequent earthquakes and severe droughts require the government to assume responsibility for relief. In the Hui community, a part of the public welfare is organized by the Muslims themselves; Muslim officials collect obligatory charity for this purpose. Since 1949 the government has made general progress in Kansu with its welfare program for workers and peasants. New residential areas, for instance, have been built in Chiu-ch'üan for families of workers in the Yü-men oil fields. Medical clinics have been established in remote areas, where most people previously relied on local herb doctors.

**Cultural life.** Kansu represents a colourful mixture of races, customs, and cultures. The land abounds with mosques, monasteries of lamas, and Chinese temples.

Communal life in Han villages is marked by religious observances, particularly rituals connected with ancestor worship; seasonal celebrations, such as the New Year, the Dragon Boat Festival, and the Moon Festival; and customs relating to birth, marriage, funerals, and burials. All of these activities are similar to those of the Han throughout the nation. Village theatricals provide another type of communal activity.

Religious festivals

Most of the Monguors and Tibetans have abandoned their nomadic way of life and have become sedentary villagers. They live in brick and mud houses resembling their former tents (yurts). Tibetans insist on simultaneous group actions within the village. Every year, when the first day of spring planting is determined by the horoscope, for instance, the villagers go to the fields in their best clothing. The fields are then plowed simultaneously, and the seeds are sown at the same time in each field. During the course of the growing season, the villagers periodically parade through the fields carrying holy books on their heads.

The Hui are faithful followers of Islām and strictly observe the month-long fast of Ramaḍān, during which they abstain from food, drink, and sexual intercourse between sunrise and sunset. Before darkness falls, pious, bearded men say their prayers in public, and one or two of the elders may preach on points of theology, quoting the Qur'ān in oddly mutilated Arabic. At nightfall a communal feast is eaten; the community fires blaze all night, and people call and shout to one another. Among the Hui, the *ḥājjīs*, those who have completed the pilgrimage to Mecca, are highly respected in the community. The number of pilgrimages has, however, decreased considerably since 1949.

The western part of Kansu has long been a region renowned for ancient and classic artistic works. Stone caves in Tun-huang have many kinds of religious paintings on their walls, dating from the T'ang dynasty (AD 618–907). In Wu-wei large numbers of writings on bamboo slips have been found on the sites of the old frontier garrisons of the Han Empire (206 BC–AD 220). In 1964 a coherent bamboo text comprising a large part of one of the classic works on ritual (the *I Li*) came to light in western Kansu. In Tun-huang, within a Buddhist cave-temple, a library was discovered that had been immured there in the year 1035. It consisted of voluminous rolls of texts, including many valuable paintings and Buddhist classics.

## HISTORY

Kansu became a part of Chinese territory during the Ch'in dynasty (221–206 BC), when Chinese power began to extend up to the Kansu Corridor and into the region of modern Ningsia and Tsinghai. In ancient times all traffic between China proper and the far west was funneled through the Kansu Corridor. Along the ancient Silk Road that began at Ch'ang-an (modern Sian) and continued through the corridor, camel caravans carried the tea, silk, and porcelain of China to bazaars in the Middle East and even to the markets of Byzantium and Rome. In the train of these caravans such travelers as the Buddhist missionary Kumārajīva and the Venetian merchant Marco Polo entered China.

The name of Kansu first came into existence in the Yüan, or Mongol, dynasty (1206–1368), when it comprised the districts of Kan-chou and Su-chou. In the Ch'ing dynasty (1644–1911/12) Kansu covered the later provinces of Kansu, Ningsia, a part of Tsinghai, and a part of Sinkiang. The area was under the administration of a governor general of Shensi-Kansu, who was stationed at Lan-chou and had authority over both provinces. One of the most prominent governors general was Tso Tsung-t'ang (1812–85), who after 1878 brought a half century of peace to Kansu. A hero in the suppression of the Taiping

Tso Tsung-t'ang

Rebellion, Tso also helped the Ch'ing court to put down the Muslim rebellion in Kansu, which lasted for 16 years (1862–78) and affected more than 10,000,000 people.

Before Tso assumed the governorship, Kansu was an area without law and order. The Hui in Kansu were in open rebellion, committing murder, arson, and numerous other crimes. After having effectively destroyed their strongholds, Tso extended Chinese educational and civil service systems into the conquered districts for the benefit of Hui and non-Hui alike. As a result, the violence subsided and peace prevailed.

Kansu remained a province of China during the period of the Chinese republic (1911–49). The territory, however, shrank substantially when Sinkiang, Tsinghai, and Ningsia became independent provinces in 1928. During the 1920s and '30s the province was controlled by Muslim warlords. The provincial leader, Ma Chung-ying, of the Ma clan of Ho-chou, Kansu, was wooed by both the Japanese and Russians, but Ma came to accept nominal Nationalist Party (Kuomintang) authority in the region.

Communist influence in Kansu began in 1935, after the Chinese Red Army withdrew from southeast China to Shensi, and a Communist-controlled Shensi–Kansu–Ningsia border government was established in the late 1930s. In the final stages of the civil war, the People's Liberation Army defeated Ma's troops and took Lan-chou in August 1949.

The area within Kansu's jursidiction has undergone several changes since 1950. In 1954 Kansu annexed the province of Ningsia. In 1956 the A-la-shan-yu Ch'i and O-chi-na Ch'i banners in northwestern Kansu were detached and incorporated into the Inner Mongolia Autonomous Region. In 1958 the affixed Ningsia Province was separated from Kansu to become the Hui Autonomous Region of Ningsia. In 1969 the two aforementioned banners were returned to Kansu again, leaving the territory of Kansu almost unchanged when compared with its 1950 area. In 1979, however, the banners received a decade earlier from Inner Mongolia were again detached from Kansu and transferred to Inner Mongolia.

(C.-y.C./V.C.F./Ed.)

## Shensi

Shensi (Shaan-hsi in Wade–Giles romanization, Shaanxi in Pinyin) is a province of China, bordered by the Inner Mongolia Autonomous Region to the north, by the Huang Ho (Yellow River) and Shansi Province in the east, by Szechwan Province in the south, by Hupeh and Honan provinces in the southeast, and by the Hui Autonomous Region of Ningsia and Kansu Province on the west. Its total area is 79,400 square miles (205,600 square kilometres). The capital is Sian.

### PHYSICAL AND HUMAN GEOGRAPHY

The three natural regions

**The land.** *Relief and drainage.* Shensi Province comprises three distinct natural regions—the mountainous southern region, the Wei River valley, and the northern upland plateau.

The mountainous southern region forms the drainage area of the upper Han River, which is a northern tributary of the Yangtze. The Han flows between two mountain complexes that structurally form part of a great, single fold zone. These complexes are the Ta-pa Mountains, forming the boundary with Szechwan Province to the south, and the Tsinling Mountains—the major environmental divide between northern and central China—to the north. The Ta-pa Mountains range from 5,000 to 6,000 feet (1,500 to 1,800 metres) in height, with individual peaks reaching altitudes of up to 8,000 feet. Its northern flank in Shensi is heavily dissected by the complex pattern of the Han River's southern tributaries. The only major break in this mountain chain occurs in the far southwest of the province where the Chia-ling River, which rises to the north in the Tsinling Mountains, cuts through the Ta-pa chain to flow into Szechwan on its way to join the Yangtze at Chungking. This valley forms the major communication route from the Wei Valley in central Shensi to Szechwan and the southwest.

The Han Valley itself broadens out near the city of Hanchung into a fertile and densely cultivated basin about 60 miles (100 kilometres) long and 10 miles broad. Farther downstream the valley again narrows, after which the river flows between mountains and through deep gorges, only emerging into the plain once more in Hupeh Province.

The Tsinling Mountains to the north of the Han Valley form an even more impressive barrier than the Ta-pa range. Structurally a continuation of the great Kunlun Mountains to the west, the range runs continuously across Shensi from west to east at an average height of some 8,000 feet, with individual peaks reaching 12,300 feet. The range merges into the Fu-niu and the Hsiung-erh Mountains in Honan. The main watershed of the range is in the north; the southern slope of the range, draining into the Han, is deeply sculptured by an extremely complex drainage pattern. Three major passes cross the Tsinling Mountains: the San-kuan Pass south of Pao-chi, which leads to the Chia-ling Valley and thus into Szechwan; the Kao-kuan Pass south of Sian, which leads to the Han-chung Basin; and the Lan-t'ien Pass southeast of Sian, which affords a route to Nan-yang in Honan and to northern Anhwei Province.

The Wei Valley

The second major region is the valley of the Wei River, a tributary of the Huang Ho, which flows from west to east across the province from its headwaters in Kansu to join the Huang Ho at the border with Shansi and Honan. This valley is a major geological trough, bounded on the south by a vast complex of faults and fractures along the base of the Tsinling Mountains; it is a zone of considerable seismic instability, especially vulnerable to earthquakes. The northern border of the Wei River trench is less abrupt, and the large northern tributaries of the Wei, Ching, and Lo rivers, have themselves formed in their lower courses quite extensive alluvial plains that are continuations of the Wei River plain. The plain consists largely of loess (which also mantles parts of the northern face of the Tsinling Mountains), as well as of redeposited loess washed off the plateau to the north. The rivers are heavily silted.

The third region, to the north, is the great upland plateau of northern Shensi. Structurally this is a basin of largely undisturbed sedimentary rocks of immense thickness. Its raised western rim forms the Liu-p'an Mountains, which extends from the far west of Shensi northward into Kansu and Ningsia. A minor northwest-to-southeast axis forms the Ta-liang and Huang-lung ranges, which constitute the watershed between the Lo River system and the northern part of the province, which drains directly into the Huang Ho. On the eastern border of the basin the Huang Ho flows from north to south through a narrow, gorgelike trough. In this section it falls some 2,000 feet in less than 500 miles, and it is mostly unnavigable, with frequent rapids, culminating in a very deep, narrow gorge and falls at Lung Gorge.

The whole of this basin plateau, which is mostly above 3,000 feet, is a peneplain (a region reduced almost to a plain by erosion) covered with a deep mantle of loess, blown from the Gobi and the Ordos Desert by the prevailing northwesterly winds of the winter season. Much of the area is covered to a depth of 150 or even 250 feet, and the loess completely masks the original relief and structure of the region. The loess, in turn, has been heavily eroded, leaving a characteristic landscape of almost vertical walls, cliff faces, and deep ravines. This erosion has been intensified by the effects of human occupation, which have destroyed the natural vegetation cover.

Climatic regions

*Climate.* The Tsinling Mountains are not only a physical divide but also separate Shensi into two sharply differentiated climatic regions. The southern mountain area has a subtropical climate, similar to that of the Middle Yangtze Basin or of Szechwan. Mean temperatures in January are from 37° to 39° F (3° to 4° C), and the frost-free growing season is from 260 to 280 days, although the summer and autumn are not so hot as in the Middle Yangtze region. Total precipitation is between 30 and 40 inches (750 and 1,000 millimetres), falling mostly between May and October. The driest part of the year is spring and early summer, when irrigation is necessary. But in general the climate is hot and moist. The rugged and varied topography, however, produces great local variations.

The Wei Valley has a much drier and somewhat colder climate. Average winter temperatures are about 32° F (0° C), and the frost-free period lasts for about 240 days. Total precipitation is between 20 to 25 inches, mostly falling between May and October, with a sharp peak in September and October. Rainfall is generally deficient in spring and early summer, but the climate is not seriously dry. It is, however, an area subject to severe and prolonged droughts. On the Loess Plateau farther north and west the climate grows progressively drier and colder. The extreme north and west have only about 10 inches of annual precipitation, most of which occurs in late summer and autumn, when evaporation loss is at its maximum. The growing season and frost-free period become progressively shorter until in the north the former is only about 190 days. In this area agriculture depends on techniques for minimizing evaporation losses and in conserving moisture in the soils. The northern frontier with the Inner Mongolia Autonomous Region, roughly coinciding with the line of the Great Wall, remains an important cultural divide. Beyond it conditions for agriculture become extremely precarious.

Vegetational zones

*Plant and animal life.* The vegetation in the northern and southern zones is also sharply differentiated. The southern mountain area forms a part of the mixed deciduous broad-leaved and evergreen forest zone that formerly covered the Lower Yangtze and Han river basins; this region is characterized by a rich variety of vegetation that includes more than 50 broad-leaved genera and a dozen or more coniferous genera. Owing to the difficulty of access, large areas of natural timber remain standing.

The northern slopes of the Tsinling Mountains and the lower Wei Valley were originally covered with deciduous broad-leaved forest. The bulk of northern Shensi, except for the pure steppe (treeless plain) of the northern and western borders, was originally part of the so-called northwestern forest-steppe area, where deciduous broad-leaved woodland grows only on the highest ground and around watercourses, most of the area being covered by grass or low, drought-resistant scrub. This northern area has been intensively cultivated since the 1st millennium BC, and its natural vegetation has been virtually destroyed.

**The people.** The people are nearly all Han (Chinese) and speak the Northwest Mandarin dialect. Some Hui (Chinese Muslim) communities remain in the south and northwest of the province. Most of the population is settled in the Wei and Han valleys; the uplands are more sparsely settled. The chief cities are Sian, Pao-chi, Hsien-yang, T'ung-ch'uan, and Han-chung. Other cities include An-k'ang, San-yüan, Yen-an, Wei-nan, and Yü-lin. Most of the province's county seats are very small. Sian is the provincial metropolis and the main communication centre and chief industrial city. It has an important university, a medical college, and an institute of art and music, as well as libraries and museums. Pao-chi is an important road and rail transportation centre. San-yüan and Hsien-yang are both satellite cities of Sian, as well as rail and road transport centres. Han-chung is the main communication and administrative centre for the southern region.

**The economy.** *Resources.* The basin in the north of the province has enormous coal reserves, second in size only to those of Shansi. Important modern mines are those at T'ung-ch'uan, on the northern slope of the Wei Valley, and at Han-ch'eng, near the Huang Ho. There are minor coal and oil-shale deposits in the Han Basin in the south, where there are also iron-ore deposits. The Tsinling Mountains contain some minor gold-producing areas (mostly in the west), as well as some minor deposits of manganese and other minerals. There are also significant deposits of molybdenum, graphite, zeolite, limestone, and barite.

*Agriculture.* The southern part of the province forms a portion of the southwestern highland and basin area, which is characterized by double-cropping, wet-field agriculture, and forestry production. Most of the cultivated land is below 3,000 feet. The Han-chung Basin grows rice intensively, followed by winter wheat, but in the mountain zones of the T'ai-pai and Tsinling ranges the main cereal crops are corn (maize) and winter wheat. Such subtropical crops as tea, tung oil, and citrus fruits are grown, as are a variety of other fruits.

The Wei Valley area is very intensively cultivated. Well over half of the total area is under cultivation, and it supports a dense agricultural population. The valley area produces some rice, good winter wheat, tobacco, and cotton, but millet, barley, and corn are also increasingly important crops, as is kaoliang (a variety of grain sorghum). On the higher ground, millet, oats, and buckwheat are common. Hemp, sesame, sugar beets, and rapeseed are important subsidiary crops, particularly in the upper Wei and the Ching valleys. Normally three crops are raised every two years.

The northern plateau is too cold in the winter for winter wheat to survive. It forms a part of the Inner Mongolia dry agricultural and pastoral zone. Spring-sown wheat and millet are the main grain crops, and these depend largely on the availability of irrigation water. Grazing becomes particularly important toward the northern and western borders, and the growing season is so short that only one crop yearly is possible.

Shensi's output and agricultural income remain below the national average, but improvements in a province known for famine and natural disaster have been considerable. Since the mid-1950s much attention has been directed toward stopping and reversing the extensive soil erosion that has long plagued the province north of the Tsinling Mountains. A large-scale multipurpose conservancy scheme has been underway on the Huang Ho, designed to reduce the enormous silt load discharged into the Huang by the Wei and its other west-bank tributaries. A great effort has been made to spread terraced cultivation in Shensi. The plan also calls for the construction of numerous dams in the loess uplands to retain silt before it reaches the Huang Ho. These small dams quickly silt up, forming new farmland. In addition, projects have been initiated to sow grass on denuded land, to plant trees for the protection of new terraced fields and slopes, and to prevent gullying. Even more ambitious is a plan to plant a belt of trees a mile or more wide, mostly consisting of drought-resistant poplar, elm, or willow, in an attempt to contain the spread of sand dunes from the Ordos Desert. This belt extends southwestward for about 375 miles from northeastern Shensi across Ningsia and into Kansu, skirting the edge of the desert.

The irrigation system

The irrigation system has also been greatly extended. The ancient irrigation systems of the Wei and Ching valleys, restored (after centuries of neglect) following the famine of 1932, have been extended, while great numbers of small dams and wells have also been constructed to increase the irrigated area.

*Industry.* The major industrial area in Shensi is that centred around Sian. Principal industries in this area include cotton and other textiles, electrical equipment, engineering and chemical manufacturing, and iron and

Richard and Sally Greenhill—Black Star



Young Han (Chinese) woman at work in a cotton mill in Sian, Shensi Province.

steel production. There are minor centres of industry at Pao-chi and at Shih-ch'üan in the Han Valley, near An-k'ang, and Yao-hsien, near T'ung-ch'uan, has a large and important cement plant. A small petroleum refinery is located at Yen-ch'ang. The production of consumer goods in Shensi has been emphasized, including bicycles, radios, televisions, watches, and apparel.

*Transportation.* The Wei Valley since prehistoric times has formed part of the main east–west route running from the North China Plain in the east to the Kansu Corridor and the steppelands in the west. Sian is a natural centre, where the great route east to west meets the routes that cross the Tsinling Mountains to the south and southeast, an alternative route to the northwest via the Ching Valley, and routes to the Ordos region in the north and to Shansi in the northeast. All these routes are now followed by modern highways. In the south a highway crosses the province from east to west, joining Han-chung with Wu-han, in Hupeh Province to the east, and Lan-chou, in Kansu Province to the west. In the far southwestern corner of Shensi, a main highway follows the route of an ancient post road from Pao-chi to Ch'eng-tu in Szechwan.

Road and rail routes

The first railway to reach Shensi was the Lung-hai line, the great east–west trunk line from the sea at Lien-yün-kang in Kiangsu, via the industrial centres of Honan. This line, extended in the 1930s through the Wei Valley to Pao-chi, was largely destroyed during the war with Japan. It was reconstructed in the late 1940s and extended westward to Kansu. A branch was also constructed from Hsien-yang to the coalfields at T'ung-ch'uan. Another major line now extends from Pao-chi to Ch'eng-tu in Szechwan, where it links with various lines to the southwest. Sian has become an important regional centre of air traffic.

The special districts

**Administration and social conditions.** The province has three prefecture-level municipalities (*shih*) directly subordinated to the provincial government. One of these municipalities includes the provincial capital, Sian. The rest of the province is organized into seven prefectures (*ti-ch'ü*); the northern plateau area is divided between the Yü-lin Prefecture in the far north and the Yen-an Prefecture farther south; the central and eastern Wei Valley is divided into the prefectures of Hsien-yang and Wei-nan; and the mountainous southern area is divided into the Han-chung, An-k'ang, and Shang-lo prefectures. At the next administrative level the province is divided into counties (*hsien*) and municipalities (*shih*) under the jurisdiction of prefectural governments. For purposes of economic planning, the province, together with Kansu and the Uighur Autonomous Region of Sinkiang, forms part of the northwestern economic region.

**Culture.** Citizens of Shensi take pride in their region as a historic centre of Chinese civilization and in their distinctive traditions in art, ceramics, and folksinging. The Yang-ko is a local form of musical folk opera with comic themes. Shensi-style Ch'in-ch'iang opera is also popular, as are shadow plays using local leather puppets.

## HISTORY

**Northern Shensi.** *The early period.* The northern parts of Shensi, particularly the Wei Valley, were some of the earliest settled parts of China. In the valley some remains of the Mesolithic Period have been found, while there are Neolithic Yang-shao culture sites spreading along the whole of the west–east corridor from Kansu to Honan, showing that this was already an important route. Chinese Neolithic culture was probably first developed in the Wei Valley. It remained an important centre of the later Neolithic Lung-shan culture and then became the first home of the Chou people, who in the late 12th century BC invaded the territories of their overlords, the Shang, to the east, and set up a dynasty in 1111 that exercised some degree of political authority over much of North China. Until 771 BC the political centre of the Chou was at Hao, near modern Sian.

For the early agriculturalists, working the ground with primitive stone-tipped tools, the slopes of loess and river terraces provided ideal farmland—light, stone-free, and fertile. The natural cover, too, was mostly grass and scrub and could be easily cleared for temporary cultivation.

After the 8th century BC the Chou lost much of their authority and moved their capital eastward to Lo-yang in Honan Province, after which Shensi became something of a backwater. Gradually, however, the predynastic Ch'in state, which controlled the area, began to develop into a strong centralized polity of a totally new kind, able to mobilize mass labour for vast construction projects, such as the part of the Great Wall of China built between Shensi and the Ordos Desert. One of the greatest of these tasks was the completion in the Wei Valley of a large and efficient irrigation system based on the Cheng-kuo and Pai-kung canals and centred around the junction of the Ching and Wei rivers. This system, completed in the 3rd century BC, watered some 450,000 acres (180,000 hectares) and provided the powerful economic base for the Ch'in's eventual conquest of the whole of China.

Ancient construction projects

*The middle period.* In 221 BC Hsien-yang, in Shensi, became the capital of the Ch'in dynasty, which unified China for the first time; it was a city of vast wealth and the focus of a nationwide road system. The area remained extremely populous and was a major centre of political authority for the next millennium. The Han (206 BC–AD 220), successors of the short-lived Ch'in dynasty, made their capital Ch'ang-an, near Hsien-yang. Later, in the 6th century, when after some centuries of disunion the Sui (581–618) again unified the empire, their capital—Ta-hsing—was on the same site as Ch'ang-an, which also was the capital of the T'ang (618–907). Ch'ang-an, as the capital was now once more known, was by far the largest and most magnificent city in the world in its day and was immensely wealthy. But by this time the irrigation system upon which Shensi primarily depended had begun to deteriorate, soil erosion and deforestation had begun to be problems, and the productivity of the area declined. The maintenance of a huge metropolis of more than 1,000,000 people in the area consequently necessitated the difficult and costly transportation of vast quantities of grain and provisions from the eastern plains and the Yangtze Valley. The capital remained in Shensi largely because the area (known as Kuan-chung—literally "Within the Passes") was easily defended and was of crucial importance, as a frontier with China's neighbours. After the sack of Ch'ang-an in 882, however, no dynasty ever again had its capital in the northwest, and the area rapidly declined in importance as the economic centre of the empire gradually gravitated toward the Yangtze Valley and the South. During the next millennium Shensi became one of the poorest and most backward of China's provinces.

*The early modern period.* Under the Mongols in the 13th century Shensi as a provincial unit assumed approximately its present form, incorporating the area formerly known as Shan-nan (literally "South of the Mountains"), or Li-chou. During this era, however, Shensi underwent many changes. In the course of the Yüan, or Mongol, dynasty (1206–1368) the province was devastated and largely depopulated as a result of the Mongol conquest. Subsequently there emerged a large Muslim element in the population. The area suffered badly from rebellion and disorders following the collapse of Mongol rule after about 1340, when two independent regimes—those of Chang Ssu-tao in the northwest and of Li Ssu-chi around Ch'ang-an—controlled most of Shensi. Later it was one of the areas in which disaffection with Ming rule (which began in 1368) first appeared in the late 1620s, and it was somewhat badly damaged in the fighting leading up to the Ch'ing conquest in 1644. Under Ming rule Shensi Province also incorporated Kansu, but under the Ch'ing dynasty (1644–1911/12) the two were separated once more.

*The 19th and 20th centuries.* By the 19th century Shensi was seriously impoverished. Although only marginally affected by the Taiping Rebellion (1850–64) in its last stages, eastern and southern Shensi were slightly disturbed by the Nien Rebellion between 1853 and 1868. It then suffered the terrible Muslim rebellion of 1862 to 1878, which affected much of the western and northern parts of the province. Although the effects of the rebellion and its savage suppression were not as terrible as in Muslim Kansu, about 600,000 were killed in Shensi, and the resulting destruction left the province in serious plight.

Periodic famines

As this rebellion was coming to an end, Shensi was also affected by one of the worst drought famines of modern times. It had virtually no rain from 1876 to 1878, and, when the government tried to remedy the situation in 1877, poor transport facilities prevented effective relief. Perhaps 4,000,000 or even 5,000,000 people died in Shensi alone, with some single counties in the fertile Wei Valley losing more than 100,000 people each. As a result of the terrible death toll in the last decades of the 19th century, Shensi became a haven for a wave of land-hungry immigrants from Szechwan and Hopeh provinces.

The end of the empire in 1911 brought yet further deterioration in living conditions. In 1912 the governors of Shensi and Kansu became engaged in a destructive civil war of an unusually brutal and violent character; the war, often affecting the whole province, continued until 1921, after which the province became involved in a still larger war between Yü-hsiang Feng and the Chihli warlords. In 1926 the capital, Sian, was besieged and badly damaged; the death toll numbered nearly 100,000 from starvation alone.

In the earlier years of the 20th century Shensi also suffered badly from periodic famines, which occurred in 1915, in 1921, and finally in 1928. This last famine was as severe as that of 1877–78; it is estimated that at least 3,000,000 people died of starvation, after which a wave of epidemics increased the death toll still further. Whole counties were virtually depopulated. This time, however, some measures of relief were forthcoming. The International Famine Relief Organization began to rehabilitate the derelict irrigation system of the Wei Valley, while the extension of the Lung-hai Railway into the province meant that, if in the future famine should threaten, relief supplies could quickly be moved into the province.

Communist base at Yen-an

A further political upheaval followed in 1936 when Communist armies, driven out of their bases in Kiangsi, passed through the western parts of Shensi. They then established themselves in Yen-an in northern Shensi, which was to be the base from which they conducted their war of resistance against the Japanese and from which, after the end of World War II, they successfully undertook the conquest of all China. In Shensi itself they controlled the territory of the present Yen-an and Yü-lin prefectures from 1937 onward.

**Southern Shensi.** The history of the southern part of the province has been considerably more placid than that of the north. Until the late 17th century the area was very sparsely peopled, and much of it, apart from the Han-chung basin, is still virgin forest. In the period after about 1680 the introduction of corn (maize) and sweet potatoes, followed in the 18th century by the introduction of the Irish potato, made upland farming possible. A pattern emerged of growing rice in the valley bottoms, corn on the lower mountain slopes, and Irish potatoes on the higher land. Southern Shensi, with its great amounts of vacant land, attracted immigrants on a large scale after severe famines and crop failures had occurred in Hupeh and Szechwan provinces in the 1770s. In the early 19th century immigrants from central and southern China constituted as much as 90 percent of the population in some parts of Southern Shensi.

Rapid and often reckless development of the uplands, however, often led to soil erosion, rapid loss of fertility, and declining crop output. Local disaffection broke out in the so-called White Lotus Rebellion of 1796–1804, which was centred in the Szechwan–Shensi–Hupeh–Honan border regions. After its suppression, however, the area remained generally peaceful: in the 20th century it escaped the worst excesses of the northwestern warlords' civil wars, as well as of the repeated famines that occurred in northern Shensi. (D.C.T./V.C.F.)

BIBLIOGRAPHY

**Physical and human geography.** *General works:* KEITH BUCHANAN, *The Transformation of the Chinese Earth* (1970); PING-CHIA KUO, *China,* 3rd ed. (1970); T.R. TREGEAR, *China: A Geographical Survey* (1980); ALBERT KOLB, *East Asia: China, Japan, Korea, Vietnam: Geography of a Cultural Region* (1971, reprinted 1977; originally published in German, 1963); BRIAN HOOK (ed.), *The Cambridge Encyclopedia of China* (1982); ROBERT L. WORDEN, ANDREA MATLES SAVADA, and RONALD E. DOLAN (eds.), *China, a Country Study,* 4th ed. (1988); and *Information China,* 3 vol. (1989), an encyclopaedic work prepared under the auspices of the Chinese Academy of Social Sciences and ed. by C.V. JAMES.

*Land and people:* Physical geography is surveyed in *The Physical Geography of China,* 2 vol., trans. from Russian (1969), prepared by the Institute of Geography, U.S.S.R. Academy of Sciences; SONGQIAO ZHAO, *Physical Geography of China* (1986); and SHIH-HSUN CH'EN, *The Climate of China* (1962). Studies of population include JUNE TEUFEL DREYER, *China's Forty Millions* (1976), an examination of China's minorities and their integration into society since 1949; C.K. LEUNG and NORTON GINSBURG (eds.), *China: Urbanization and National Development* (1980); LIU ZHENG *et al., China's Population: Problems and Prospects* (1981), containing surveys and case studies; JUDITH BANISTER, *China's Changing Population* (1987), an analysis of modern demographic trends; and *New China's Population* (1988), studies by Chinese demographers of three censuses. Atlases include CHIAO-MIN HSIEH, *Atlas of China,* ed. by CHRISTOPHER L. SALTER (1973); P.J.M. GEELAN and D.C. TWITCHETT (eds.), *The Times Atlas of China* (1974); CAROLINE BLUNDEN and MARK ELVIN, *Cultural Atlas of China* (1983); *The Population Atlas of China* (1987), on the 1982 census, prepared by the Population Census Office of the State Council of the People's Republic of China and the Chinese Academy of Sciences; and NATHAN SIVIN (ed.), *The Contemporary Atlas of China* (1988), including maps, illustrations, essays, and a chronology.

*Economy:* Analyses of Chinese economic development include AUDREY DONNITHORNE, *China's Economic System* (1967); ALEXANDER ECKSTEIN, *China's Economic Revolution* (1977); and WORLD BANK, *China,* 9 vol. (1982). Also see KENNETH RUDDLE and WU CHUANJUN (eds.), *Land Resources of the People's Republic of China* (1983); STANLEY D. RICHARDSON, *Forestry in Communist China* (1966); FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS, *Forestry in China* (1982); CHRISTOPHER HOWE, *China's Economy* (1978); A. DOAK BARNETT, *China's Economy in Global Perspective* (1981); *China Quarterly,* no. 100 (December 1984), a special issue on economic reform; UNITED STATES. CONGRESS. JOINT ECONOMIC COMMITTEE, *An Economic Profile of Mainland China* (1968), *China: A Reassessment of the Economy* (1975), *Chinese Economy Post-Mao* (1978), *China Under the Four Modernizations* (1982), and *Chinese Economy in the Eighties* (1985), collections of studies by China specialists; EDWARD L. WHEELWRIGHT and BRUCE MCFARLANE, *The Chinese Road to Socialism: Economics of the Cultural Revolution* (1970); THEODORE SHABAD, *China's Changing Map: National and Regional Development, 1949–71,* 2nd ed. (1972); NICHOLAS R. LARDY and KENNETH LIEBERTHAL (eds.), *Ch'en Yun's Strategy for China's Development: A Non-Maoist Alternative* (1983); DOROTHY J. SOLINGER, *Chinese Business Under Socialism* (1984), an analysis of controversies over the organization and role of commerce in China's development to 1980; NICHOLAS R. LARDY, *Economic Growth and Distribution in China* (1978); and SAMUEL P.S. HO and RALPH W. HUENEMANN, *China's Open Door Policy: The Quest for Foreign Technology and Capital: A Study of China's Special Trade* (1984). LILIAN CRAIG HARRIS and ROBERT L. WORDEN (eds.), *China and the Third World* (1986), is a collection of essays on China's role in the redistribution of the resources of the industrialized world. Reference works in English published by or with the cooperation of the Chinese government include *Almanac of China's Economy* (annual) by the ECONOMIC RESEARCH CENTRE; and *Statistical Yearbook of China* (annual) by the STATE STATISTICAL BUREAU.

Information on railway, air, and highway communication may be found in general works on the Chinese economy (see above). For the role of transportation in the history of Western influence on China and on regional economics, see SHUN-HSIN CHOU, "Railway Development and Economic Growth in Manchuria," *China Quarterly,* no. 45, pp. 57–84 (1971). Articles on transportation appear intermittently in *Beijing Review* (weekly).

*Administration, social conditions, and cultural life:* Two general surveys are JOHN K. FAIRBANK, *The United States and China,* 4th enlarged ed. (1983); and C.T. HU *et al., China: Its People, Its Society, Its Culture* (1960), with a topical bibliography. See also FREDRIC M. KAPLAN and JULIAN M. SOBIN, *Encyclopaedia of China Today,* 3rd rev. ed. (1982); WILLIAM HINTON, *Fanshen: A Documentary of Revolution in a Chinese Village* (1966), an eyewitness account of 1948 revolutionary activity in a village in Shansi; MARC BLECHER, *China, Politics, Economics, and Society* (1986), a survey of China since 1949; FRANZ SCHURMANN, *Ideology and Organization in Communist China,* 2nd enlarged ed. (1968), a sociological analysis of Communist ideology and institutions on the eve of the Cultural Revolution; PING-TI HO and TANG TSOU (eds.), *China's Heritage and the Communist Political System* in *China in Crisis,*

vol. 1 (1968), a critical analysis of the Chinese political system during the early phase of the Cultural Revolution; JOHN WILSON LEWIS (ed.), *Party Leadership and Revolutionary Power in China* (1970), analyses of the early impact of the Cultural Revolution on the Chinese Communist leadership and its policies; RICHARD H. SOLOMON, *Mao's Revolution and the Chinese Political Culture* (1971), an examination of Mao's thought as applied to the problems of political socialization of the masses; JOHN M.H. LINDBECK (ed.), *China: Management of a Revolutionary Society* (1971), studies of cultural, ideological, and institutional changes; HARRY HARDING, *Organizing China: The Problem of Bureaucracy, 1949–1976* (1981); JAMES R. TOWNSEND, *Politics in China,* 2nd ed. (1980), a summary of works on all aspects of political history; BENEDICT STAVIS, *China's Political Reforms* (1988), an analysis of some of the reforms of the 1980s; HARVEY W. NELSEN, *The Chinese Military System,* 2nd rev. ed. (1981), an examination of decision making in the People's Liberation Army; ELLIS JOFFÉ, *The Chinese Army After Mao* (1987), an assessment of the PLA's modernization; PI-CHAO CHEN, *Population and Health Policy in the People's Republic of China* (1976); DANIEL L. OVERMYER, *Religions of China: The World as a Living System* (1986), an introduction to religion in everyday life; and KENNETH LIEBERTHAL and MICHEL OKSENBERG, *Policy Making in China* (1988), an analysis of the influence of political institutions on policy-making.

The following periodicals give current information on political, cultural, and economic affairs: *China Quarterly; Far Eastern Economic Review* (weekly); *China Business Review* (bimonthly); *China Reconstructs* (monthly); *China Pictorial* (monthly); and *China News Analysis* (biweekly).

(C.-S.Ch./K.G.L./K.P.S./E.I.U./K.J.De.W.)

**History.** *General works: The Cambridge History of China,* DENIS TWITCHETT and JOHN K. FAIRBANK (general eds.), will be the standard multivolume reference work for all aspects of Chinese history when it is completed; the following volumes were published from 1978 to 1988: vol. 1, *The Ch'in and Han Empires, 221 B.C.–A.D. 220;* vol. 3, *Sui and T'ang China, 589–906,* part 1; vol. 7, *The Ming Dynasty, 1368–1644,* part 1; vol. 10, *Late Ch'ing, 1800–1911,* part 1; vol. 11, *Late Ch'ing, 1800–1911,* part 2; vol. 12, *Republican China 1912–1949,* part 1; vol. 13, *Republican China 1912–1949,* part 2; and vol. 14, *The People's Republic, Part 1: The Emergence of Revolutionary China, 1949–1965.* The following are comprehensive works: WOLFRAM EBERHARD, *A History of China,* 4th ed. (1977; originally published in German, 1948); JOHN K. FAIRBANK and EDWIN O. REISCHAUER, *China: Tradition and Transformation* (1978); OTTO FRANKE, *Geschichte des chinesischen Reiches,* 2nd ed., 5 vol. (1948–65); JACQUES GERNET, *A History of Chinese Civilization* (1982; originally published in French, 1972); CHARLES O. HUCKER, *China's Imperial Past: An Introduction to Chinese History and Culture* (1975); and JOHN MESKILL (ed.), *An Introduction to Chinese Civilization* (1973).

*Prehistory:* CHI LI, *The Beginnings of Chinese Civilization* (1957, reprinted 1968); KWANG-CHIH CHANG, *The Archaeology of Ancient China,* 3rd rev. enlarged ed. (1977), a pioneering, comprehensive synthesis based on the coverage of excavations reported to the end of 1976; DAVID N. KEIGHTLEY (ed.), *The Origins of Chinese Civilization* (1983), a collection of articles about environment and agriculture, cultures and peoples, and the genesis of the state; JESSICA RAWSON, *Ancient China: Art and Archeology* (1980), a study of artistic significance of Neolithic and Bronze Age artifacts; WILLIAM WATSON, *Cultural Frontiers in Ancient East Asia* (1971), an illustrated analysis of cultural interaction; and RUKANG WU and JOHN W. OLSEN (eds.), *Palaeoanthropology and Palaeolithic Archaeology in the People's Republic of China* (1985), an account of Chinese research on early man.

*The first historical dynasty: the Shang:* In addition to the works of Chang, Watson, and Rawson mentioned above, see also KWANG-CHIN CHANG, *Early Chinese Civilization: Anthropological Perspectives* (1976), and *Shang Civilization* (1980), on Shang archaeology and culture; TSUNG-TUNG CHANG, *Der Kult der Shang-Dynastie im Spiegel der Orakelinschriften* (1970); DAVID N. KEIGHTLEY, *Sources of Shang History: The Oracle-Bone Inscriptions of Bronze Age China* (1978, reprinted 1985); WEN FONG (ed.), *The Great Bronze Age of China* (1980), an analytical catalog of an exhibition from the People's Republic of China; and PAUL WHEATLEY, *The Pivot of the Four Quarters* (1971), a comparative study of the origins of urban society.

(D.N.K.)

*The Chou and Ch'in dynasties:* CHI LI, *The Formation of the Chinese People* (1928, reprinted 1967), useful for information on the ethnic history of ancient China; HERLEE GLESSNER CREEL, *The Birth of China* (1937, reprinted 1954), still regarded as one of the standard references, and *The Origins of Statecraft in China,* vol. 1 (1970), the first Western book to include extensive materials from bronze inscriptions—especially significant on the activities of non-Chou peoples; CHO-YUN HSU, *Ancient China in Transition* (1965), a standard work on social structure and social mobility in the Chou period; DERK BODDE, *China's First Unifier* (1938, reprinted 1967), and *Statesman, Patriot, and General in Ancient China* (1940, reprinted 1967), works covering the effort of the first emperors and their courts to accomplish unification; ARTHUR COTTERELL, *The First Emperor of China* (1981), a popular history that includes a description and historical analysis of 7,000 terra-cotta life-size figures buried in 210 BC; YU-NING LI, *The First Emperor of China* (1975), YU-LAN FUNG, *A History of Chinese Philosophy,* 2nd ed., 2 vol., trans. from the Chinese (1952–53, reprinted 1983), coverage of the thought of different schools in ancient China; and XUEQIN LI, *Eastern Zhou and Qin Civilizations* (1986), an account of archaeological findings of the period from the 8th to the 3rd century BC.

(C.-y.H.)

*The Han dynasty:* Translations of historical source material can be found in ÉDOUARD CHAVANNES (ed. and trans.), *Les Mémoires historiques de Se-ma Ts'ien* (vol. 1–5, 1895–1905, reprinted 1967; vol. 6, 1969); HOMER H. DUBS (ed. and trans.), *The History of the Former Han Dynasty,* 3 vol. (1938–55); ESSON M. GALE (ed. and trans.), *Discourses on Salt and Iron* (1931, reprinted 1973); BURTON WATSON (ed. and trans.), *Records of the Grand Historian of China,* 2 vol. (1961); HSIEN-YI YANG and GLADYS YANG (trans.), *Selections from Records of the Historian* (1979); and A.F.P. HULSEWÉ, *Remnants of Ch'in Law* (1985). Works on specific topics of Han history include HANS BIELENSTEIN, *The Restoration of the Han Dynasty,* 4 vol. (1953–79); MICHAEL LOEWE, *Records of Han Administration,* 2 vol. (1967), *Everyday Life in Early Imperial China During the Han Period 202 B.C.–A.D. 220* (1968, reissued 1973), and *Crisis and Conflict in Han China 104 B.C. to A.D. 9* (1974); T'UNG-TSU CH'U, *Han Social Structure* (1972); A.F.P. HULSEWÉ, *Remnants of Han Law* (1955); CHO-YUN HSU, *Han Agriculture: The Formation of Early Chinese Agrarian Economy, 206 B.C.–A.D. 220* (1980); and YING-SHIH YU, *Trade and Expansion in Han China* (1967). ZHONGSBU WANG, *Han Civilization* (1982), is a survey of archaeological investigations of the Han dynasty.

(J.L.D.)

*Period of division:* The most extensive modern account of the period of division is that in OTTO FRANKE (*op. cit.*), vol. 2, with copious notes in vol. 3. This is a traditional chronological history, which pays little attention to nonpolitical matters and absolutely none to modern historical writing on the period in Chinese and Japanese. Other works include WOLFRAM EBERHARD, *Das Toba-Reich Nordchinas: Eine soziologische Untersuchung* (1949), a Western-language study on the T'o-pa Wei—controversial, interesting, but highly technical; W.J.F. JENNER, *Memories of Loyang* (1981), a political history of the Wei dynasty during the Pei-ch'ao (Northern Dynasties) period; ÉTIENNE BALAZS, "Les Courants intellectuels en Chine au III$^{e}$ siècle de notre ère," *Études Asiatiques,* vol. 2 (1948), the best Western-language study on the rise of "Neo-Taoism" and other schools of thought after the breakdown of the Han Empire; HENRI MASPERO, *Taoism and Chinese Religion* (1983; originally published in French, 1950), a collection of essays dealing mainly with T'ang dynasty Taoism, still the most important general survey of the Taoist religion of this period, written mainly in the 1930s and '40s by a great authority for the general public; HOLMES WELCH, *Taoism: The Parting of the Way,* rev. ed. (1965), a general history of the Taoist movement, with about one-third of the book devoted to the development of Taoist religion in the Six Dynasties period; MICHEL STRICKMANN, *Le Taoïsme du Mao Chan: chronique d'une révélation* (1981), a scholarly account of one of the main schools of medieval Taoism; ARTHUR F. WRIGHT, *Buddhism in Chinese History* (1959, reprinted 1971), a popular but authoritative survey of Chinese Buddhism as a whole, two chapters of which are devoted to the Six Dynasties period; KENNETH K.S. CHEN, *Buddhism in China* (1964, reprinted 1972), an extensive history of Chinese Buddhism by an eminent specialist; ERIK ZÜRCHER, *The Buddhist Conquest of China,* 2 vol. (1959, reprinted 1972), a detailed, rather technical study of the formation of gentry Buddhism; and JACQUES GERNET, *Les Aspects économiques du bouddhisme dans la société chinoise du V$^{e}$ au X$^{e}$ siècle* (1956), an indispensable but rather technical work on the economic functions of the Buddhist monasteries from the 5th to the 10th century.

(E.Z./D.C.T.)

*Sui and T'ang periods:* By far the best general history of the periods is in volume 3 of the *Cambridge History of China* (1979). The most important work on the Sui period is ARTHUR F. WRIGHT, *The Sui Dynasty* (1978). See also his two studies "The Formation of Sui Ideology 581–604," in JOHN K. FAIRBANK (ed.), *Chinese Thought and Institutions* (1957); and "Sui Yang-ti: Personality and Stereotype," in ARTHUR F. WRIGHT (ed.), *The Confucian Persuasion* (1960). WOODBRIDGE BINGHAM, *The Founding of the T'ang Dynasty: The Fall of Sui and Rise*

*of T'ang* (1941, reprinted 1970), gives a clear account of the period from 607 to 624. Two books by C.P. FITZGERALD, *The Son of Heaven: A Biography of Li Shih-Min, Founder of the T'ang Dynasty* (1933, reprinted 1971), and *The Empress Wu,* 2nd ed. (1968), give a traditional account of the political history of the early T'ang, without advantage of later scholarship. More modern analyses of the early T'ang are presented in HOWARD J. WECHSLER, *Mirror to the Son of Heaven: Wei Cheng at the Court of T'ang T'ai-tsung* (1974), and in his *Offerings of Jade and Silk: Ritual and Symbol in the Legitimation of the T'ang Dynasty* (1985). R.W.L. GUISSO, *Wu Tse-t'ien and the Politics of Legitimation in T'ang China* (1978), is a comprehensive study of Empress Wu's reign. On the Empress' use of Buddhist ideology, see A. FORTE, *Political Propaganda and Ideology in China at the End of the Seventh Century* (1976). EDWIN G. PULLEYBLANK, *The Background to the Rebellion of An Lu-shan* (1955, reprinted 1982), gives a full account of every aspect of the reign of Hsüan-tsung. Also interesting on the politics of the same period is P.A. HERBERT, *Under the Brilliant Emperor: Imperial Authority in T'ang China as Seen in the Writings of Chang Chiu-ling* (1978). There is no modern full-scale study of the An Lu-shan rebellion itself, but the two principal sources exist in well-annotated translations by HOWARD S. LEVY (ed. and trans.), *Biography of An Lu-shan* (1960); and by ROBERT DES ROTOURS (trans.), *Histoire de Ngan Lou-chan* (1962). There is no satisfactory book-length account of the following period. The rebellions of the 780s are described briefly in DENIS TWITCHETT, "Lu Chih (754–805)," in ARTHUR F. WRIGHT and DENIS TWITCHETT (eds.), *Confucian Personalities* (1962). The mysterious reign of Shuntsung has not been subjected to a modern analytic study, but the principal source is translated in BERNARD S. SOLOMON (ed. and trans.), *The Veritable Record of the T'ang Emperor Shuntsung* (1955). There is some account of the subsequent reigns in ARTHUR WALEY, *The Life and Times of Po Chü-i, 772–846 A.D.* (1949), but the historical analysis is somewhat outdated. For the period after 847 even the Chinese primary documentation becomes very thin. The only events that have attracted attention of Western scholars are the rebellions; on these see ROBERT DES ROTOURS, "La Révolte de P'ang Hiun," *T'oung Pao,* 56:229–240 (1970); and HOWARD S. LEVY (ed. and trans.), *Biography of Huang Ch'ao,* 2nd rev. ed. (1961). On the Huang Ch'ao rebellion and the subsequent disorders the best account is GUNGWU WANG, *The Structure of Power in North China During the Five Dynasties* (1963, reissued 1967). A number of important studies on T'ang political history, taking account of modern Japanese and Chinese scholarship, are included in ARTHUR F. WRIGHT and DENIS TWITCHETT (eds.), *Perspectives on the T'ang* (1973). JOHN CURTIS PERRY and BARDWELL L. SMITH (eds.), *Essays on T'ang Society: The Interplay of Social, Political, and Economic Forces* (1976), is also a significant collection. On T'ang institutions in general there is a perceptive though outdated (the book was actually completed in 1951) account in HENRI MASPERO and ÉTIENNE BALAZS, *Histoire et institutions de la Chine ancienne,* pp. 160–262 (1967). The traditional sources on the administrative system are translated by ROBERT DES ROTOURS, *Traité des fonctionnaires, et Traité de l'armée,* 2 vol., 2nd rev. ed. (1974), and on the examination system in his *Traité des examens,* 2nd rev. ed. (1976). On finances and general economic problems, see DENIS TWITCHETT, *Financial Administration Under the T'ang Dynasty,* 2nd ed. (1970), which contains a full bibliography. On finance under the Sui, see ÉTIENNE BALAZS, *Études sur la société et l'économie de la Chine médiévale,* 2 vol. (1953–54). For translations of the main documents on T'ang law, see KARL BÜNGER, *Quellen zur Rechtsgeschichte der T'ang-Zeit* (1946); and WALLACE JOHNSON (trans.), *The T'ang Code,* vol. 1 (1979). On general social history, see DENIS TWITCHETT, *Land Tenure and the Social Order in T'ang and Sung China* (1962), "The T'ang Market System," *Asia Major,* 12:202–248 (1966), "Merchant, Trade and Government in Late T'ang," *Asia Major,* 14:63–95 (1968), and *Birth of the Chinese Meritocracy: Bureaucrats and Examinations in T'ang China* (1976); also see the relevant sections of ÉTIENNE BALAZS, *Chinese Civilization and Bureaucracy,* trans. from the French (1964, reprinted 1972). More information is to be found in ROBERT M. HARTWELL, "Demographic, Political, and Social Transformations of China, 750–1550," *Harvard Journal of Asiatic Studies,* 42:365–442 (December 1982); and in the eyewitness account by a contemporary Japanese monk in EDWIN O. REISCHAUER (trans.), *Ennin's Diary* (1955), with a companion volume, *Ennin's Travels in T'ang China* (1955). See also, on the very important evidence from Tun-huang, LIONEL GILES, *Six Centuries of Tunhuang* (1944); and DENIS TWITCHETT, "Chinese Social History from the Seventh to the Tenth Centuries: The Tunhuang Documents and Their Implications," *Past and Present,* 35:28–53 (1966). The following are extremely important for the light they throw on the cosmopolitan nature of T'ang society: EDWARD H. SCHAFER, *The Vermilion Bird: T'ang Images of the South* (1967), *The Golden Peaches of Samarkand: A Study of T'ang Exotics* (1963, reprinted 1985), and *Shore of Pearls* (1970), dealing with the early history of Hai-nan Island. The best general account of Chinese relations with its northern neighbours in the steppes is RENE GROUSSET, *The Empire of the Steppes* (1970; originally published in French, 1939; 4th French ed., 1960). On Chinese overseas trade and relations with Southeast Asia, see GUNGWU WANG, *The Nanhai Trade: A Study of the Early History of Chinese Trade in the South China Sea* (1958). (D.C.T.)

*Five Dynasties, Ten Kingdoms, and Sung periods:* EDWARD H. SCHAFER, *The Empire of Min* (1954), has the best sinological summary on this kingdom in the South. On conquest dynasties in the North, see HOK-LAM CHAN, *The Historiography of the Chin Dynasty: Three Studies* (1970), his *Legitimation in Imperial China: Discussions Under the Jurchen-Chin Dynasty, 1115–1234* (1984); and JING-SHEN TAO, *The Jurchen in Twelfth-Century China: A Study of Sinicization* (1976). On the Sung period, listings of works in Western languages are available in YVES HERVOUET, *Bibliographie des travaux en langues occidentales sur les Song parus de 1946 à 1965* (1969). Sequels to it appear in various volumes of *Sung Studies Newsletter* (ceased publication in 1978), and its successor, the *Bulletin of Sung and Yüan Studies* (annual). C.P. FITZGERALD, "The Chinese Middle Ages in Communist Historiography," in ALBERT FEUERWERKER (ed.), *History in Communist China* (1968), provides information on some earlier work done by historians in mainland China. On the significance of the Sung and varying interpretations, see JAMES T.C. LIU and PETER J. GOLAS (eds.), *Change in Sung China: Innovation or Renovation?* (1969). Further general information is available in the following: EDWARD A. KRACKE, JR., *Civil Service in Early Sung China, 960–1067* (1953), a monumental work; JAMES T.C. LIU, *Ou-yang Hsiu, an Eleventh-Century Neo-Confucianist,* trans. from the Chinese (1967), and *Reform in Sung China: Wang An-shih (1021–1086) and His New Policies* (1959), both works attempting to relate general trends through historical figures; and BRIAN E. MCKNIGHT, *Village and Bureaucracy in Southern Sung China* (1971), an account coming down to the later period as well as the local scene. JACQUES GERNET, *Daily Life in China on the Eve of the Mongol Invasion, 1250–1276* (1962; originally published in French, 1959), provides vivid descriptions. See also PATRICIA BUCKLEY EBREY (ed. and trans.), *Family and Property in Sung China: Yüan Ts'ai's Precepts for Social Life* (1984), an excellent picture of family relations as studied in the 12th-century work by Yüan Ts'ai; THOMAS H.S. LEE, *Government Education and Examinations in Sung China* (1985), a description of education and its relationship to the civil service; and BRIAN E. MCKNIGHT (trans.), *The Washing Away of Wrongs: Forensic Medicine in Thirteenth-Century China* (1985), a translation of an early work on the topic. (J.T.C.L./B.E.McK.)

*The Yüan dynasty:* The standard Chinese source is the official dynastic history, *Yüan shih,* completed in 1370. Supplementing this, for the history of Genghis Khan and Ögödei Khan, is the Chinese rendition of the Mongolian historical literary narrative, *Yüan ch'ao pi shih.* It is known in the West through several translations, such as *The Secret History of the Mongols,* trans. by FRANCIS WOODMAN CLEAVES (1982); or *The Secret History of the Mongols,* adapted by PAUL KAHN (1984). A partial listing of Western works on Yüan history is available in HENRY G. SCHWARZ, *Bibliotheca Mongolica: Works in English, French, and German* (1978). On the Mongol operations against China, see the early work of H. DESMOND MARTIN, *The Rise of Chingis Khan and His Conquest of North China* (1950, reissued 1971), now partially superseded by IGOR DE RACHEWILTZ, "Personnel and Personalities in North China in the Early Mongol Period," *Journal of the Economic and Social History of the Orient,* 9:88–144 (1966), a study of the Mongol conquest. On the political history of the later Yüan, see JOHN W. DARDESS, *Conquerors and Confucians: Aspects of Political Change in Late Yüan China* (1973). For works on governmental institutions and legal and military systems, see PAUL RATCHNEVSKY (ed.), *Un Code des Yuan,* 2 vol. (1937–72); FRANZ SCHURMAN (ed. and trans.), *Economic Structure of the Yüan Dynasty* (1956, reissued 1967); HOK-LAM CHAN, "Liu Ping-chung (1216–74): A Buddhist-Taoist Statesman at the Court of Khubilai Khan," *T'oung Pao,* 53:98–146 (1967); HERBERT FRANKE, *From Tribal Chieftain to Universal Emperor and God: The Legitimation of the Yüan Dynasty* (1978); CH'I-CH'ING HSIAO, *The Military Establishment of the Yüan Dynasty* (1978); PAUL HENG-CHAO CH'EN, *Chinese Legal Tradition Under the Mongols: The Code of 1291 as Reconstructed* (1979); and JOHN D. LANGLOIS, JR. (ed.), *China Under Mongol Rule* (1981). For religious and intellectual life of the period, see ARTHUR WALEY (trans.), *The Travels of an Alchemist* (1931, reprinted 1963), on Ch'ang-ch'un; FREDERICK W. MOTE, "Confucian Eremitism in the Yüan period," in ARTHUR F. WRIGHT, *The Confucian Persuasion* (1960); CH'EN YUAN, *Western and Central Asians in China Under the Mongols* (1966), an annotated translation of an authoritative Chinese work; WM.

THEODORE DE BARY, "The Rise of Neo-Confucian Orthodoxy in Yüan China," in his *Neo-Confucian Orthodoxy and the Learning of the Mind-and-Heart* (1981); and HOK-LAM CHAN and WM. THEODORE DE BARY (eds.), *Yüan Thought: Chinese Thought and Religion Under the Mongols* (1982). For works on art and literature of the Yüan dynasty, see SHERMAN E. LEE and WAI-KAM HO, *Chinese Art Under the Mongols: The Yüan Dynasty, 1279–1368* (1968); WAYNE SCHLEPP, *San-ch'ü: Its Technique and Imagery* (1970); CHUNG-WEN SHIH, *The Golden Age of Chinese Drama: Yüan Tsa-chü* (1976); JAMES CAHILL, *Hills Beyond a River: Chinese Painting of the Yüan Dynasty, 1279–1368* (1976); RICHARD JOHN LYNN, *Kuan Yün-shih* (1980), a biography of the poet; and J.I. CRUMP, *Chinese Theater in the Days of Khublai Khan* (1980), and *Songs from Xanadu: Studies in Mongol-Dynasty Song-Poetry (San-ch'u)* (1983). On Chinese contacts with Asia and the West, see E. BRETSCHNEIDER, *Mediaeval Research from Eastern Asiatic Sources: Fragments Towards the Knowledge of the Geography and History of Central and Western Asia from the 13th to the 17th Century,* 2 vol. (1888, reprinted 1967); HERBERT FRANKE, "Sino-Western Contacts Under the Mongol Empire," *Journal of the Hong Kong Branch of the Royal Asiatic Society,* 6:49–72 (1966); LEONARD OLSCHKI, *Marco Polo's Asia: An Introduction to His "Description of the World" called "Il Milione"* (1960; originally published in Italian, 1957); and IGOR DE RACHEWILTZ, *Papal Envoys to the Great Khans* (1971). (H.Fr./H.Ch.)

*Ming dynasty:* The standard reference guide to traditional Chinese-language materials is WOLFGANG FRANKE, *An Introduction to the Sources of Ming History* (1968). Essential information is contained in L. CARRINGTON GOODRICH and CHAO-YING FANG (eds.), *Dictionary of Ming Biography* (1976). Reviews and articles regularly appear in the journal *Ming Studies* (semiannual). ALBERT CHAN, *The Glory and Fall of the Ming Dynasty* (1982), is a book-length survey of the whole period, but it does not make use of recent critical scholarship. The early Ming years are described and analyzed in EDWARD L. DREYER, *Early Ming China: A Political History, 1355–1435* (1982); and CHARLES O. HUCKER, *The Ming Dynasty: Its Origins and Evolving Institutions* (1978). Additional light is shed on early Ming life, thought, and institutions in EDWARD L. FARMER, *Early Ming Government: The Evolution of Dual Capitals* (1976); FREDERICK W. MOTE, *The Poet Kao Ch'i* (1962); and JOHN W. DARDESS, *Confucianism and Autocracy: Professional Elites in the Founding of the Ming Dynasty* (1983). Early Ming overseas expeditions and foreign relations are dealt with in J.J.L. DUYVENDAK, *China's Discovery of Africa* (1949); and YI-T'UNG WANG, *Official Relations Between China and Japan, 1368–1549* (1953). Specialized studies of mature Ming include CHARLES O. HUCKER, *The Traditional Chinese State in Ming, 1368–1644* (1961), *The Censorial System of Ming China* (1966), and CHARLES O. HUCKER (ed.), *Chinese Government in Ming Times: Seven Studies* (1969); RAY HUANG, *Taxation and Governmental Finance in Sixteenth-Century Ming China* (1974); KWAN-WAI SO, *Japanese Piracy in Ming China During the 16th Century* (1975); PING-TI HO, *Studies on the Population of China, 1368–1953* (1959, reprinted 1967), and *The Ladder of Success in Imperial China: Aspects of Social Mobility, 1368–1911* (1962, reprinted 1976); and AYAO HOSHI *The Ming Tribute Grain System,* trans. from the Chinese by MARK ELVIN (1969). RAY HUANG, *1587, A Year of No Significance: The Ming Dynasty in Decline* (1981), is a wide-ranging critical discussion of Ming governance and the ruling class; a somewhat more approving view of Ming China of the same period is *China in the Sixteenth Century: The Journals of Matthew Ricci, 1583–1610,* trans. from the Latin by LOUIS J. GALLAGHER (1953). Modern studies of China's contacts with Europeans in Ming times notably include T'IEN-TSE CHANG, *Sino-Portuguese Trade from 1514 to 1644* (1934, reprinted 1973); CHARLES R. BOXER (ed. and trans.), *South China in the Sixteenth Century* (1953, reprinted 1967); and GEORGE H. DUNNE, *Generations of Giants: The Story of the Jesuits in China in the Last Decades of the Ming Dynasty* (1962). The last Ming years and the struggles of post-Ming loyalists are dealt with in JAMES B. PARSONS, *The Peasant Rebellions of the Late Ming Dynasty* (1970); LYNN A. STRUVE, *The Southern Ming, 1644–1662* (1984); and JONATHAN D. SPENCE and JOHN E. WILLS, JR. (eds.), *From Ming to Ch'ing: Conquest, Region, and Continuity in Seventeenth-Century China* (1979). Studies in Ming intellectual and religious history are found in WM. THEODORE DE BARY *et al., Self and Society in Ming Thought* (1970), and *The Unfolding of Neo-Confucianism* (1975).

(C.O.Hu.)

*The Ch'ing period:* (*The rise of the Ch'ing dynasty*): IMMANUEL C.Y. HSU, *The Rise of Modern China,* 3rd ed. (1983); ROBERT H.G. LEE, *The Manchurian Frontier in Ch'ing History* (1970); SILAS H.L. WU, *Communication and Imperial Control in China: Evolution of the Palace Memorial System, 1693–1735* (1970); and FREDERIC WAKEMAN, JR., *The Great Enterprise: The Manchu Reconstruction of the Imperial Order in Seventeenth-Century China* (1985). (*Early foreign relations*): LUCIANO PETECH, *China and Tibet in the Early 18th Century,* 2nd rev. ed. (1972); CHUSEI SUZUKI, "China's Relations with Inner Asia: The Hsiung-nu, Tibet," pp. 180–197 in JOHN K. FAIRBANK (ed.), *The Chinese World Order* (1968); EARL H. PRITCHARD, *Anglo-Chinese Relations During the Seventeenth and Eighteenth Centuries* (1929, reissued 1970), and *The Crucial Years of Early Anglo-Chinese Relations, 1750–1800* (1936, reissued 1970); ANTONIO S. ROSSO, *Apostolic Legations to China of the Eighteenth Century* (1948); and MARC MANCALL, *Russia and China: Their Diplomatic Relations to 1728* (1971). (*Mid-Ch'ing society and economy*): CHUNG-LI CHANG, *The Chinese Gentry* (1955, reprinted 1974); T'UNG-TSU CHU, *Local Government in China Under the Ch'ing* (1962, reissued 1969); MADELEINE ZELIN, *The Magistrate's Tael: Rationalizing Fiscal Reform in Eighteenth-Century Ch'ing China* (1984); and YEH-CHIEN WANG, *Land Taxation in Imperial China, 1750–1911* (1973). (*Intellectual and cultural aspects*): L. CARRINGTON GOODRICH, *The Literary Inquisition of Ch'ien-lung,* 2nd ed. (1966); JOSEPH R. LEVENSON, *Confucian China and Its Modern Fate,* 3 vol. (1958–65, reissued in 1 vol., 1968), and "The Abortiveness of Empiricism in Early Ch'ing Thought," *Far Eastern Quarterly,* 13:155–165 (1954); CH'I-CH'AO LIANG, *Intellectual Trends in the Ch'ing Period,* trans. by IMMANUEL C.Y. HSU (1959); and EVELYN SAKAKIDA RAWSKI, *Education and Popular Literacy in Ch'ing China* (1979). (*Dynastic degeneration*): JEAN CHESNEAUX (comp.), *Secret Societies in China in the Nineteenth and Twentieth Centuries* (1971; originally published in French, 1965); KUNG-CHUAN HSIAO, *Rural China: Imperial Control in the Nineteenth Century* (1960, reprinted 1967); GILBERT ROZMAN (ed.), *The Modernization of China* (1981), an interdisciplinary study of China's modernization between the Opium Wars and 1980; PHILIP A. KUHN, *Rebellion and Its Enemies in Late Imperial China: Militarization and Social Structure, 1794–1864* (1970, reprinted 1980); and SUSAN NAQUIN, *Millenarian Rebellion in China: The Eight Trigrams Uprising of 1813* (1976). (*Western challenges*): MASATAKA BANNO, *China and the West, 1858–1861: The Origins of the Tsungli Yamen* (1964); HSIN-PAO CHANG, *Commissioner Lin and the Opium War* (1964, reprinted 1970); W.C. COSTIN, *Great Britain and China, 1833–1860* (1937, reprinted 1968); JOHN K. FAIRBANK, *Trade and Diplomacy on the China Coast,* 2 vol. (1953, reissued in 1 vol., 1969); MICHAEL GREENBERG, *British Trade and the Opening of China, 1800–42* (1951, reprinted 1979); IMMANUEL C.Y. HSU, *China's Entrance into the Family of Nations: The Diplomatic Phase, 1858–1880* (1960); and PETER WARD FAY, *The Opium War, 1840–1842* (1975).

(*The Taipings*): VINCENT Y.C. SHIH, *The Taiping Ideology: Its Sources, Interpretations, and Influences* (1967, reprinted 1972); FRANZ H. MICHAEL and CHANG CHUNG-LI, *The Taiping Rebellion: History and Documents,* 3 vol. (1966–71); and ALBERT FEUERWERKER, *Rebellion in Nineteenth-Century China* (1975). (*Nien Rebellion*): SIANG-TSEH CHIANG, *The Nien Rebellion* (1954, reprinted 1967); S.Y. TENG, *The Nien Army and Their Guerrilla Warfare, 1851–1868* (1961, reprinted 1984); and ELIZABETH J. PERRY, *Rebels and Revolutionaries in North China, 1845–1945* (1980). (*Chinese response*): ALBERT FEUERWERKER, *China's Early Industrialization: Sheng Hsüan-huai (1844–1916) and Mandarin Enterprise* (1958, reissued 1970); YEN-P'ING HAO, *The Comprador in Nineteenth Century China: Bridge Between East and West* (1970); PAUL A. COHEN, *Between Tradition and Modernity: Wang T'ao and Reform in Late Ch'ing China* (1974); WILLIAM T. ROWE, *Hankow: Commerce and Society in a Chinese City, 1796–1889* (1984); LUKE S.K. KWONG, *A Mosaic of the Hundred Days: Personalities, Politics, and Ideas of 1898* (1984); and KUNG-CH'UAN HSIAO, *A Modern China and a New World: K'ang Yu-wei, Reformer and Utopian, 1858–1927* (1975). (*Boxer Rebellion*): PAUL A. COHEN, *China and Christianity: The Missionary Movement and the Growth of Chinese Antiforeignism, 1860–1870* (1963); and VICTOR PURCELL, *The Boxer Uprising* (1963, reprinted 1974).

For a comprehensive survey of the development of state power, see JOHN K. FAIRBANK, *The Great Chinese Revolution, 1800–1985* (1986). (*Revolutionary movements at the end of the Ch'ing*): MICHAEL GASSTER, *Chinese Intellectuals and the Revolution of 1911: The Birth of Modern Chinese Radicalism* (1969); ROBERT A. SCALAPINO and GEORGE T. YU, *The Chinese Anarchist Movement* (1961, reprinted 1980); HAROLD Z. SCHIFFRIN, *Sun Yat-sen and the Origins of the Chinese Revolution* (1968); JOHN H. FINCHER, *Chinese Democracy, the Self-Government Movement in Local, Provincial, and National Politics, 1905–1914* (1981); MARY B. RANKIN, *Early Chinese Revolutionaries: Radical Intellectuals in Shanghai and Chekiang, 1902–1911* (1971); EDWARD J.M. RHOADS, *China's Republican Revolution: The Case of Kwangtung, 1895–1913* (1975); JOSEPH W. ESHERICK, *Reform and Revolution in China: The 1911 Revolution in Hunan and Hubei* (1976); and EDMUND S.K. FUNG, *The Military*

*Dimension of the Chinese Revolution: The New Army and Its Role in the Revolution of 1911* (1980). (C.Su./A.Fe.)

*Twentieth-century China: (Republican China)*: An accomplished text on this period is JAMES E. SHERIDAN, *China in Disintegration: The Republican Era in Chinese History, 1912–1949* (1975); more detailed and politically centred is O. EDMUND CLUBB, *20th Century China,* 3rd ed. (1978). There are two excellent collections of biographies of the leading figures of the period: HOWARD L. BOORMAN and RICHARD C. HOWARD (eds.), *Biographical Dictionary of Republican China,* 5 vol. (1967–79); and DONALD W. KLEIN and ANNE B. CLARKE, *Biographic Dictionary of Chinese Communism, 1921–65,* 2 vol. (1971). For the revolution of 1911–12 and the early republican period, scholarly essays offering fresh interpretations are found in MARY C. WRIGHT (ed.), *China in Revolution: The First Phase, 1900–1913* (1968). The immediate consequences of the revolution of 1911–12 and the search for an appropriate political form for China are analyzed in EDWARD FRIEDMAN, *Backward Toward Revolution: The Chinese Revolutionary Party* (1974, reprinted 1977); and ERNEST P. YOUNG, *The Presidency of Yuan Shih-K'ai: Liberalism and Dictatorship in Early Republican China* (1977). The dimensions of the ebullient intellectual and political movements in China in the late teens are still best presented in a classic work, TSE-TUNG CHOW, *The May Fourth Movement: Intellectual Revolution in Modern China* (1960, reissued 1967). For a more critical view of some of the leading participants, see YÜ-SHENG LIN, *The Crisis of Chinese Consciousness: Radical Antitraditionalism in the May Fourth Era* (1979).

The period of warlordism has been comprehensively treated by HSI-SHENG CH'I, *Warlord Politics in China, 1916–1928* (1976). Among studies of particular warlord formations, especially notable are JAMES E. SHERIDAN, *Chinese Warlord: The Career of Feng Yü-hsiang* (1966); and DONALD E. SUTTON, *Provincial Militarism and the Chinese Republic: The Yunnan Army, 1905–25* (1980). Economic conditions are discussed in R.H. TAWNEY'S prophetic and still valuable study, *Land and Labour in China* (1932, reprinted 1966). Essays in DWIGHT H. PERKINS (ed.), *China's Modern Economy in Historical Perspective* (1975), are relevant. A penetrating analysis of a critical sector of China's rural social order is PHILIP C.C. HUANG, *The Peasant Economy and Social Change in North China* (1985). On the Nationalist revolution, see C. MARTIN WILBUR and JULIE LIEN-YING HOW (eds.), *Documents on Communism, Nationalism, and Soviet Advisers in China, 1918–1927* (1956, reissued with corrections, 1972). The struggles that displaced warlords from centre stage are set forth in C. MARTIN WILBUR, *The Nationalist Revolution in China, 1923–1928* (1984). RICHARD W. RIGBY, *The May 30 Movement: Events and Themes* (1980), is a study of one of the most energizing of these struggles. The new national government that emerged has been brilliantly portrayed in LLOYD E. EASTMAN, *The Abortive Revolution: China Under Nationalist Rule, 1927–1937* (1974). Among the interesting rebuttals to some of Eastman's arguments is JOSEPH FEWSMITH, *Party, State, and Local Elites in Republican China: Merchant Organizations and Politics in Shanghai, 1890–1930* (1985).

Events of the war with Japan are discussed in HSI-SHENG CH'I, *Nationalist China at War: Military Defeats and Political Collapse, 1937–45* (1982); and LLOYD E. EASTMAN, *Seeds of Destruction: Nationalist China in War and Revolution, 1937–1949* (1984). Works dealing with wartime conditions in Nationalist areas are CHIA-AO CHANG, *The Inflationary Spiral: The Experience in China, 1939–1950* (1958); and CHARLES F. ROMANUS and RILEY SUNDERLAND, *The United States Army in World War II: China–Burma–India Theatre,* 3 vol. (1952–58), which is carefully documented and well illustrated.

*(Communist China)*: The emergence and growth of the Communist opposition to Nationalist rule can be traced in BENJAMIN I. SCHWARTZ, *Chinese Communism and the Rise of Mao* (1951, reprinted 1979); JOHN E. RUE, *Mao Tse-tung in Opposition, 1927–1935* (1966); JEROME CH'EN, *Mao and the Chinese Revolution* (1965); STUART R. SCHRAM, *Mao Tse-tung* (1967); TRYGVE LÖTVEIT, *Chinese Communism 1931–1934: Experience in Civil Government,* 2nd ed. (1979); and MARK SELDEN, *The Yenan Way in Revolutionary China* (1971). The broadest work treating the history of the Communist movement is CONRAD BRANDT, BENJAMIN I. SCHWARTZ, and JOHN K. FAIRBANK, *A Documentary History of Chinese Communism* (1952, reprinted 1966). A stimulating reportorial work is EDGAR SNOW, *Red Star over China,* rev. ed. (1968, reprinted 1973); while a broad survey is presented by LYMAN P. VAN SLYKE, *Enemies and Friends: The United Front in Chinese Communist History* (1967). The particular role of women and the consequences of the limits placed on their full participation have been the topic of DELIA DAVIN, *Woman-Work: Women and the Party in Revolutionary China* (1976); and JUDITH STACEY, *Patriarchy and Socialist Revolution in China* (1983). The final phase of contest between the two parties is interpreted in SUSANNE PEPPER, *Civil War in China: The Political Struggle, 1945–1949* (1978).

On China's relations with foreign powers, see AKIRA IRIYE, *After Imperialism: The Search for a New Order in the Far East, 1921–1931* (1965, reprinted 1978). O. EDMUND CLUBB, *China and Russia: The "Great Game"* (1971), is a general work on Sino-Soviet relations; while a fine study of American policy toward China during a limited period between the world wars is given in DOROTHY BORG, *The United States and the Far Eastern Crisis of 1933–1938* (1964). See also MICHAEL SCHALLER, *The U.S. Crusade in China, 1938–1945* (1979); and DOROTHY BORG and WALDO HEINRICHS (eds.), *Uncertain Years: Chinese-American Relations, 1947–1950* (1980). F.C. JONES, *Japan's New Order in Asia: Its Rise and Fall, 1937–45* (1954, reprinted 1978); and F.F. LIU, *A Military History of Modern China, 1924–1949* (1956, reprinted 1981), provide information on the military aspects of Sino-Japanese relations. Also see AKIRA IRIYE (ed.), *The Chinese and the Japanese: Essays in Political and Cultural Interactions* (1980). A scholarly study of foreign assistance to China during the Sino-Japanese War is ARTHUR N. YOUNG, *China and the Helping Hand, 1937–1945* (1963). JONATHAN R. ADELMAN, *The Revolutionary Armies: The Historical Development of the Soviet and Chinese People's Liberation Armies* (1980), is a comparative study. Other treatments include WILLIAM C. KIRBY, *Germany and Republican China* (1984); and JAMES REARDON-ANDERSON, *Yenan and the Great Powers: The Origins of Chinese Communist Foreign Policy, 1944–1946* (1980).

For developments in the Communist regions during the Sino-Japanese War, in addition to the general works on the Communist movement cited above, see LYMAN P. VAN SLYKE (ed.), *The Chinese Communist Movement: A Report of the United States War Department, July 1945* (1968), a close factual account based on intelligence studies. CHALMERS A. JOHNSON, *Peasant Nationalism and Communist Power: The Emergence of Revolutionary China* (1962), provides a scholarly interpretation of Communist successes behind Japanese lines. The final civil war period is covered by LIONEL MAX CHASSIN, *The Communist Conquest of China: A History of the Civil War, 1945–1949* (1965; originally published in French, 1952).

The most influential work of Sun Yat-sen has been translated under the title *San Min Chu I: The Three Principles of the People,* trans. by FRANK W. PRICE (1927). The most revealing work of Chiang Kai-shek is his *China's Destiny and Chinese Economic Theory,* published in English with comments by PHILIP JAFFE (1947). The official translation of Mao Zedong's (Mao Tse-tung's) *Selected Works,* 4 vol., is that published by the Foreign Languages Press, Peking (1961–65). STUART R. SCHRAM, *The Political Thought of Mao Tse-tung,* rev. ed. (1969), is the best introduction to Chinese Communist ideology and contains selections from the writings of Mao.

The bibliography on contemporary Chinese history is vast and growing rapidly. Although publications on modern China are uneven in coverage and quality, most subjects of interest have been discussed in the writings of scholars and visitors to China and Taiwan. A comprehensive bibliography of these writings is compiled annually by the *Journal of Asian Studies.* See also GEORGE GINSBURGS and MICHAEL MATHOS, *Communist China and Tibet: The First Dozen Years* (1964); EDGAR SNOW, *Red China Today* (1971); JOHN WILSON LEWIS, *Leadership in Communist China* (1963, reprinted 1978), and (ed.), *The City in Communist China* (1971); BENJAMIN I. SCHWARTZ, *Communism and China: Ideology in Flux* (1968); C.K. YANG, *The Chinese Communist Society: The Family and the Village* (1965); BYUNG-JOON AHN, *Chinese Politics and the Cultural Revolution: Dynamics of Policy Processes* (1976); LOWELL DITTMER, *Liu Shao-Ch'i and the Chinese Cultural Revolution: The Politics of Mass Criticism* (1974); KENNETH LIEBERTHAL, *Revolution and Tradition in Tientsin, 1949–1952* (1980); MAURICE MIESNER, *Mao's China: History of the People's Republic* (1977); WILLIAM E. GRIFFITH, *The Sino-Soviet Rift, Analysed and Documented* (1964), and *Sino-Soviet Relations, 1964–1965* (1967); RICHARD BAUM, *Prelude to Revolution: Mao, the Party, and the Peasant Question, 1962–66* (1975); RODERICK MACFARQUHAR, *The Origins of the Cultural Revolution,* 2 vol. (1974–83); FREDERICK C. TEIWES, *Politics and Purges in China: Rectifications and the Decline of Party Norms, 1950–1965* (1979); TANG TSOU, *The Cultural Revolution and Post-Mao Reforms* (1986); EDWARD E. RICE, *Mao's Way* (1972); ROBERT A. SCALAPINO (ed.), *Elites in the People's Republic of China* (1972); STUART SCHRAM (ed.), *Mao Tse-tung Unrehearsed: Talks and Letters: 1956–71* (1974, U.S. title, *Chairman Mao Talks to the People: Talks and Letters, 1956–1971*); FREDERIC WAKEMAN, JR., *History and Will: Philosophical Perspectives of Mao Tse-tung's Thought* (1973); WILLIAM W. WHITSON and CHEN-HSIA HUANG, *The Chinese High Command: A History of Communist Military Politics, 1927–71* (1973); MARTIN KING WHYTE, *Small Groups and Political Rituals in China* (1974); MARGERY WOLF and ROXANE WITKE (eds.), *Women in Chinese Society* (1975); and MARGERY WOLF, *Revolution Postponed: Women in Contemporary China* (1985).

(C.M.W./E.P.Y./K.G.L.)

**Provinces.** *General works:* For the study of provinces, two map collections are helpful: UNITED STATES. CENTRAL INTELLIGENCE AGENCY, *Communist China Map Folio* (1967), and *Communist China Administrative Atlas* (1969). In addition to materials listed in the geographical section at the beginning of this bibliography, the following studies contain information on the provinces: CHIAO-MIN HSIEH, *China: Ageless Land and Countless People* (1967); LEO J. MOSER, *The Chinese Mosaic: The Peoples and Provinces of China* (1985); GEORGE BABCOCK CRESSEY, *Asia's Lands and Peoples,* 3rd ed. (1963); YI-FU TUAN, *China* (1969); ALBERT HERRMANN, *An Historical Atlas of China,* new ed. (1966); AUREL STEIN, *Innermost Asia,* 5 vol. (1928, reprinted 1981); PING-HAI CHU, *Climate of China,* trans. from the Chinese (1967); JAMES S. LEE, *The Geology of China* (1939); UNITED STATES DEPARTMENT OF THE INTERIOR, *Foreign Minerals Survey: Regional Review,* vol. 2, *Mineral Resources of China* (1948); RONALD HSIA, *Steel in China* (1971); YUAN-LI WU, *The Steel Industry in Communist China* (1965); FREDERICK C. TEIWES, *Provincial Leadership in China: The Cultural Revolution and Its Aftermath* (1974); PARRIS H. CHANG, *Power and Policy in China,* 2nd ed. (1978); and CH'ANG-TU HU, *The Education of National Minorities in Communist China* (1970).

*Northeast and North China:* WASHINGTON (STATE) UNIVERSITY, FAR EASTERN AND RUSSIAN INSTITUTE, *A Regional Handbook on Northeast China* (1956); NAI-RUENN CHEN, *Chinese Economic Statistics: A Handbook for Mainland China* (1967); and DONALD G. GILLIN, *Warlord: Yen Hsi-shan in Shansi Province, 1911–1949* (1967).

*Lower Yangtze Valley:* CHING-CHIH SUN (ed.), *Economic Geography of Central China* (1960), and *Economic Geography of the East China Region* (1961); JONATHAN K. OCKO, *Bureaucratic Reform in Provincial China: Ting Jih-Ch'ang in Restoration Kiangsu, 1867–1870* (1983); ANDREA LEE MCELDERRY, *Shanghai Old-Style Banks (ch'ien-chuang), 1800–1935: A Traditional Institution in a Changing Society* (1976); PARKS M. COBLE, JR., *The Shanghai Capitalists and the Nationalist Government, 1927–1937* (1980); NICHOLAS R. CLIFFORD, *Shanghai, 1925: Urban Nationalism and the Defense of Foreign Privilege* (1979); and "Hubei, Anatomy of a Province," *China Business Review,* 7(5):39–47 (September–October 1980).

*South China:* R. KEITH SCHOPPA, *Chinese Elites and Political Change: Zhejiang Province in the Early Twentieth Century* (1982); KEITH FORSTER, "The Reform of Provincial Party Committees in China: The Case of Zhejiang," *Asian Survey,* 24:618–636 (June 1984); MICHEL OKSENBERG and SAI-CHEUNG YEUNG, "Hua Kuo-feng's Pre-Cultural Revolution Hunan Years, 1949–66," *China Quarterly,* 69:3–53 (March 1977); REWI ALLEY, *Land and Folk in Kiangsi: A Chinese Province in 1961* (1962); DIANA LARY, *Religion and Nation: The Kwangsi Clique in Chinese Politics, 1925–1937* (1974); and EZRA F. VOGEL, *Canton Under Communism: Programs and Politics in a Provincial Capital, 1949–1968* (1969, reprinted 1980).

*Southwest China:* HAN-SHENG CH'EN, *Frontier Land Systems in Southernmost China* (1949); ERNEST HENRY WILSON, *A Naturalist in Western China,* 2 vol. (1913, reprinted 1977); DOROTHY J. SOLINGER, *Regional Government and Political Integration in Southwest China, 1949–1954: A Case Study* (1977); HENRY R. DAVIES, *Yunnan, the Link Between India and the Yangtze* (1909, reissued 1970); and ROBERT A. KAPP, *Szechwan and the Chinese Republic: Provincial Militarism and Central Power, 1911–1938* (1973).

*Western China:* JACK CHEN, *The Sinkiang Story* (1977); DONALD H. MCMILLEN, *Chinese Communist Power and Policy in Xinjiang, 1949–1977* (1979); COLIN MACKERRAS, "Uygur Performing Arts in Contemporary China," *China Quarterly,* 101:58–77 (March 1985); CHARLES BELL, *Tibet Past and Present* (1924, reprinted 1968); TIEH-TSENG LI, *Tibet, Today and Yesterday,* rev. ed. (1960); W.D. SHAKABPA, *Tibet, a Political History* (1967, reissued 1984); DAVID SNELLGROVE and HUGH RICHARDSON, *A Cultural History of Tibet,* rev. ed. (1980); ALASTAIR LAMB, *Britain and Chinese Central Asia: The Road to Lhasa, 1767 to 1905* (1960); GEORGE GINSBURGS and MICHAEL MATHOS, *Communist China and Tibet* (1964); DALAI LAMA (XIV) OF TIBET, *My Land and My People* (1962, reissued 1977); SARAT CHANDRA DAS, *Journey to Lhasa and Central Tibet* (1902, reissued 1970); HUGH RICHARDSON, *Tibet and Its History,* 2nd rev. ed. (1984); and JOHN F. AVEDON, *In Exile from the Land of Snows* (1984).

*Northwest China:* WASHINGTON (STATE) UNIVERSITY. FAR EASTERN AND RUSSIAN INSTITUTE, *A Regional Handbook on Northwest China,* 2 vol. (1956), and *A Regional Handbook on the Inner Mongolia Autonomous Region* (1956); WEN-DJANG CHU, *The Moslem Rebellion in Northwest China, 1862–1878: A Study of Government Minority Policy* (1966); ROBERT J. MILLER, *Monasteries and Culture Change in Inner Mongolia* (1959); OWEN LATTIMORE, *Mongol Journeys* (1941, reprinted 1975), and *Nomads and Commissars: Mongolia Revisited* (1962); and SECHIN JAGCHID and PAUL HYER, *Mongolia's Culture and Society* (1979).

(T.R.T./F.Hu./B.Bo./Jo.E.S./P.-c.K./C.Hu./Y.-G.G.H./N.S.G./C.-y.C./C.-M.H./D.C.T./T.W.D.S./T.V.W./H.E.R./V.C.F.)

# Chinese Literature

Chinese literature is one of the major literary heritages of the world, with an uninterrupted history of more than 3,000 years, dating back at least to the 14th century BC. Its medium, the Chinese language, has retained its unmistakable identity in both its spoken and written aspects in spite of generally gradual changes in pronunciation, the existence of regional and local dialects, and several stages in the structural representation of the written graphs, or "characters." Even the partial or total conquests of China for considerable periods by non-Chinese ethnic groups from outside the Great Wall failed to disrupt this continuity, for the conquerors were forced to adopt the written Chinese language as their official medium of communication because they had none of their own. Since the Chinese graphs were inherently nonphonetic, they were at best unsatisfactory tools for the transcription of a non-Chinese language; and attempts at creating a new alphabetic–phonetic written language for empire building proved unsuccessful on three separate occasions. The result was that after a period of alien domination, the conquerors were culturally assimilated (except the Mongols, who retreated en masse to their original homeland after the collapse of the Yüan [or Mongol] dynasty in 1368). Thus, there was no disruption in China's literary development.

This article is divided into the following sections:

General characteristics 231
History 232
  Origins: *c.* 1400–221 BC 232
  Ch'in and Han dynasties: 221 BC–AD 220 233
  The Six Dynasties and Sui dynasty: AD 220–618 234
  T'ang and Five Dynasties: 618–960 234
  Sung dynasty: 960–1279 235
  Yüan dynasty: 1206–1368 236
  Ming dynasty: 1368–1644 237
  Ch'ing dynasty: 1644–1911/12 237
  Modern Chinese literature 238
Bibliography 240

## GENERAL CHARACTERISTICS

Through cultural contacts, Chinese literature has profoundly influenced the literary traditions of other Asian countries, particularly Korea, Japan, and Vietnam. Not only was the Chinese script adopted for the written language in these countries but some writers adopted the Chinese language as their chief literary medium.

The graphic nature of the written aspect of the Chinese language has produced a number of noteworthy effects upon Chinese literature and its diffusion: (1) Chinese literature, especially poetry, is recorded in handwriting or in print and purports to make an aesthetic appeal to the reader that is visual as well as aural. (2) This visual appeal of the graphs has in fact given rise to the elevated status of calligraphy in China, where it has been regarded for at least the last 16 centuries as a fine art comparable to painting. Scrolls of calligraphic renderings of poems and prose selections have continued to be hung alongside paintings in the homes of the common people as well as the elite, converting these literary gems into something to be enjoyed in everyday living. (3) On the negative side, such a writing system has been an impediment to education and the spread of literacy, thus reducing the number of readers of literature; for even a rudimentary level of reading and writing requires knowledge of more than 1,000 graphs, together with their pronunciation. (4) On the other hand, the Chinese written language, even with its obvious disadvantages, has been a potent factor in perpetuating the cultural unity of the growing millions of the Chinese people, including assimilated groups in far-flung peripheral areas. Different in function from recording words in an alphabetic–phonetic language, the graphs are not primarily indicators of sounds and can therefore be pronounced in variant ways to accommodate geographical diversities in speech and historical phonological changes without damage to the meaning of the written page. As a result, the major dialects in China never developed into separate written languages as did the Romance languages, and, although the reader of a Confucian Classic in southern China might not understand the everyday speech of someone from the far north, Chinese literature has continued to be the common asset of the whole Chinese people. By the same token, the graphs of China could be utilized by speakers of other languages as their literary mediums.

The pronunciation of the Chinese graphs has also influenced the development of Chinese literature. The fact that each graph had a monophonic pronunciation in a given context created a large number of homonyms, which led to misunderstanding and confusion when spoken or read aloud without the aid of the graphs. One corrective was the introduction of tones or pitches in pronunciation. As a result, metre in Chinese prosody is not concerned with the combination of syllabic stresses, as in English, but with those of syllabic tones, which produce a different but equally pleasing cadence. This tonal feature of the Chinese language has brought about an intimate relationship between poetry and music in China. All major types of Chinese poetry were originally sung to the accompaniment of music. Even after the musical scores were lost, the poems were, as they still are, more often chanted—in order to approximate singing—than merely read.

Relationship between poetry and music

Chinese poetry, besides depending on end rhyme and tonal metre for its cadence, is characterized by its compactness and brevity. There are no epics of either folk or literary variety and hardly any narrative or descriptive poems that are long by the standards of world literature. Stressing the lyrical, as has often been pointed out, the Chinese poet refrains from being exhaustive, marking instead the heights of his ecstasies and inspiration or the depths of sorrow and sympathy. A short poem in Chinese sometimes resembles a cablegram, wherein verbal economy is highly desirable. Generally, pronouns and conjunctions are omitted, and one or two words often allude to highly complex thoughts or situations. This explains why many poems have been differently interpreted by learned commentators and competent translators.

The line of demarcation between prose and poetry is much less distinctly drawn in Chinese literature than in other national literatures. This is clearly reflected in three genres. The *fu,* for example, is on the borderline between poetry and prose, containing elements of both. It uses rhyme and metre and not infrequently also antithetic structure, but, despite occasional flights into the realm of the poetic, it retains the features of prose without being necessarily prosaic. This accounts for the variety of labels given to the *fu* in English by writers on Chinese literature—poetic prose, rhyme prose, prose poem, rhapsody, and prose poetry.

Another genre belonging to this category is *p'ien-wen* ("parallel prose"), characterized by antithetic construction and balanced tonal patterns without the use of rhyme; the term is suggestive of "a team of paired horses," as is implied in the Chinese word *p'ien.* Despite the polyphonic effect thus produced, which approximates that of poetry, it has often been made the vehicle of proselike exposition and argumentation. Another genre, a peculiar mutation in this borderland, is the *pa-ku wen-chang* ("eight-legged essay"). Now generally regarded as unworthy of classification as literature, for centuries (from 1487 to 1901) it dominated the field of Chinese writing as the principal

yardstick in grading candidates in the official civil-service examinations. It exploited antithetical construction and contrasting tonal patterns to the limit by requiring pairs of columns consisting of long paragraphs, one responding to the other, word for word, phrase for phrase, sentence for sentence.

The two kinds of prose

Chinese prose writing has been diverted into two streams, separated at least for the last 1,000 years by a gap much wider than the one between folk songs and so-called literary poems. Classical, or literary, prose (*ku-wen,* or *wen-yen*) aims at the standards and styles set by ancient writers and their distinguished followers of subsequent ages, with the Confucian Classics and the early philosophers as supreme models. While the styles may vary with individual writers, the language is always far removed from their spoken tongues. Sanctioned by official requirement for the competitive examinations and dignified by traditional respect for the cultural accomplishments of past ages, this medium became the linguistic tool of practically all Chinese prose writers. Vernacular prose (*pai-hua*), in contrast, consists of writings in the living tongue, the everyday language of the authors. Traditionally considered inferior, the medium was piously avoided for creative writing until it was adopted by novelists and playwrights from the 13th century on.

HISTORY

**Origins: c. 1400–221 BC.** The oldest specimens of Chinese writing extant are inscriptions on bones and tortoise shells dating back to the last three centuries of the Shang dynasty (18th–12th centuries BC) and recording divinations performed at the royal capital. These inscriptions, like those engraved on ceremonial bronze vessels toward the end of the Shang period, are usually brief and factual and cannot be considered literature. Nonetheless, they are significant in that their sizable vocabulary (about 3,400 characters, of which nearly 2,000 have been reliably deciphered) has proved to be the direct ancestor of the modern Chinese script. Moreover, the syntactical structure of the language bears a striking resemblance to later usages. From the frequent occurrences in the bone inscriptions of such characters as "dance" and "music," "drum" and "chimes" (of stone), "words" and "southern" (airs), it can safely be inferred that, by the Shang dynasty, songs were sung to the accompaniment of dance and music; but these songs are now lost. (T.-y.L./W.H.N.)

*Literary use of myths.* Early Chinese literature does not present, as the literatures of certain other world cultures do, great epics embodying mythological lore. What information exists is sketchy and fragmentary and provides no clear evidence that an organic mythology ever existed; if it did, all traces have been lost. Attempts by scholars, Eastern and Western alike, to reconstruct the mythology of antiquity have consequently not advanced beyond probable theses. Shang dynasty material is limited. Chou dynasty (*c.* 1111–255 BC) sources are more plentiful, but even these must at times be supplemented by writings of the Han period (206 BC–AD 220), which, however, must be read with great caution. This is the case because Han scholars reworked the ancient texts to such an extent that no one is quite sure, aside from evident forgeries, how much was deliberately reinterpreted and how much was changed in good faith in an attempt to clarify ambiguities or reconcile contradictions.

The early state of Chinese mythology was also molded by the religious situation that prevailed in China at least since the Chou conquest (12th century BC), when religious observance connected with the cult of the dominant deities was proclaimed a royal prerogative. Because of his temporal position, the king alone was considered qualified to offer sacrifice and to pray to these deities. Shang-ti ("Supreme Ruler"), for example, one of the prime dispensers of change and fate, was inaccessible to persons of lower rank. The princes, the aristocracy, and the commoners were thus compelled, in descending order, to worship lesser gods and ancestors. Though this situation was greatly modified about the time of Confucius in the early part of the 5th century BC, institutional inertia and a trend toward rationalism precluded the revival of a mythological world. Confucius prayed to Heaven (T'ien) and was concerned about the great sacrifices, but he and his school had little use for genuine myths.

Nevertheless, during the latter centuries of the Chou, Chinese mythology began to undergo a profound transformation. The old gods, to a great extent already forgotten, were gradually supplanted by a multitude of new ones, some of whom were imported from India with Buddhism or gained popular acceptance as Taoism spread throughout the empire. In the process, many early myths were totally reinterpreted to the extent that some deities and mythological figures were rationalized into abstract concepts and others were euhemerized into historical figures. Above all, a hierarchical order, resembling in many ways the institutional order of the empire, was imposed upon the world of the supernatural. Many of the archaic myths were lost; others survived only as fragments, and, in effect, an entirely new mythological world was created.

These new gods generally had clearly defined functions and definite personal characteristics and became prominent in literature and the other arts. The myth of the battles between Huang-ti ("The Yellow Emperor") and Ch'ih Yu ("The Wormy Transgressor"), for example, became a part of Taoist lore and eventually provided models for chapters of two works of vernacular fiction, *Shui-hu chuan* (*The Water Margin,* also translated as *All Men Are Brothers*) and *Hsi-yu chi* (1592; *Journey to the West,* also partially translated as *Monkey*). Other mythological figures such as K'ua-fu and the Hsi-wang-mu subsequently provided motifs for numerous poems and stories.

Historical personages were also commonly taken into the pantheon, for Chinese popular imagination has been quick to endow the biography of a beloved hero with legendary and eventually mythological traits. Ch'ü Yüan, the ill-fated minister of the state of Ch'u (771–221 BC), is the most notable example. Mythmaking consequently became a constant, living process in China. It was also true that historical heroes and would-be heroes arranged their biographies in a way that lent themselves to mythologizing. (He.W./W.H.N.)

First anthology of poetry: *Shih Ching*

*Poetry.* The first anthology of Chinese poetry, known as the *Shih Ching* ("Classic of Poetry") and consisting of temple, court, and folk songs, was given definitive form somewhere around the time of Confucius (551–479 BC). But its 305 songs are believed to range in date from the beginning of the Chou dynasty to the time of their compiling.

The *Shih Ching* is generally accounted the third of the Five Classics (*Wu Ching*) of Confucian literature, the other four of which are: the *I Ching* ("Classic of Changes"), a book of divination and cosmology; the *Shu Ching* ("Classic of History"), a collection of official documents; the *Li chi* ("Record of Rites"), a book of rituals with accompanying anecdotes; and the *Ch'un-ch'iu* ("Spring and Autumn") annals, a chronological history of the feudal state of Lu, where Confucius was born, consisting of topical entries of major events from 722 to 481 BC. The Five Classics have been held in high esteem by Chinese scholars since the 2nd century BC. (For a discussion of the *I Ching* and *Shu Ching,* see below *Prose.*)

The poems of the *Shih Ching* were originally sung to the accompaniment of music; and some of them, especially temple songs, were accompanied also by dancing. (In all subsequent periods of Chinese literary history, new trends in poetry were profoundly influenced by music.) Most of the poems of the *Shih Ching* have a preponderantly lyrical strain whether the subject is hardship in military service or seasonal festivities, agricultural chores or rural scenes, love or sports, aspirations or disappointments of the common folk and of the declining aristocracy. Apparently, the language of the poems was relatively close to the daily speech of the common people, and even repeated attempts at refinement during the long process of transmission have not spoiled their freshness and spontaneity. In spite of this, however, when the songs are read aloud and not sung to music their prevailing four-syllable lines conduce to monotony, hardly redeemed by the occasional interspersion of shorter or longer lines.

If there ever was an epic tradition in ancient China

comparable to that of early India or the West, only dim traces of it persist in the written records. The *Shih Ching* has a few narrative poems celebrating heroic deeds of the royal ancestors, but these are rearranged in cycles and only faintly approximate the national epics of other peoples. One cycle, for example, records the major stages in the rise of the Chou kingdom, from the supernatural birth of its remote founder to its conquest of the Shang kingdom. These episodes, which, according to traditional history, cover a period of more than 1,000 years, are dealt with in only about 400 lines. Other cycles, which celebrate later military exploits of the royal Chou armies, are even briefer.

The *Shih Ching* exerted a profound influence on Chinese poetry that, generally speaking, has stressed the lyrical rather than the narrative element; a dependence more on end rhymes for musical effect than on other rhetorical devices; regular lines, consisting of a standard number of syllables; and the utilization of intonation that is inherent in the language for rhythm, instead of the alternation of stressed and unstressed syllables as is the norm in Western poetry. The high regard in which this anthology has been held in China results both from its antiquity and from the legend that Confucius himself edited it. It was elevated in 136 BC to the position of a major classic in the Confucian canon.

Meanwhile, another type of poetry, also originating in music and dance, had developed in the south, in the basin of the Yangtze River, an area dominated by the principality of Ch'u—hence the generic appellation *Ch'u tz'u,* or "songs of Ch'u." These southern songs, though adorned with end rhymes like the songs of the *Shih Ching,* follow a different metrical pattern: the lines are usually longer and more irregular and are commonly (though not always) marked by a strong caesura in the middle. Their effect is thus rather plaintive, and they lend themselves to chanting instead of singing. The beginning of this tradition is obscure because most of the early samples were eclipsed by the brilliant 4th/3rd-century-BC compositions of the towering genius Ch'ü Yüan, China's first known poet.

*Ch'u tz'u*

Among some 25 elegies that are attributed to Ch'ü Yüan, the most important and longest is *Li sao* ("On Encountering Sorrow"), which has been described as a politico-erotic ode, relating by means of a love allegory the poet's disappointment with his royal master and describing his imaginary travels in distant regions and the realms of heaven, in an attempt to rid himself of his sorrow. Ch'ü Yüan committed suicide by drowning in the Mi-lo River; and his tragic death, no less than his beautiful elegies, helped to perpetuate the new literary genre. In contrast to the poems of the *Shih Ching,* which had few successful imitators, the genre created by Ch'ü Yüan was cultivated for more than five centuries, and it also experienced later revivals.

*Prose.* Prior to the rise of the philosophers in the 6th century BC, brief prose writings were reported to be numerous; but of these only two collections have been transmitted: the *Shu,* or *Shu Ching* ("Classic of History"), consisting of diverse kinds of primitive state papers, such as declarations, portions of charges to feudal lords, and orations; and the *I,* or *I Ching* ("Classic of Changes"), a fortune-telling manual. Both grew by accretion and, according to a very doubtful tradition, were edited by Confucius himself. Neither can be considered literature, but both have exerted influence on Chinese writers for more than 2,000 years as a result of their inclusion in the Confucian canon.

The earliest writings that can be assigned to individual "authorship," in the loose sense of the term, are the *Lao-tzu,* or *Tao-te Ching* ("Classic of the Way of Power"), which is attributed to Lao-tzu, who is credited with being the founder of Taoism and who might have been an older contemporary of Confucius; and the *Lun yü* ("Conversations"), or *Analects* (selected miscellaneous passages), of Confucius. Neither of the philosophers wrote extensively, and their teachings were recorded by their followers. Thus, the *Lao-tzu* consists of brief summaries of Lao-tzu's sayings, many of which are in rhyme and others in polished prose to facilitate memorization. Likewise, the *Analects* is composed of collections of the sage's sayings, mostly as answers to questions or as a result of discussions because writing implements and materials were expensive and scarce. The circumstances of the conversations, however, were usually omitted; and as a consequence the master's words often sound cryptic and disjointed, despite the profundity of the wisdom.

Earliest writings claiming individual authorship

By about 400 BC, writing materials had improved, and a change in prose style resulted. The records of the discourses became longer, the narrative portions more detailed; jokes, stories, anecdotes, and parables, interspersed in the conversations, were included. Thus, the *Mencius,* or *Meng-tzu,* the teachings of Mencius, not only is three times longer than the *Analects* of Confucius but also is topically and more coherently arranged. The same characteristic may be noticed in the authentic chapters of the *Chuang-tzu,* attributed to the Taoist sage Chuang-tzu, who "in paradoxical language, in bold words, and with subtle profundity, gave free play to his imagination and thought. . . . Although his writings are inimitable and unique, they seem circuitous and innocuous. Although his utterances are irregular and formless, they are unconventional and readable . . ." (from the epilogue of the *Chuang-tzu*).

The first example of the well-developed essay, however, is found neither in the *Mencius* nor in the *Chuang-tzu* but in the *Mo-tzu,* attributed to Mo Ti, or Mo-tzu, a predecessor of Mencius and Chuang-tzu, whose singular attainments in logic made him a forceful preacher. His recorded sermons are characterized by simplicity of style, clarity of exposition, depth of conviction, and directness of appeal.

Well-developed essay

The prose style continued to be developed by such outstanding philosopher-essayists as Hsün-tzu and his pupil, the Legalist Han-fei-tzu. The peak of this development, however, was not reached until the appearance of the first expertly arranged full-length book, *Lü-shih Ch'un-ch'iu* ("The Spring and Autumn [Annals] of Mr. Lü"), completed in 240 BC under the general direction of Lü Pu-wei. The work, 60 essays in 26 sections, summarizes the teachings of the several schools of philosophy as well as the folklore of the various regions of China.

**Ch'in and Han dynasties: 221 BC–AD 220.** *Poetry.* Following the unification of the empire by the Ch'in dynasty (221–206 BC) and the continuation of the unified empire under the Han, literary activities took new directions. At the Imperial and feudal courts, the *fu* genre, a combination of rhyme and prose, began to flourish. Long and elaborate descriptive poetic compositions, the *fu* were in form a continuation of the Ch'u elegies, now made to serve a different purpose—the amusement of the new aristocracy and the glorification of the empire—by dwelling on such topics as the low table and the folding screen or on descriptions of the capital cities. But even the best *fu* writing, by such masters of the art as Mei Sheng and Ssu-ma Hsiang-ju, bordered on the frivolous and bombastic. Another major *fu* writer, Yang Hsiung, in the prime of his career remorsefully realized that the genre was a minor craft not worthy of a true poet. Nonetheless, the *fu* was almost universally accepted as the norm of creative writing, and nearly 1,000 pieces were produced.

The *fu* genre

A more important contribution to literature by the Han government was the reactivation in 125 BC of the Yüeh Fu, or Music Bureau, which had been established at least a century earlier to collect songs and their musical scores. Besides temple and court compositions of ceremonial verse, this office succeeded in preserving a number of songs sung or chanted by the ordinary people, including songs from the border areas, which reveal alien influences. This category—called *yüeh-fu,* for the Music Bureau—includes not only touching lyrics but also charming ballads.

Ballads

One such ballad, "The Orphan," tells of an orphan's hardships and disappointments; the form of the poem—lines of irregular length, varying from three to six syllables (or graphs)—represents the singer's attempt to simulate the choking voice of the sufferers. *Lo-fu hsing* ("The Song of Lo-fu"; also called *Mo-shang sang,* "Roadside Mulberry Tree"), recounts how a pretty young lady declined a carriage ride offered her by a government commissioner. The most outstanding folk ballad of this period is *K'ung-ch'üeh tung-nan fei* ("Southeast the Peacock Flies"). The longest

poem of early Chinese literature (353 lines), it relates the tragedy of a young married couple who had committed suicide as the result of the cruelty of the husband's mother. The ballad was probably first sung shortly after AD 200 and grew by accretion and refinement in oral transmission until it was recorded in final form for the first time in about 550. *Yüeh-fu* songs, most of which are made up mainly of five-syllable lines, became the fountainhead of a new type of poetry, *ku-shih* ("ancient-style poems"); contemporary Han dynasty poets at first merely refined the originals of the folk songs without claiming credit and later imitated their fresh and lively metre.

*Prose.* Prose literature was further developed during the Ch'in and Han dynasties. In addition to a prolific output of philosophers and political thinkers—a brilliant representative of whom is Liu An, prince of Huai-nan, whose work is called *Huai-nan-tzu* (*c.* 140 BC; "The Master of Huai-nan")—an important and monumental category of Han dynasty literature consists of historical works. Outstanding among these is the *Shih-chi* (*c.* 85 BC; "Historical Records," Eng. trans., *The Records of the Grand Historian of China,* 2 vol.) by Ssu-ma Ch'ien. A masterpiece that took 18 years to produce, it deals with major events and personalities of about 2,000 years (down to the author's time), comprising 130 chapters and totaling more than 520,000 words. The *Shih-chi* was not only the first general history of its kind attempted in China, it also set a pattern in organization for dynastic histories of subsequent ages. An artist as well as a historian, Ssu-ma Ch'ien succeeded in making events and personalities of the past into living realities for his readers; his biographies subsequently became models for authors of both fiction and history. Ssu-ma's great successor, the poet-historian-soldier Pan Ku, author of the *Han shu* ("Han Documents"), a history of the Former Han dynasty containing more than 800,000 words, performed a similar tour de force but did not equal Ssu-ma Ch'ien in either scope or style.

Pan Ku's prose style, though not necessarily archaic, was more consciously literary—a result of the ever-widening gap between the spoken and written aspects of the language. This anomaly was more evident in China than elsewhere, and it was to have far-reaching effects on the evolution of Chinese literary tradition. In an attempt to resolve the difficulties of communication among speakers of many dialects in the empire, a standard literary language, *wen-yen,* was promoted from the Han dynasty on. Perpetuated for more than 2,000 years, the literary language failed to keep pace with changes in the spoken tongue, and eventually it became almost unintelligible to the illiterate masses.

Promotion of a standard literary language

**The Six Dynasties and Sui dynasty: AD 220–618.** After the fall of the Han dynasty, there was a long period of political division (AD 220–589), with barely four decades of precarious unification (AD 280–316/17). Despite the social and political confusion and military losses, however, the cultural scene was by no means dismal. Several influences on the development of literature are noteworthy. First, Buddhism, introduced earlier, had brought with it religious chants and Indian music, which helped to attune Chinese ears to the finer distinctions of tonal qualities in their own language. Second, aggressive northern tribes, who invaded and dominated the northern half of the country from 316, were being culturally absorbed and converted. Third, the political division of the empire between the South and the North (as a result of the domination of non-Chinese in the north) led to an increase in cultural differences and to a subsequent rivalry to uphold what was regarded as cultural orthodoxy, frequently resulting in literary antiquarianism.

*Poetry.* Folk songs flourished in both regions. In the South, popular love songs, originating in the coastal areas, which now came increasingly under Chinese political and cultural domination, attracted the attention of poets and critics. The songs of the North were more militant. Reflecting this spirit most fully is the *Mu-lan shih* ("Ballad of Mu Lan"), which sings of a girl who disguised herself as a warrior and won glory on the battlefield.

Ts'ao Chih and T'ao Ch'ien

Soon the number of writers of "literary" poetry greatly increased. Among them, two poets deserve special mention. Ts'ao Chih (3rd century), noted for his ethereal lyricism, gave definite artistic form to the poetry of the five-syllable line, already popularized in folk song. T'ao Ch'ien (4th–5th centuries), also known as T'ao Yüan-ming, is one of China's major poets and was the greatest of this period. A recluse, he retired from a post in the bureaucracy of the Chin dynasty at the age of 33 to farm, contemplate nature, and write poetry. His verse, written in a plain style, was echoed by many poets who came after him. Using several verse forms with seemingly effortless ease—including the *fu,* for *Kuei-ch'ü-lai tz'u* ("Homeward Bound")—he was representative of the trend of the age to explore various genres for lyrical expression. One of his best loved poems is the following *ku-shih,* translated by Arthur Waley; it is one of 12 he wrote at different times after he had been drinking.

I built my hut in a zone of human habitation,
Yet near me there sounds no noise of horse or coach.
Would you know how this is possible?
A heart that is distant creates a wilderness round it.
I pluck chrysanthemums under the eastern hedge,
Then gaze long at the distant hills.
The mountain air is fresh at the dusk of day;
The flying birds two by two return.
In these things there lies a deep meaning;
Yet when we would express it, words suddenly fail us.

*Prose.* As orthodox Confucianism gradually yielded to Taoism and later to Buddhism, nearly all of the major writers began to cultivate an uninhibited individuality. Lu Chi, 3rd-century poet and critic, in particular emphasized the importance of originality in creative writing and discredited the long-established practice of imitating the great masters of the past. Still, his celebrated essay on literature (*Wen fu*), in which he enunciated this principle, was written as a *fu,* showing after all that he was a child of his own age. The 3rd/4th-century Taoist philosopher Ko Hung insisted that technique is no less essential to a writer than moral integrity. The revolt of the age against conventionality was revealed in the new vogue of *ch'ing-t'an* ("pure conversation"), intellectual discussions on lofty and nonmundane matters, recorded in a 5th-century collection of anecdotes entitled *Shih-shuo hsin-yü* ("A New Account of Tales of the World") by Liu Yi-ch'ing. Though prose writers as a whole continued to be most concerned with lyrical expression and rhetorical devices for artistic effect, there were notable deviations from the prevailing usage in the polyphonic *p'ien-wen* ("parallel prose"). In this form, parallel construction of pairs of sentences and counterbalancing of tonal patterns were the chief requirements. *P'ien-wen* was used especially in works concerned with philosophical disputes and in religious controversies; but it was also used in the first book-length work of literary criticism, *Wen-hsin tiao-lung* ("The Literary Mind and the Carving of the Dragon"), by the 6th-century writer Liu Hsieh.

Development of polyphonic prose: *p'ien-wen*

Among prose masters of the 6th century, two northerners deserve special mention: Yang Hsien-chih, author of *Lo-yang Chia-lan chi* ("Record of Buddhist Temples in Lo-yang"), and Li Tao-yüan, author of *Shui Ching chu* ("Commentary on the Water Classic"). Although both of these works seem to have been planned to serve a practical, utilitarian purpose, they are magnificent records of contemporary developments and charming storehouses of accumulated folklore, written with great spontaneity and artistry. This age also witnessed the first impact of Buddhist literature in Chinese translation, which had been growing in size and variety since the 2nd century.

**T'ang and Five Dynasties: 618–960.** During the T'ang dynasty (618–907), Chinese literature reached its golden age.

*Poetry.* In poetry, the greatest glory of the period, all the verse forms of the past were freely adopted and refined, and new forms were crystallized. One new form was perfected early in the dynasty and given the definitive name *lü-shih* ("regulated verse"). A poem of this kind consists of eight lines of five or seven syllables—each line set down in accordance with strict tonal patterns—calling for parallel structure in the middle, or second and third, couplets.

Another verse form much in vogue was the *chüeh-chü* ("truncated verse"). An outgrowth and a shortened ver-

Chüeh-chü verse form

sion of the *lü-shih,* it omitted either the first four lines, the last four lines, the first two and the last two lines, or the middle four lines. Thus, the tonal quality of the *lü-shih* was retained, whereas antithetic structure was made optional. These poems of four lines, each consisting of five or seven words (syllables or characters), had to depend for their artistry on suggestiveness and economy comparable to the *robāʿīyāt* ("quatrains") of Omar Khayyam and the Japanese haiku.

The fine distinctions of tonal variations in the spoken language had reached their height during this period, with eight tones; and rules and regulations concerning the sequence of lighter and heavier tones had been formulated. But since the observance of strict rules of prosody was not mandatory in the *ku-shih* ("ancient style") form still in use, it was possible for an individual poet to enjoy conformity or freedom as he saw fit.

Of the more than 2,200 T'ang poets whose works—totaling more than 48,900 pieces—have been preserved, only a few can be mentioned. Wang Wei, a musician and the traditional father of monochrome landscape painting, was also a great poet. Influenced by Buddhism, he wrote exquisite meditative verse of man's relation to nature that exemplified his own dictum that poetry should have the beauty of painting and vice versa. Li Po, one of the two major poets of the T'ang dynasty, a lover of detachment and freedom, deliberately avoided the *lü-shih* and chose the less formal verse forms to sing of friendship or wine. An example is the poem "To Tan-Ch'iu," translated by Arthur Waley.

My friend is lodging high in the Eastern Range,
Dearly loving the beauty of valleys and hills.
At green Spring he lies in the empty woods,
And is still asleep when the sun shines on high.
A pine-tree wind dusts his sleeves and coat;
A pebbly stream cleans his heart and ears.
I envy you, who far from strife and talk
Are high-propped on a pillow of blue cloud.

The works of Tu Fu

Generally considered the greatest poet of China was Tu Fu, a keen observer of the political and social scene who criticized injustice wherever he found it and who clearly understood the nature of the great upheaval following the rebellion of dissatisfied generals in 755, which was a turning point in the fortunes of the T'ang. As an artist, Tu Fu excelled in all verse forms, transcending all rules and regulations in prosody while conforming to and exploiting them. His power and passion can perhaps be suggested by a single line (translated by Robert Payne): "Blue is the smoke of war, white the bones of men."

One of the admirers of Tu Fu as a poet-historian was Po Chü-i who, like his great predecessor, was deeply concerned with the social problems of his age. Po Chü-i sought to learn from ordinary folk not only naturalness of language but also their feelings and reactions, especially at the height of his career when he wrote what he called the *Hsin yüeh-fu shih* ("New Yüeh-fu Poems").

Development of *tz'u*

At the end of the T'ang and during the Five Dynasties, another new verse form developed. Composed normally of lines of irregular length and written as lyrics to musical tunes, this form came to be known as *tz'u,* in contrast with *shih,* which includes all the verse forms mentioned above. Since the lines in a *tz'u* might vary from one to nine or even 11 syllables, they were comparable to the natural rhythm of speech and therefore easily understood when sung.

First sung by ordinary folk, they were popularized by professional women singers and, during the T'ang, attracted the attention of poets. It was not, however, until the transitional period of the Five Dynasties (907–960), a time of division and strife, that *tz'u* became the major vehicle of lyrical expression. Of *tz'u* poets in this period, the greatest was Li Yü, last monarch of the Southern T'ang, who was seized in 976 as the new Sung dynasty consolidated its power. Li Yü's *tz'u* poetry is saturated with a tragic nostalgia for better days in the South; it is suffused with sadness—a new depth of feeling notably absent from earlier *tz'u,* which had been sung at parties and banquets. The following is typical, translated by Jerome Ch'en and Michael Bullock:

*Lin hua hsieh liao ch'un hung*
*T'ai ch'ung ch'ung*
*Wu nai chao lai han yü wan lai feng*
*Yen chih lei*
*Hsiang liu tsui*
*Chi shih ch'ung*
*Tzu shih jen sheng ch'ang hen shui ch'ang tung*
The red of the spring orchard has faded.
Far too soon!
The blame is often laid
on the chilling rain at dawn
and the wind at dusk.
The rouged tears
That intoxicate and hold in thrall—
When will they fall again?
As a river drifts toward the east
So painful life passes to its bitter end.

*Folk literature.* Besides the early *tz'u,* the end of the T'ang saw the evolution of another new folk form: *pien-wen* ("popularizations," not to be confused with *p'ien-wen,* or parallel prose), utilizing both prose and verse to retell episodes from the Buddha's life and, later, non-Buddhist stories from Chinese history and folklore.

Han Yü's reform of prose

*Prose.* In prose writing a major reform was led by Han Yü against the peculiarly artificial prose style of *p'ien-wen,* which, cultivated for almost 1,000 years, had become so burdened with restrictive rules as to make forthright expression virtually impossible. Han Yü boldly advocated the use of Chou philosophers and early Han writers as models for prose writing. This seemingly conservative reform had, in fact, a liberalizing effect; for the sentence unit in prose writing was now given perfect freedom to seek its own length and structural pattern as logic and content might dictate, instead of slavishly conforming to the rules of *p'ien-wen.* This new freedom enabled Liu Tsung-yüan, Han Yü's chief associate in the literary reform, to write charming travel and landscape pieces. It also accelerated the development of a new genre in prose: well-made tales of love and romance, of heroic feats and adventures, of the mysterious and supernatural, and of imaginary incidents and fictionalized history. Among the 9th-century writers of such prose romances were Han Yü's pupil Shen Ya-chih and Po Hsing-chien, younger brother of the poet Po Chü-i. These prose romances, generally short, were written in the classical prose style for the amusement of the literati and did not reach the masses until some of the popular ones were adapted by playwrights in later ages.

**Sung dynasty: 960–1279.** The Sung dynasty was marked by cultural advancement and military weakness. During this period, literary output was spectacularly increased, thanks mainly to the improvement of printing (invented in the 8th century) and to the establishment of public schools throughout the empire (from 1044). Nearly all the literary genres in verse and prose were continued; and some trends, begun in T'ang times, were accelerated.

*Prose.* In prose, the reform initiated by Han Yü in the name of ancient, more straightforward style (*ku-wen*) was reemphasized by such 11th-century writers as Ou-yang Hsiu and Su Tung-p'o. Both men held high rank in the civil service and were great painters as well as leading poets. Nevertheless, their contribution to prose writing in *ku-wen* style was as important as their poetry. The *ku-wen* movement was further supported by men whose primary interest was not belles lettres, such as Ssu-ma Kuang, the statesman-historian, and Chu Hsi, the scholar-philosopher and principal formulator of Neo-Confucianism.

Sung trends in prose fiction

In prose fiction there were two distinct trends. Short tales in *ku-wen* were written in ever greater bulk but failed to maintain the level achieved in the T'ang dynasty. The subject matter became more fragmentary and anecdotal and the style duller. In sharp contrast to the *ku-wen* school, which was still a literary language despite the movement toward naturalness of expression, there arose a school of storytelling in the vernacular. Almost purely oral in origin, these tales reflected the style of the storyteller who entertained audiences gathered in marketplaces, fairgrounds, or temple yards. In the 12th century they became fairly lengthy, connected stories, especially those dealing with fictionalized history. This elevation of the everyday speech of the common people as a medium of story writing of

the *hua-pen* ("vernacular story") type was to open up new vistas in prose fiction in later periods.

*Poetry.* Poetry of the conventional type (*shih*) was cultivated by numerous rival schools, each claiming many illustrious members. On the whole, the rival literary movements were significant as steps toward greater naturalness in syntax, and a few outstanding writers approximated the spoken vernacular language. Among the many *shih* poets of the Sung dynasty, Lu Yu, who flourished in the 12th century, was a towering figure. A traveler and patriot, he wrote throughout his long career no fewer than 20,000 poems, of which more than 9,000 have been preserved.

But it was in their utilization of the newer verse form, *tz'u,* that Sung poets achieved their greatest distinction, making *tz'u* the major genre of the dynasty. As noted above, the *tz'u* form had been popularized at first orally by women singers; and the first generation of *tz'u* writers had been inspired and guided by them in sentiment, theme, and diction; their lyrics were thus redolent with the fragrance of these women. Later in the 12th century, as men (and one great woman) of letters began to take over, the *tz'u* form reached the heights of great art. Ou-yang Hsiu and Li Ch'ing-chao, the latter generally considered the greatest woman poet of China, may be considered representatives of this trend. Li Ch'ing-chao's poems, paralleling her life, are intensely personal. They at first dealt with the joys of love, but gradually their tone darkened to one of despair, caused first by frequent and lengthy separations from her husband, who was in government service, and then by his untimely death.

Other masters of the *tz'u* were Su Tung-p'o and Hsin Ch'i-chi, the latter a soldier turned recluse. It was Hsin Ch'i-chi who imbued the writing of *tz'u* with new characteristics by rising above rules without breaking them, surpassing in this respect his contemporaries as well as those who came after him.

**Yüan dynasty: 1206–1368.** Fleeing from the Chin (Juchen) Tatars, who captured their capital in 1127, the Sung officials and courtiers retreated southward. For almost a century and a half, China was again divided. And in spite of political reunification by Kublai Khan, founder of the Yüan, or Mongol, dynasty (beginning in 1206 in the North and comprising the whole of China by 1280), the cultural split persisted. In the South, where China's historic traditions found asylum, racial and cultural homogeneity persisted. In fact, the centre of Chinese philosophy and traditional literature never again returned north of the Yangtze Delta. But in the North new developments arose, which led to wholly new departures. First, the migration and fusion of the various ethnic groups gave birth to a common spoken language with fewer tones, which later was to become the basis of a national language; second, with the southward shift of the centre of traditional culture, the prestige of the old literature began to decline in the North, especially in the eyes of the conquerors. Thus, in contrast to the South, North China under the Yüan dynasty provided a unique milieu for unconventional literary activities.

*Drama.* In this period, dramatic literature came into a belated full flowering. The skits and vaudeville acts, the puppet shows and shadow plays of previous ages had laid the foundation for a full-fledged drama; but the availability of Indian and Iranian models during the Yüan dynasty may have been a more immediate cause for its accelerated growth. Many Chinese men of letters refused to cooperate with the alien government, seeking refuge in painting and writing. As the new literary type developed—the drama of four or five acts, complete with prologue and epilogue and including songs and dialogue in language fairly close to the daily speech of the people—many men of letters turned to playwriting. Between 1234 and 1368, more than 1,700 musical plays were written and staged, and 105 dramatists were recorded; moreover, there is an undetermined number of anonymous playwrights whose unsigned works have been preserved but discovered only in the 20th century. This remarkable burst of literary innovation, however, failed to win the respect of the orthodox critics and official historians. No mention of it was made in the copious dynastic history, *Yüan shih;* and casual references in the collected works of contemporary writers were few. Many plays were allowed to fall into oblivion. It was not until 1615 that a bibliophile undertook to reprint, as a collection, 100 of the 200 plays he had seen. Even after ardent searches by 20th-century librarians and specialists, the number of extant Yüan dramas has been increased to only 167, hardly 10 percent of the number produced. Moreover, since the musical scores have been lost, the plays cannot be produced on the stage in the original manner.

Development of a new literary type

Among the Yüan dramatists, the following deserve special mention. Kuan Han-ch'ing, the author of some 60 plays, was the first to achieve distinction. His *Tou-o yüan* ("Injustice Suffered by Tou-o") deals with the deprivations and injustices suffered by the heroine, Tou-o, which begin when she is widowed shortly after her marriage to a poor scholar and culminate in her execution for a crime she has not committed. Wang Shih-fu, Kuan's contemporary, wrote *Hsi-hsiang chi* (*Romance of the Western Chamber*), based on a popular T'ang prose romance about the amorous exploits of the poet Yüan Chen, renamed Chang Chun-jui in the play. Besides its literary merits and its influence on later drama, it is notable for its length, two or three times that of the average Yüan play. Ma Chih-yüan, another contemporary, wrote 14 plays, of which the most celebrated is *Han-kung ch'iu* ("Sorrow of the Han Court"). It deals with the tragedy of a Han dynasty court lady, Wang Chao-chün, who, through the intrigue of a vicious portrait painter, was picked by mistake to be sent away to Central Asia as a chieftain's consort. Like the *Romance of the Western Chamber,* this play has been translated into western European languages.

Characteristics of the new literary genre

This new literary genre acquired certain distinct characteristics: (1) All extant compositions may be described as operas; (2) each play normally consists of four acts following a prologue; (3) the language of both the dialogue (for the most part in prose) and the arias—which alternate throughout the play—are fairly close to the daily speech of ordinary people; (4) all of the arias are in rhymed verse, and only one end rhyme is used throughout an act; (5) all of the arias in an act are sung by only one actor; (6) nearly all of the plays have a happy ending; (7) the characters in most of the plays are people of the middle and underprivileged classes—poor scholars, bankrupt merchants, Buddhist nuns, peasants, thieves, kidnappers, abductors, and women entertainers—antedating a similar trend in European drama by nearly four centuries.

At least 12 of the playwrights thus far identified were Sinicized members of originally non-Chinese ethnic groups—Mongols, Juchens, Uighurs, and other Central Asians.

*Poetry.* Another literary innovation, preceding but later interacting with the rise of the drama, was a new verse form known as *san-ch'ü* ("nondramatic songs"), a liberalization of the *tz'u,* which utilized the spoken language of the people as fully as possible. Although line length and tonal pattern were still governed by a given tune, extra words could be inserted to make the lyrics livelier and to clarify the relationship between phrases and clauses of the poem. The major dramatists were all masters of this genre.

*Vernacular fiction.* Similarly, fiction writers who wrote in a semivernacular style began to emerge, continuing the tradition of storytellers of the past or composing lengthy works of fiction written almost entirely in the vernacular. All of the early pieces of this type of book-length fiction were poorly printed and anonymously or pseudonymously published. Although many early works were attributed to such authors as Lo Kuan-chung, there is little reliable evidence of his authorship in any extant work. These novels exist in numerous, vastly different versions that can best be described as the products of long evolutionary cycles involving several authors and editors. The best known of the works attributed to Lo are *San-kuo chih yen-i* (*Romance of the Three Kingdoms*), *Shui-hu chuan* (*The Water Margin*), and *P'ing-yao chuan* ("The Subjugation of the Evil Phantoms"). The best of the three from a literary standpoint is the *Shui-hu chuan,* which gives full imaginative treatment to a long accretion of stories and anecdotes woven around a number of enlightened bandits—armed social and political dissenters.

Composite authorship of novels

**Ming dynasty: 1368–1644.** The Yüan dynasty was succeeded by the Ming dynasty, under which cultural influences from the South—expressed in movements toward cultural orthodoxy—again became important. Nearly all the major poets and prose writers in traditional literature were southerners, who enthusiastically launched and supported antiquarian movements based on a return to models of various ages of the past. With the restoration of competitive literary examinations, which had been virtually discontinued under the Mongols, the highly schematic *pa-ku wen-chang* ("eight-legged essay") was adopted as the chief yardstick in measuring a candidate's literary attainments. Despite occasional protests, it continued to engage the attention of aspirants to official literary honours from 1487 to 1901.

*Classical literature.* Although Ming poets wrote both *shih* and *tz'u* and their output was prodigious, poetry on the whole was imitative rather than freshly creative. Tirelessly, the poets produced verses imitating past masters, with few individually outstanding attainments.

Prose writers in the classical style were also advocates of antiquarianism and conscious imitators of the great masters of past ages. Rival schools were formed, but few writers were able to rise above the ruts of conventionalism. The Ch'in-Han school tried to underrate the achievements of Han Yü and Liu Tsung-yüan, along with the Sung essayists, and proudly declared that post-Han prose was not worth reading. The T'ang-Sung school, on the other hand, accused its opponents of limited vision and reemphasized Han Yü's dictum that literature should be the vehicle of Tao, equated with the way of life taught by orthodox Confucianism. These continuous squabbles ultimately led nowhere, and the literary products were only exquisite imitations of their respective models.

The Kung-an school of prose writers

The first voice of protest against antiquarianism was not heard until the end of the 16th century; it came from the Kung-an school, named for the birthplace of three brothers, of whom the middle one was the best known. Yüan Hung-tao challenged all of the prevailing literary trends, advocating that literature should change with each age and that any attempt at erasing the special stamp of an era could result only in slavish imitation. Declaring that he could not smile and weep with the multitude, he singled out "substantiality" and "honesty with oneself" as the chief prerequisites of a good writer.

This same spirit of revolt was shared by Chung Hsing and T'an Yüan-ch'un, of a later school, who were so unconventional that they explored the possibilities of writing intelligibly without observing Chinese grammatical usages. Although their influence was not long lasting, these two schools set the first examples of a new subgenre in prose—the familiar essay.

*Vernacular literature.* It was in vernacular literature that the writers of this period made a real contribution. In drama, a tradition started in the Sung dynasty and maintained in southern China during the period of Mongol domination was revitalized. This southern drama, also musical and known as *ch'uan-ch'i* ("tales of marvels"), had certain special traits: (1) a *ch'uan-ch'i* play contains from 30 to 40 changes of scene; (2) the change of end rhymes in the arias is free and frequent; (3) the singing is done by many actors instead of by the hero or heroine alone; (4) many plots, instead of being extracted from history or folklore, are taken from contemporary life.

Since there were no rules regulating the structure of the *ch'uan-ch'i,* playlets approaching the one-act variety were also written. This southern theatre movement, at first largely carried on by anonymous amateurs, won support gradually from the literati until finally, in the 16th century, a new and influential school was formed under the leadership of the poet-singer Liang Ch'en-yü and his friend the great actor Wei Liang-fu. The K'un school, initiating a style of soft singing and subtle music, was to dominate the theatre to the end of the 18th century.

Aside from drama and *ta-ch'ü* (a suite of melodies sung in narration of stories), which in the South were noticeably modified in spirit and structure, becoming more ornate and bookish—it was prose fiction that made the greatest progress in the 16th century. Two important novels took shape at that time. Wu Ch'eng-en's *Hsi-yu chi* is a fictionalized account of the pilgrimage of the Chinese monk Hsüan-tsang to India in the 7th century. The subject matter was not new; it had been used in early *hua-pen,* or "vernacular story," books and Yüan drama; but it had never been presented at length in such a lively and rapid-moving narration. Of all of the 81 episodes of trial and tribulation experienced by the pilgrim, no two are alike. Among the large number of monsters introduced, each has unique individuality. Like the *Shui-hu chuan,* it reveals the influence of the style of the oral storytellers, for each chapter ends with the sentence "in case you are interested in what is to follow, please listen to the next installment, which will reveal it." Unlike the *Shui-hu chuan,* which was written in a kind of semivernacular, the language used was the vernacular of the living tongue. For the author the choice must have been a deliberate but difficult one, for he had the novel first published anonymously to avoid disapproval. Besides eliciting numerous commentaries and "continuations" in China, it has two English translations.

Important novels of the 16th century

The title of the second novel (the author of which is unknown), *Chin P'ing Mei,* is composed of graphs from the names of three female characters. Written in an extremely charming vernacular prose style, the novel is a well-knit, long narrative of the awful debaucheries of the villain Ch'ing Hsi-men. The details of the different facets of life in 16th-century China are so faithfully portrayed that it can be read almost as a documentary social history of that age. The sexual perversions of the characters are so elaborately depicted that several Western translators have rendered a number of indelicate passages in Latin. The novel has been banned in China more than once, and all copies of the first edition of 1610 were destroyed.

**Ch'ing dynasty: 1644–1911/12.** The conquest of China by the Manchus, a Mongol people from the region north of China who set up the Ch'ing dynasty in 1644, did not disrupt the continuation of major trends in traditional literature. (During the literary inquisition of the 18th century, however, many books suspected of anti-Manchu sentiments were destroyed; and numerous literati were imprisoned, exiled, or executed.) Antiquarianism dominated literature as before, and excellent poetry and prose in imitation of ancient and medieval masters continued to be written, many works rivaling the originals in archaic beauty and cadence. Although the literary craftsmanship was superb, genuine creativity was rare.

*Poetry and prose nonfiction.* In the field of *tz'u* writing, the 17th-century Manchu poet Nara Singde (Sinicized name, Na-lan Hsing-te) was outstanding; but even he lapsed into conscious imitation of Southern T'ang models except when inspired by the vastness of open space and the beauties of nature. In nonfictional prose, Chin Jen-jui continued the familiar essay form.

*Prose fiction.* P'u Sung-ling continued the prose romance tradition by writing in *ku-wen* ("classical language") a series of 431 charming stories of the uncanny and the supernatural entitled *Liao-chai chih-i* (1766; "Strange Stories from the Liao-chai Studio"; Eng. trans., *Strange Stories from a Chinese Studio*). This collection, completed in 1679, was reminiscent of the early literary tale tradition, for it contained several T'ang stories retold with embellishments and minor changes to delineate the characters more realistically and to make the plots more probable. Such traditional supernatural beings as fox spirits, assuming in these stories temporary human form in the guise of pretty women, became for the first time in Chinese fiction humanized and likable. Despite the seeming success of these tales, the author soon became aware of the limitations of the *ku-wen* style for fiction writing and proceeded to produce a vernacular novel of some 1,000,000 words, the *Hsing-shih yin-yüan chuan* ("A Marriage to Awaken the World"). This long story of a shrew and her henpecked husband was told without any suggestion of a solution to the problems of unhappy marriages. Unsure of the reaction of his colleagues to his use of the vernacular as a literary medium, P'u Sung-ling had this longest Chinese novel of the old school published under a pseudonym.

The prose romances of P'u Sung-ling

Wu Ching-tzu satirized the 18th-century literati in a re-

alistic masterpiece, *Ju-lin wai-shih* (*c.* 1750; "Unofficial History of the Literati"; Eng. trans., *The Scholars*), 55 chapters loosely strung together in the manner of a picaresque romance. Unlike P'u Sung-ling, whom he far surpassed in both narration and characterization, he adopted the vernacular as his sole medium for fiction writing.

Better known and more widely read was Ts'ao Chan's *Hung-lou meng* (*Dream of the Red Chamber*), a novel of a love triangle and the fall of a great family, also written in the vernacular and the first outstanding piece of Chinese fiction with a tragic ending. Because its lengthy descriptions of poetry contests, which interrupt the narrative, may seem tiresome, especially to non-Chinese readers, they have been largely deleted in Western translations. Nevertheless, some Western critics have considered it one of the world's finest novels.

*Drama.* In drama, the Ming tradition of *ch'uan-ch'i* was worthily continued by several leading poets of the conventional school, though as a whole their dramatic writings failed to appeal to the masses. Toward the end of the 18th century, folk dramas of numerous localities began to gain popularity, converging finally at the theatres of Peking and giving rise to what came to be designated as Peking drama—a composite product that has continued to delight large audiences in China.

Popularity of folk dramas

*19th-century translations of Western literature.* By the early 19th century, China could no longer ward off the West and, after the first Opium War (1839–42), China's port cities were forcibly opened to increased foreign contacts. In due course, many Western works on diverse subjects were translated into Chinese. The quality of some of these was so outstanding that they deserve a place in the history of Chinese literature. One distinguished translator was Yen Fu, who had studied in Great Britain and whose renderings of Western philosophical works into classical Chinese were acclaimed as worthy of comparison, in literary merit, with the Chou philosophers. Another great translator was Lin Shu, who, knowing no foreign language himself but depending on oral interpreters, made available to Chinese readers more than 170 Western novels, translated into the literary style of Ssu-ma Ch'ien.

*19th-century native prose and poetry.* Meanwhile, writers of native fiction, especially in central and southern China, began to be seriously influenced by Western models. Using the vernacular and mostly following the picaresque romance structure of the *Ju-lin wai-shih,* they wrote fiction usually intended for serial publication and satirizing Chinese society and culture. One of these writers was Liu E, whose *Lao Ts'an yu-chi* (1904–07; *The Travels of Lao Ts'an*), a fictional account of contemporary life, pointed to the problems confronting the tottering Ch'ing dynasty.

19th-century developments in poetry

Poetry, long stagnant, at last began to free itself from the shackles of traditionalism. The most prominent poet, Huang Tsun-hsien, inspired by folk songs and foreign travel, tried to write poetry in the spoken language and experimented with new themes, new diction, and new rhythm. His young friend Liang Ch'i-ch'ao not only fervently supported Huang and his associates in what they called "the revolution in Chinese poetry" but also ventured forth in new directions in prose. Liang's periodical publications, especially, exerted an extensive influence on the Chinese people in the early years of the 20th century. Fusing all the unique and attractive features of the various schools of prose writing of the past into a new compound, Liang achieved a vibrant and widely imitated style of his own, distinguished by several characteristics: flexibility in sentence structure so that new terms, transliterations of foreign words and phrases, and even colloquial expressions could be accommodated; a natural liveliness; a touch of infectious emotionalism, which the majority of his readers enjoyed. Although he was too cautious to use the vernacular, except in fiction and plays, he did attempt to approximate the living speech of the people, as Huang Tsun-hsien had done in poetry.

As part of a westernization movement, the competitive literary examination system, which had been directly responsible for excessive conservatism and conventionality in thought as well as in literature, was abolished in 1905.

(T.-y.L./W.H.N.)

**Modern Chinese literature.** *May Fourth period.* Following the overthrow of the Manchu dynasty and the establishment of the Republic in 1912, many young intellectuals turned their attention to the overhauling of literary traditions, beginning with the language itself. In January 1917 an article by Hu Shih, a student of philosophy at Columbia University, entitled "Wen-hsüeh kai-liang ch'u-i" ("Tentative Proposal for Literary Reform") was published in the Peking magazine *Hsin ch'ing-nien* (*New Youth*). In it Hu called for a new national literature written not in the classical language but in the vernacular, the living "national language" (*kuo-yü*). Ch'en Tu-hsiu, the editor of *Hsin ch'ing-nien,* supported Hu's views in his own article "Wen-hsüeh ko-ming lun" ("On Literary Revolution"), which emboldened Hu to hone his arguments further in a second article (1918), entitled "Chien-she te wen-hsüeh ko-ming" ("Constructive Literary Revolution"), in which he spelled out his formula for a "literary renaissance." The literary reform movement that began with these and other "calls to arms" was a part of the larger May Fourth Movement for cultural and sociopolitical reform, whose name commemorates a 1919 student protest against the intellectual performance of the Chinese delegates to the Paris Peace Conference formally terminating World War I. At the outset, the literary reformers met with impassioned but mostly futile opposition from classical literati such as the renowned translator Lin Shu, who would largely give up the battle within a few years.

The first fruits of this movement were seen in 1918 and 1919 with the appearance in *Hsin ch'ing-nien* of such stories as "K'uang-jen jih-chi" ("The Diary of a Madman"), a Gogol-inspired piece about a "madman" who suspects that he alone is sane and the rest of the world is mad, and "Yao" ("Medicine"), both by Chou Shu-jen. Known by the pseudonym Lu Hsün, Chou had studied in Japan and, with his younger brother, the noted essayist Chou Tso-jen, had become a leader of the literary revolution soon after returning to China. Lu Hsün's acerbic, somewhat westernized, and often satirical attacks on China's feudalistic traditions established him as China's foremost critic and writer. His "Ah Q cheng-chuan" (1921; "The True Story of Ah Q"), a damning critique of early 20th-century conservatism in China, is the representative work of the May Fourth period and has become an international classic.

These early writings provided the impetus for a number of youthful intellectuals to pool their resources and promote shared ideals by forming literary associations. The Wen-hsüeh yen-chiu hui ("Literary Research Association"), generally referred to as the "realist" or "art-for-life's-sake" school, assumed the editorship of the established literary magazine *Hsiao-shuo yüeh-pao* (*Short Story Monthly*), in which most major fiction writers published their works throughout the 1920s, until the magazine's headquarters was destroyed by Japanese bombs in 1932. The socially reflective, critical-realist writing that characterized this group held sway in China well into the 1940s, when it was gradually eclipsed by more didactic, propagandistic literature. Members of the smaller Ch'uang-tsao she ("Creation Society"), on the other hand, were followers of the "Romantic" tradition who eschewed any expressions of social responsibility by writers, referring to their work as "art for art's sake." In 1924, however, the society's leading figure, Kuo Mo-jo, converted to Marxism, and the Creation Society evolved into China's first Marxist literary society. Much of the energy of members of both associations was expended in translating literature of other cultures, which largely replaced traditional Chinese literature as the foundation upon which the new writing was built. This was particularly true in drama and poetry, in which figures such as Henrik Ibsen and Rabindranath Tagore, respectively, were as well known to Chinese readers as indigenous playwrights and poets. In drama, the Nan-kuo she ("South China Society"), founded by the former Creationist T'ien Han, produced and performed several short plays that were a mixture of critical realism and melodrama, while poets of the Hsin-yüeh she ("Crescent Moon Society") such as the British-educated Hsü Chih-mo and the American-educated Wen I-to were creating new forms based on Western models, introducing the beauty of mu-

Literary associations

sic and colour into their extremely popular lyrical verse.

*1927–37.* Political events of the mid-1920s, in which Nationalist, Communist, and warlord forces clashed frequently, initiated a shift to the left in Chinese letters, culminating in 1930 in the founding of the Tso-i tso-chia lien-meng ("League of Leftist Writers"), whose membership included most influential writers. Lu Hsün, the prime organizer and titular head throughout the league's half-decade of activities, had stopped writing fiction in late 1925 and, after moving from Peking to Shanghai in 1927, directed most of his creative energies to translating Russian literature and writing the bitingly satirical random essays (*tsa-wen*) that became his trademark. Among the many active prewar novelists, the most successful were Mao Tun, Lao She, and Pa Chin.

Mao Tun, a founder of the Literary Research Association, was the prototypical Realist. The subjects of his socially mimetic tableaux included pre-May Fourth urban intellectual circles, bankrupt rural villages, and, in perhaps his best known work, *Tzu-yeh* (1933; *Midnight*), metropolitan Shanghai in all its financial and social chaos during the post-Depression era.

Lao She, modern China's foremost humorist, whose early novels were written while he was teaching Chinese in London, was deeply influenced by traditional Chinese storytellers and the novels of Charles Dickens. His works are known for their episodic structure, racy northern dialect, vivid characterizations, and abundant humour. Yet it was left to him to write modern China's classic novel, the moving tale of the gradual degeneration of a seemingly incorruptible denizen of China's "lower depths"—*Lo-t'o hsiang-tzu* (1936; "Camel Hsiang-tzu," published in English in a bowdlerized translation as *Rickshaw Boy,* 1945).

Pa Chin, a prominent Anarchist, was the most popular novelist of the period. A prolific writer, he is known primarily for his autobiographical novel *Chia* (1931; *The Family*), which traces the lives and varied fortunes of the three sons of a wealthy, powerful family. The book is a revealing portrait of China's oppressive patriarchal society, as well as of the awakening of China's youth to the urgent need for social revolution.

The 1930s also witnessed the meteoric rise of a group of novelists from Northeast China (Manchuria) who were driven south by the Japanese annexation of their homeland in 1932. The sometimes rousing, sometimes nostalgic novels of Hsiao Chün and Hsiao Hung and the powerful short stories of Tuan-mu Hung-liang became rallying cries for anti-Japanese youth as signs of impending war mounted.

Poetry of the 1930s underwent a similar politicization, as more and more students returned from overseas to place their pens in the service of the "people's resistance against feudalism and imperialism." The lyrical verse of the early Crescent Moon poets was replaced by a more socially conscious poetry by the likes of Ai Ch'ing, T'ien Chien, and Tsang K'o-chia that appealed to the readers' patriotic fervour. Others, particularly those who had at first gravitated toward the Crescent Moon Society, began striking out in various directions: notable works of these authors include the contemplative sonnets of Feng Chih, the urbane songs of Peking by Pien Chih-lin, and the romantic verses of Ho Ch'i-fang. Less popular, but more daring, were Tai Wang-shu and Li Chin-fa, poets of the Hsien-tai ("Contemporary Age") group, who wrote very sophisticated, if frequently baffling, poetry in the manner of the French Symbolists.

While fiction reigned supreme in the 1930s, as the art of the short story was mastered by growing numbers of May Fourth writers, and novels were coming into their own, the most spectacular advances were made in drama, owing largely to the efforts of a single playwright. Although realistic social drama written in the vernacular had made its appearance in China long before the 1930s, primarily as translations or adaptations of Western works, it did not gain a foothold on the popular stage until the arrival of Ts'ao Yü, whose first play, *Lei-yü* (1934; *Thunderstorm*), a tale of fatalism, retribution, and incestual relations among members of a rich industrialist's family, met with phenomenal success. It was followed over the next several years by other critically and popularly acclaimed plays, including *Jih-ch'u* (1936; *Sunrise*) and *Yüan-yeh* (1937; *Wilderness*), all of which examined pressing social issues and universal human frailties with gripping tension and innovative dramaturgy. Political realities in future decades would force a steady decline in dramatic art, so that Ts'ao Yü's half-dozen major productions still stand as the high-water mark of modern Chinese theatre. Yet, even though movies, television, and other popular entertainments would weaken the resiliency of this literary form, it would still serve the nation as an effective propaganda medium, particularly during the war of resistance.

The plays of Ts'ao Yü

*The war years: 1937–45.* During the Sino-Japanese War, most writers fled to the interior, where they contributed to the war effort by writing patriotic literature under the banner of the Chung-hua ch'üan-kuo wen-i chieh k'ang-ti hsieh-hui ("All-China Anti-Japanese Federation of Writers and Artists"), founded in 1938 and directed by Lao She. All genres were represented, including reportage (*pao-kao wen-hsüeh*), an enormously influential type of writing that was a natural outgrowth of the federation's call for writers to go to the countryside and the front lines. Literary magazines were filled with short, easily produced and adaptable plays, topical patriotic verse, and war-zone dispatches. Among the major writers who continued to produce work of high quality during this period were Pa Chin, Ts'ao Yü, Mao Tun, and Ting Ling. The latter's fictional explorations of the female psyche and the social condition of women had caught the public's imagination in the 1920s, and in the late 1930s she established herself as the major literary figure in the Communist stronghold of Yen-an.

The growing dissatisfaction of intellectuals with the Nationalist government in Chungking surfaced dramatically during the civil war that raged throughout China following Japan's surrender, ending with the Nationalists' retreat to Taiwan and the establishment, in October 1949, of the People's Republic of China. Most writers, feeling intense pride and welcoming the challenge, chose to remain on the mainland and serve the new government.

*1949 to the present.* Literature on the China mainland since 1949 has largely been a reflection of political campaigns and ideological battles. This state of affairs can be traced to Mao Tse-tung's 1942 "Tsai Yen-an wen-i tso-t'an-hui shang te chiang-hua" ("Talks at the Yenan Forum on Literature and Art"), in which he articulated his position that literature, which existed to serve politics, was to be popularized while the people's level of literary appreciation was gradually being elevated. Mao's call for a truly proletarian literature—written by and for workers, peasants, and soldiers—gave rise to a series of rectification campaigns that further defined and consolidated party control over literary activities. In 1949, the First National Congress of Writers and Artists was convened, and the All-China Federation of Literature and Art Circles was founded, with Kuo Mo-jo elected as its first chairman.

Mao's "Talks at the Yenan Forum"

Mao's literary ideals had first been realized in the 1940s by Chao Shu-li, whose early stories, such as "Li Yu-ts'ai pan-hua" ("The Rhymes of Li Yu-ts'ai"), were models of proletarian literature, both in form and in content. As the civil war neared its conclusion, novels of land reform, such as Ting Ling's prizewinning *T'ai-yang chao tsai Sang-kan-ho shang* (1949; *The Sun Shines over the Sangkan River*) and *Pao-feng tsou-yü* (1949; *The Hurricane*) by Chou Li-po, became quite popular. Few of the established May Fourth writers continued to produce fiction after 1949, for their experience as social critics did not prepare them for Socialist Realism, a method of composition, borrowed from the Soviet Union, according to which society is described as it should be, not necessarily as it is. Many of the older poets, however, were successful during the early postliberation years, writing poetry in praise of land reform, modernization, and Chinese heroes of the Korean War. Playwrights were also active, introducing more proletarian themes into their works, some of which incorporated music. By this time, Lao She had begun writing plays, such as *Lung-hsü kou* (1951; *Dragon Beard Ditch*), which earned him the prestigious title of People's Artist. Another very popular play, *Pai-mao nü*

(1953; *White-Haired Girl*) by Ho Ching-chih, was taken from a contemporary folk legend.

During the mid-1950s, an experiment in liberalization—the Hundred Flowers Campaign—was abruptly terminated as criticism of the party went beyond all expectations; it was followed by an anti-rightist movement that purged the cultural ranks of most preliberation writers and artists. The literary nadir, however, was not reached until the Cultural Revolution (1966–76), when the only literature available were a few carefully screened works by Lu Hsün, a handful of model revolutionary Peking operas, and the revolutionary-romantic novels of Hao Jan. After the death of Mao and the fall of the Gang of Four, literature made a comeback and most surviving writers were rehabilitated, although the progress was as rocky as the political scene Chinese literature continued to reflect.

The accusatory "scar literature," a sort of national catharsis that immediately followed the 10-year "holocaust," gave way to more professional and more daring writing, as exemplified in the stories of Wang Meng, with their stylistic experiments in stream of consciousness; the symbolic "obscure" poetry of Pei Tao and others; the relatively bold dramas, both for the stage and for the screen, of several playwrights; and the innovative investigative reportage of Liu Pin-yen. In addition to translated literature from the West, literature from Taiwan also began to reach mainland writers and readers as literary restrictions continued to fall gradually.

*Taiwanese literature after 1949.* The first decade of literary activities in Taiwan after 1949 was characterized by stereotypical anti-Communist fiction and drippingly sentimental essays and poetry, producing little memorable literature other than novels such as *Yang-ko* (1954; *The Rice-Sprout Song*) by Chang Ai-ling, a story of peasant life under Communist rule, and *Hsüan-feng* (1959; *The Whirlwind*), Chiang Kuei's novel of power struggles in Shantung. In the 1960s, however, a group of Taiwan University students ushered in the modernist era by publishing their own craftsmanlike stories, which were heavily indebted to such Western masters as Franz Kafka, James Joyce, and Virginia Woolf. Many of these writers, such as Pai Hsien-yung, author of *Yu-yüan ching-meng* (1982; *Wandering in the Garden, Waking from a Dream*), remained active and influential in the mid-1980s. Vernacular poetry in Taiwan developed around several societies in which modernist, even surrealist, verse was in vogue. These poets, while not widely accepted by the reading public, strongly influenced the more accessible poets who followed. The late 1960s witnessed the rise of regional (*hsiang-t'u*) writing, in which the Taiwanese countryside served as the setting for fiction and poetry that effectively captured the dramatic social and psychological effects of transition from a rural to an urban-based society. Huang Ch'un-ming's *Ni-szu i-chih lao-mao* (1980; *The Drowning of an Old Cat*) is representative of this nativist school, which in later years gave way to a more nationalistic literature that reflected Taiwan's current political situation. Mainland literature occasionally appears in Taiwanese periodicals, while firsthand experiences and observations by mainland émigrés and overseas Chinese, such as the collection of stories *Yin hsien-chang* (1976; *The Execution of Mayor Yin*) by Ch'en Jo-hsi, are given broad exposure.

Taiwanese regionalism

(H.C.G.)

**BIBLIOGRAPHY**

*General works:* KARL LO, *A Guide to the Ssŭ pu ts'ung k'an: Being an Index to Authors, Titles, and Subjects* (1965); HU SHIH, *Pai-hua wen-hsüeh shih* ("History of Vernacular Literature," 1929); WU-CHI LIU, *An Introduction to Chinese Literature* (1966, reprinted 1967); YUANJUN FENG, *An Outline History of Classical Chinese Literature,* trans. by XIANYI YANG and GLADYS YANG (1983); BURTON WATSON, *Early Chinese Literature* (1962); PATRICK HANAN, *The Chinese Vernacular Story* (1981); C.T. HSIA, *The Classic Chinese Novel: A Critical Introduction* (1968, reissued 1980); COLIN MACKERRAS (ed.), *Chinese Theater: From Its Origins to the Present Day* (1983), a collection of essays on various eras and genres of traditional drama, its performance, and its audience; WILLIAM H. NIENHAUSER, JR. (ed.), *Indiana Companion to Traditional Chinese Literature* (1985), containing essays and entries with extensive bibliographies on all aspects of traditional Chinese literature; translations by ARTHUR WALEY and BURTON WATSON, too numerous to be listed; WU-CHI LIU and IRVING YUCHENG LO (eds.), *Sunflower Splendor: Three Thousand Years of Chinese Poetry* (1975, reprinted 1977), the most extensive collection of translations; Y.W. MA and JOSEPH S.M. LAU (eds.), *Traditional Chinese Stories: Themes and Variations* (1978), the standard collection of translations; EUGEN FEIFEL (ed. and trans.), *Geschichte der chinesischen Literatur: Mit Berücksichtigung ihres geistesgeschichtlichen Hintergrundes,* 3rd ed. (1967); GEORGES MARGOULIÈS, *Évolution de la prose artistique chinoise* (1929), *Histoire de la littérature chinoise: prose* (1949), and *Histoire de la littérature chinoise: poésie* (1951); DERK BODDE, "Myths of Ancient China," in SAMUEL NOAH KRAMER (ed.), *Mythologies of the Ancient World,* pp. 367–408 (1961), a good critical introduction but limited to five classical myths; and WOLFGANG MÜNKE, *Die klassische chinesische Mythologie* (1976).

*Modern Chinese literature:* All the works mentioned in this section of the article are available in English translation and can be located in DONALD A. GIBBS and YUN-CHEN LI, *A Bibliography of Studies and Translations of Modern Chinese Literature, 1918–1942* (1975); and WINSTON L.Y. YANG and NATHAN K. MAO (eds.), *Modern Chinese Fiction: A Guide to Its Study and Appreciation: Essays and Bibliographies* (1981). The most useful historical works are TSE-TSUNG CHOU, *The May Fourth Movement: Intellectual Revolution in Modern China* (1960, reissued 1967); C.T. HSIA, *A History of Modern Chinese Fiction,* 2nd ed. (1971); D.W. FOKKEMA, *Literary Doctrine in China and Soviet Influence, 1956–1960* (1965); and MERLE GOLDMAN, *Literary Dissent in Communist China* (1967, reissued 1971). Synopses of representative works are given in JOSEPH SCHYNS, *1500 Modern Chinese Novels & Plays* (1948, reissued 1970); and MEISHI TSAI, *Contemporary Chinese Novels and Short Stories, 1949–1974: An Annotated Bibliography* (1979). Poetry is treated in KAI-YU HSU (ed. and trans.), *Twentieth Century Chinese Poetry: An Anthology* (1963, reissued 1970); and ANGELA C.Y. JUNG PALANDRI (ed. and trans.), *Modern Verse from Taiwan* (1972). The best anthologies of translated literature are JOSEPH S.M. LAU, C.T. HSIA, and LEO OU-FAN LEE (eds.), *Modern Chinese Stories and Novellas, 1919–1949* (1981); JOSEPH S.M. LAU (ed.), *The Unbroken Chain: An Anthology of Taiwan Fiction Since 1926* (1983); KAI-YU HSU (ed.), *Literature of the People's Republic of China* (1979); PERRY LINK (ed.), *Roses and Thorns: The Second Blooming of the Hundred Flowers in Chinese Fiction, 1979–1980* (1984); and EDWARD M. GUNN (ed.), *Twentieth-Century Chinese Drama: An Anthology* (1983).

(W.H.N./H.C.G.)

# Chordates

The phylum Chordata (chordates) consists of three subphyla: Tunicata (also called Urochordata), Cephalochordata (also called Acrania), and Vertebrata (also called Craniata). As the name implies, at some time in the life cycle a chordate possesses a stiff, dorsal supporting rod (the notochord). Also characteristic of the chordates are a tail that extends behind and above the anus, a hollow nerve cord above (or dorsal to) the gut, gill slits opening from the pharynx to the exterior, and an endostyle (a mucus-secreting structure) or its derivative between the gill slits. (A characteristic feature may be present only in the developing embryo and may disappear as the embryo matures into the adult form.) A somewhat similar body plan can be found in the closely related phylum Hemichordata.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 313, and the *Index.*

This article is divided into the following sections:


Chordates: the phylum Chordata 241
- General features
- Natural history
  - Reproduction and life cycle
  - Ecology and habitats
  - Locomotion
  - Associations
- Form and function
  - General features
  - External features
  - Internal features
- Evolution and paleontology
- Classification
  - Annotated classification
  - Critical appraisal

Major chordate groups 243
- Tunicates 243
  - General features
  - Natural history
  - Form and function
  - Evolution and paleontology
  - Classification
- Cephalochordates 245
  - General features
  - Natural history
  - Form and function
  - Evolution and paleontology
  - Classification
- Vertebrates 247
  - General features
  - Natural history
  - Form and function
  - Evolution and paleontology
  - Classification

Bibliography 250


## CHORDATES: THE PHYLUM CHORDATA

### GENERAL FEATURES

Tunicates are small animals, typically one to five centimetres (0.4 to 2.0 inches) long, with a minimum length of about one millimetre (0.04 inch) and a maximum length slightly more than 20 centimetres; colonies may grow to 18 metres (59 feet) in length. Cephalochordates range from one to three centimetres. Vertebrates range in size from tiny fish to the whales, which include the largest animals ever to have existed.

Tunicates are marine animals, either benthic (bottom dwellers) or pelagic (inhabitants of open water), that often form colonies by asexual reproduction. They feed by taking water in through the mouth, using the gill slits as a kind of filter. The feeding apparatus in cephalochordates is similar. They have a well-developed musculature and can swim rapidly by undulating the body. Cephalochordates usually live partially buried in marine sand and gravel.

Evolutionary changes

Vertebrates retain traces of a feeding apparatus like that of tunicates and cephalochordates. The gill slits, however, ceased to function as feeding structures, and then later as respiratory devices, as the vertebrate structure underwent evolutionary changes. Except in some early branches of the vertebrate lineage (*i.e.*, agnathans) a pair of gill arches has become modified so as to form jaws. The fishlike habitus that evidently began with cephalochordates became modified by the development of fins that were later transformed into limbs. With the invasion of the vertebrates into fresh water and then onto land, there was a shift in means of breathing—from gills to lungs. Other modifications, such as an egg that could develop on land, also emancipated the vertebrates from water. Elaboration of the locomotory apparatus and other developments allowed a diversification of structure and function that produced the amphibians, reptiles, birds, and mammals.

### NATURAL HISTORY

**Reproduction and life cycle.** The chordate life cycle begins with fertilization (the union of sperm and egg). In its primitive form, fertilization occurs externally, in the water. Asexual reproduction takes place in tunicates and in some vertebrates (females of some fish and lizards can reproduce without fertilization). Hermaphroditism (possessing both male and female reproductive organs) is found in tunicates and some fishes, but otherwise the sexes are separate. Larvae (very young forms that differ considerably from the juveniles and adults), when they do occur, differ in structure from the larvae of nonchordates. Internal fertilization, viviparity (giving birth to young that have undergone embryological development), and parental care are common in tunicates and vertebrates.

**Ecology and habitats.** Chordates are common in all major habitats. Tunicate larvae either seek out a place where they can attach and metamorphose into an adult or develop into adults that float in the open water. Cephalochordates develop in the open water, but as adults they lie partially or entirely buried in sand and gravel. In either case, they are filter feeders with simple behaviour. Vertebrates are much more complex and, in keeping with their more active manner of obtaining food, highly varied in their ecology and habits.

Muscular movements

**Locomotion.** Chordates are capable of locomotion by means of muscular movements at some stage in life. In tunicate larvae, this is accomplished using a tail; in cephalochordates, by undulations of the body; and in vertebrates, by general body movements (as in eels and snakes) and by the action of fins and limbs, which in birds and some mammals are modified into wings.

**Associations.** Chordates enter into a wide variety of symbiotic relationships and are especially noteworthy as hosts for parasites. Family groups and societal relationships, in both a broad and narrow sense, are particularly well developed in vertebrates, due primarily to their elaborate nervous systems. This phenomenon is seen in schools of fish, flocks of birds, and herds of mammals, as well as in the primate associations that suggest the beginnings of human society.

## FORM AND FUNCTION

**General features.** Chordates have many distinctive features, suggesting that there has been extensive modification from simple beginnings. The early stages of chordate development show features shared with some invertebrate phyla, especially the mouth that forms separately from the anus, as it does in the phyla Hemichordata, Echinodermata, and Chaetognatha. Likewise, as in these phyla, the coelom, or secondary body cavity around the viscera, develops as outpouchings of the gut. A coelom also is present in some more distantly related phyla, including Annelida, Arthropoda, and Mollusca, but the main organs of the body are arranged differently in these phyla. In chordates the main nerve cord is single and lies above the alimentary tract, while in other phyla it is paired and lies below the gut. Cephalochordates and vertebrates are segmented, as are the annelids and their relatives; however, segmentation in the two groups probably evolved independently. The gill slits and some other features that are common among the hemichordates and the chordates originated before the chordates became a separate group. Hemichordates have no tail above the gut and no mucus-secreting endostyle between the gill slits.

Coelom

**External features.** An ancestral chordate, as suggested by the adult lancelet and the tadpole larva of tunicates, had a distinct front and hind end, an anterior mouth, a posterior tail above an anus, unpaired fins, and gill slits that opened directly to the exterior. A free-swimming tunicate larva metamorphoses into an attached, sessile adult with an atrium that surrounds the gills. The atrium of lancelets probably evolved independently.

**Internal features.** *Skeleton and support.* The chordate notochord is a stiff rod with a turgid core and fibrous sheath. It keeps the animal from shortening when locomotory waves are produced through muscular contraction. The chordate body is supported by fluid in the body cavities. In tunicates, added support is provided by the tunic. Cartilaginous material supports the gills and other body parts of tunicates and cephalochordates. Immature vertebrate skeletons generally consist largely of cartilage, which becomes increasingly bony with age. The cartilaginous skeletons of sharks and some other vertebrates are thought to have evolved from more highly mineralized ones.

Cartilage

*Tissues and muscles.* In both cephalochordates and vertebrates, muscles used in locomotion are well developed and organized segmentally. The tail musculature of tunicates is simpler and without clear indications of segmentation. There is at least a small amount of musculature throughout the body of all chordates. As jaws, limbs, and other body parts have evolved in vertebrates, so have the muscles that operate them.

*Nervous system and sense organs.* The anterior end of the main nerve cord in chordates is enlarged to form at least the suggestion of a brain, but a brain is well developed only in vertebrates. Tunicate larvae have visual organs sensitive to light and sense organs responsive to the direction of gravity. Pigment spots and light receptors in the nerve cord of lancelets detect sudden changes in light intensity. The eyes and other sense organs of vertebrates are more elaborate and complex.

The presence in cephalochordates and vertebrates of a nervous system with segmentally repeated nerves arising from the dorsal hollow nerve cord is suggestive of a common ancestry. The tunicate nervous system does not have the segmentally repeated nerves. The brains of all vertebrates are greatly enlarged and subdivided into functionally specialized regions.

*Digestion and nutrition.* Both tunicates and cephalochordates are filter feeders of small particles of food suspended in the water. Beating cilia (hairlike cellular extensions) on the gill slits draw a current of water into the mouth and through the pharynx, where a sheet of mucus, secreted by the endostyle (a glandular organ lying below the two rows of gill slits), filters suspended food particles from the water. Cilia lining the pharynx move the food-rich sheet of mucus upward over the gill slits, and it is then rolled up and transported to the posterior part of the gut. The water current passes into the atrium and exits through the atrial opening.

Endostyle

Something similar to this arrangement occurs in the vertebrates in the "ammocoetes" larva stage of the primitive jawless fish called the lamprey. The difference is that the food consists of somewhat larger particles that have been deposited on the bottom (detritus), and, instead of the feeding current being driven by cilia, the pharyngeal musculature pumps water and food particles across the gill slits. The earliest fishes probably fed on detritus, and a sucking action is retained by their extant representatives (lampreys and hagfishes). With the development of jaws, it became possible for the vertebrates to capture and seize larger food items.

The lower digestive tract of the primitive chordate is a simple tube with a saclike stomach. There are only indications of the specialized areas and of glandlike structures, such as the liver and pancreas, that occur in vertebrates.

*Excretion.* The excretion of wastes and the control of the chemical composition of the internal environment are largely effected by kidneys, although other parts of the body, including the gills, may play an important role. Tunicates and cephalochordates have a salt content essentially the same as seawater, but vertebrates, even marine species, have body fluids of low salt content, with the exception of hagfishes. A possible explanation is that the vertebrates evolved in fresh water, but it seems reasonable that hagfishes branched off while still marine and that the freshwater form evolved later.

*Respiration.* A primitive chordate gill is present in tunicates and cephalochordates, where it serves in both respiration and feeding. The vertebrate gill may retain some role in feeding, although the current is now produced by the action of muscles, not cilia. The gills became reduced in number in various lineages, and they were strengthened by supporting elements, some of which evolved into jaws. Lungs, already present in fishes, became the main respiratory organs of terrestrial vertebrates.

Evolution in gill structure

*Circulatory system.* The circulatory system in chordates has a characteristic pattern. In tunicates and vertebrates the blood is propelled by a distinct heart; in cephalochordates, by contraction of the blood vessels. Unoxygenated blood is driven forward via a vessel called the ventral aorta. It then passes through a series of branchial arteries in the gills, where gas exchange takes place, and the oxygenated blood flows to the body, much of it returning to its origin via a dorsal aorta. The blood of vertebrates passes through the tissues via tiny vessels called capillaries. In tunicates and cephalochordates, capillaries are absent and the blood passes through spaces in the tissues instead.

*Hormones.* In vertebrates, endocrine glands (those of internal secretion) produce hormones that regulate many physiological activities. In tunicates and cephalochordates, organs have been identified that correspond in anatomical position to the pituitary gland of vertebrates, but which hormones, if any, they secrete is uncertain. In vertebrates, the thyroid gland produces thyroxine, an iodine-containing hormone that helps regulate metabolism. The thyroid is a modified endostyle, as can be illustrated by larval lampreys in which the thyroid still secretes mucus for use in feeding. The endostyles of lancelets take up iodine and form thyroxine, but the thyroxine formed may not function as a hormone in the lancelets themselves.

*Features of defense and aggression.* Tunicates largely rely upon the passive defense afforded by their heavy tunic. Lancelets move rapidly through the substrate, and their well-developed locomotory apparatus evolved largely to provide a means of escaping predators. Vertebrates have ceased to feed on detritus brought to them by water currents. They have shifted to consuming larger foodstuffs and to actively locating, pursuing, and subduing what they eat.

## EVOLUTION AND PALEONTOLOGY

Chordates originated sometime earlier than 590 million years ago; that is, they predate the fossil record. Early representatives were soft-bodied, and they therefore left a poor fossil record. Fossils dubiously attributed to all three chordate subphyla have been found in Cambrian rocks (more than 505 million years old), and an extensive vertebrate fossil record begins around 400 million years ago.

Embryological evidence places the phylum Chordata within the deuterostomes (bilaterally symmetrical animals with undeterminate cleavage and whose mouth does not arise from the blastopore), which also includes the phyla Hemichordata, Echinodermata, and Chaetognatha. The closest relatives of the chordates are probably the hemichordates, since these animals possess gill slits and other features not found in other animal phyla. A slightly more remote relationship to the echinoderms is inferred on the basis of resemblances between the larvae in some groups of hemichordates and echinoderms. The derivation of chordates from certain fossil echinoderms has been argued on the basis of features such as what appear to be gill slits. Theories that derive them from other phyla (*e.g.,* Annelida, Nemertea, Arthropoda) have been proposed, but such theories have few contemporary advocates.

Whether the ancestral chordate was more like a tunicate or a cephalochordate has been extensively debated. The classical theory is that the ancestor was like a cephalochordate, and that one lineage became attached to hard surfaces and evolved into tunicates, whereas another remained unattached and evolved into vertebrates. An alternative theory is that the ancestor was like a tunicate and that the other two subphyla arose by modification of the tadpole larva. There is some preference for the classical theory because it provides the most satisfactory way of accounting for the similarities between chordates and hemichordates of the subphylum Enteropneusta. Within the chordates the tunicates probably branched off before the common ancestor of cephalochordates and vertebrates arose, for the latter resemble each other in some details of neuroanatomy and biochemistry.

### CLASSIFICATION

**Annotated classification.**

PHYLUM CHORDATA
Deuterostomatous eucoelomates; gill clefts; endostyle or its derivative in pharynx; notochord; hollow dorsal nerve cord; tail posterior and dorsal to anus.

**Subphylum Tunicata** (or **Urochordata**; tunicates)
Notochord, when present, restricted to tail; body covered with tunic, but sometimes only cuticle; atrium, absent in Appendicularia, dorsal and often paired in embryonic development; heart present; generally sessile (attached) as adults; see below *Tunicates.*

*Class Ascidiacea* (sea squirts)
Sessile; benthic; solitary or colonial within a common tunic.

*Class Appendicularia* (larvacea)
Free-swimming; pelagic; resembles tadpole larvae of ascidians; 1 pair of gill slits; no distinct atrium.

*Class Thaliacea*
Pelagic; forms aggregations or colonies.

**Subphylum Cephalochordata** (or **Acrania**; lancelets)
Notochord extends entire body length, with tip anterior to nerve cord; atrium a single cavity with single, ventral opening; segments well developed; head poorly developed; no paired fins; no heart; see below *Cephalochordates.*

**Subphylum Vertebrata** (or **Craniata**; vertebrates)
Notochord extends to the back of a well-developed head; no atrium; segments well developed; paired fins or limbs usually present; heart present; see below *Vertebrates.*

**Critical appraisal.** This outline gives the major groups of chordates; for lesser units see the sections below on *Tunicates, Cephalochordates,* and *Vertebrates.* Modern systematic biology attempts to arrange groups of organisms in a way that suggests the genealogical relationships (branching sequences) and therefore presents an epitome of evolutionary history. It also may attempt to show where there are important differences among the various groups. These goals often conflict. In a purely genealogical system, each group must correspond to a single lineage (clade) composed of the common ancestor and all of its descendants. A group that does not meet both of these requirements is called a grade and may be used as an informal group. Groups that do not contain the common ancestor, and therefore had two separate origins, are said to be polyphyletic. Such polyphyletic grades, which would put whales together with fish or birds together with bats, have generally been abandoned as soon as they were recognized. Another kind of grade, which does not include all the descendants of the common ancestor, is said to be paraphyletic and is retained in more conservative systems. Within the vertebrates the class Aves is a clade, but the class Reptilia is a grade, for the birds are modified dinosaurs. Some systems do not recognize Reptilia as a formal group. Likewise, birds, mammals, reptiles, and amphibians are all modified fish, and the old class of fishes (Pisces) is now rarely used. Vertebrata is a single clade, but "invertebrate" is a grade consisting of all animals except vertebrates. Therefore there is no formal group called Invertebrata.

Many differences among systems are quite subjective. This is often the case when a group may be ranked either as a class or as a subphylum. The organizational limits of some groups are also largely a matter of opinion. Some authors have placed the phylum Hemichordata within the Chordata, expressing the close genealogical relationship. Others prefer to keep them as a separate phylum because hemichordates lack what are considered important chordate features.

# MAJOR CHORDATE GROUPS

## Tunicates

Adult members of the subphylum Tunicata (also called Urochordata) commonly are embedded in a tough, secreted tunic containing cellulose (a glucose polysaccharide not normally found in animals). All tunicates are marine; the less modified forms are benthic (bottom-dwelling and sessile), while the more advanced forms are pelagic (floating and swimming in open water). A characteristic tadpole larva develops in the life cycle, and in one group the adult closely resembles this larva, which has many features in common with other chordates.

### GENERAL FEATURES

Tunicate classes

**Size range and diversity of structure.** The tunicates are divided into three classes: Ascidiacea (ascidians, or sea squirts), Appendicularia, and Thaliacea. Ascidians are largely benthic animals. They often form colonies, comprising a few to many individuals (zooids), which reach up to two metres in length. Solitary (noncolonial) forms range from one millimetre to over 20 centimetres in length. The adult appendicularian resembles the tadpole larva of other tunicates. The body is enveloped in a "house," with which the animal nets food. Small (usually around five millimetres in length, including the tail) and simple, appendicularians do not form colonies. They spend their entire lives in the open sea. The thaliaceans (pyrosomes, dolioloids, and salps) are also pelagic. Their structure suggests that they are ascidians modified in adaptation to conditions in open water. They have specialized modes of reproduction, sometimes with a complicated alteration of sexual and asexual generations. Pyrosomes form long, tubular colonies. Dolioloids and salps occur both as solitary individuals and as chains.

**Distribution and abundance.** Tunicates are distributed in ocean waters from the polar regions to the tropics. Free-swimming tunicates are found throughout the oceans as plankton, while sessile forms grow mainly on solid surfaces such as wharf piles, ship hulls, rocks, and the shells of various sea creatures.

**Importance.** Although rarely eaten by humans, tunicates are an important link in the food chain and thus indirectly provide humans with a source of food. Tunicates contain some unusual chemicals, and some of these may prove useful as drugs. Some tunicates are fouling organisms that grow on ships' hulls. Their main interest to humans is in providing clues to the possible ancestry of vertebrates.

NATURAL HISTORY

**Reproduction and life cycle.** With rare exceptions, tunicates are hermaphrodites, but reproduction may be by sexual or asexual (budding) means. In general, hermaphroditic animals do not self-fertilize (*i.e.,* provide both the male and female gametes) if they can avoid doing so, a rule that seems also to be true of tunicates. In primitive forms the eggs are fertilized, and development takes place, in the surrounding water, but often embryos are retained in the female's atrium or elsewhere until the larva is developed.

Larvae

The larval stage is brief; the larva does not feed, but concentrates on finding an appropriate place for the adult to live. In keeping with this motile phase, the muscular tail comprises two-thirds of the larval body; it is supported by a notochord and contains a nerve cord. Gravity- and light-sensitive sensory vesicles along the dorsal surface of the larval body orient the animal as it swims. After a period of up to a few days, the larva will settle and attach itself to a surface using three anterior adhesive papillae. As the larva metamorphoses into an adult, the tail resorbs, providing food reserves for the developing animal. Free-swimming tunicates metamorphose without attachment.

Colonies are formed by asexual reproduction, with zooids usually being formed by budding. In thaliaceans, two groups (dolioloids and salps) have a complex system of alternating generations; the first generation reproduces by budding, and the products of this budding release sperm and eggs.

**Locomotion.** Tadpole larvae and appendicularians swim by undulating the tail. Despite their sessile life-styles, some adult ascidians can move by attaching with one area of the body and letting go with another. Movement of colonies up to 1.5 centimetres per day has been recorded. In thaliaceans an exhalant current of water, which in dolioloids and salps is combined with a strong muscular contraction, creates a jet stream that propels the animal forward.

**Food and feeding.** An internal mucous sheet, secreted by the endostyle, allows ascidians and thaliaceans to utilize a variety of organisms, especially small plants (phytoplankton) as their food source (see below *Internal features: Digestion, nutrition, and excretion*). Some trap small animals. The feeding mechanism is different in appendicularians. Glands on the surface of the body secrete a complex house made up of mucus, which surrounds the animal. Undulations of the tail produce a feeding current that draws water into the house and through a fine sheet of mucus, which serves as a net to filter the food. Appendicularians feed on microscopic organisms (nannoplankton).

**Associations.** Tunicates often are hosts for various parasitic animals. Some tunicates, especially in the tropics, live symbiotically with unicellular plants and blue-green algae that may provide them with a supply of food.

FORM AND FUNCTION

**General features.** A tunicate tadpole larva contains several chordate features, such as the notochord, dorsal nerve cord, and tail. These features are lost, however, as the larva metamorphoses into the adult form (Figure 1). The tunicate larva has special organs of sense and attachment, which it uses to find and occupy a suitable habitat. Once settled, the larval features regress and considerable changes in size and proportion of parts take place. Larvae attach at the anterior end (under the mouth) and, by differential growth, the larva loses most of the chordate features. For example, the notochord, nerve cord, and most of the tail are resorbed. The area between the mouth and the point of attachment grows rapidly until the mouth comes to be directed away from the point of attachment, which now becomes the posterior end of the animal. The atrium usually forms from a pair of pouches that grow inward and fuse into a single cavity that opens near the mouth on what is technically the dorsal area of the body.

Characteristic features

**External features.** A solitary tunicate has two major openings, or siphons, on the surface away from the area of attachment: a branchial aperture, through which water enters the body, and an atrial aperture, through which water, wastes, and gametes leave. The animal is covered with a thick tunic, which consists of some cells, blood vessels, and a secretion of a variety of proteins and carbohydrates, including cellulose, which, although abundant in plants, is unusual in animals.

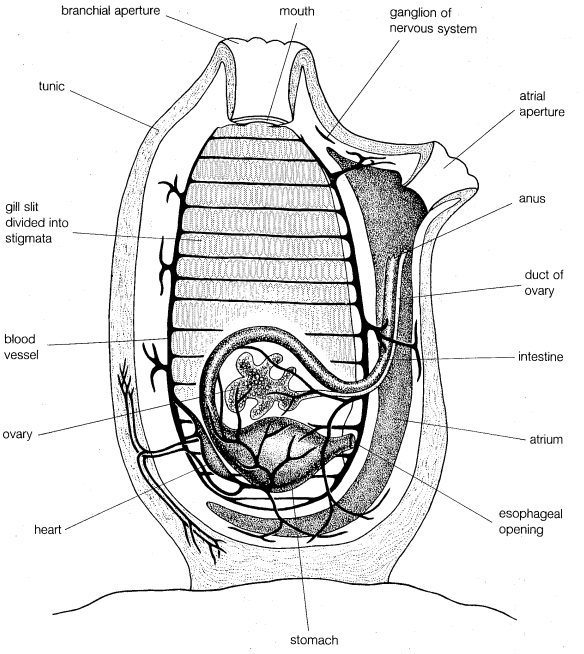



Figure 1: Anatomy of *Ascidia,* a solitary tunicate.
After a drawing by W.A. Herdman, 1882

Some solitary, sessile ascidians are stalked, and budding commonly occurs by growth at the base of the animal. In "social" colonial ascidians the zooids are relatively independent, whereas in "compound" colonial ascidians budding gives rise to a colony in which the zooids are embedded in a common tunic. Several zooids may share a single, common cloacal aperture through which water exits, but each zooid has its own branchial aperture through which water enters.

**Internal features.** *Skeleton, tissues, and muscles.* The tunic functions to some extent as an external skeleton that supports and protects the body. Additional support is provided by body fluids and connective tissue. Firm proteinaceous rods also may support the branchial apparatus.

Although musculature is poorly developed in tunicates, there are muscles that retract the body and constrict the atrial cavity, allowing it to eject water. In dolioloids and salps, these muscles have become modified so as to produce the characteristic jet propulsion.

*Nervous system and organs of sensation.* In the tadpole larvae and appendicularians, the dorsal nerve cord is well developed. At the anterior end there are usually sensory structures, which detect light and orient the animal to gravity. Similar sensory structures can be found in adult thaliaceans. Special organs of sense are otherwise poorly developed. When the larva metamorphoses into an adult, the original nervous system and sensory organs degenerate, leaving a single ganglion between the oral and atrial openings. Nerves grow to the various organs of the body from this ganglion.

Sensory structures

*Digestion, nutrition, and excretion.* In ascidians and thaliaceans the beating action of cilia creates a water current. As the water is driven from the branchial sac into the atrial cavity, a sheet of mucus, secreted by the endostyle, traps a variety of very small organisms suspended in the water current, especially small plants (phytoplankton). The mucus is rolled into a cord and then conveyed to the intestine, where it is digested and absorbed. A stomach and glands may be present. The intestine ends as an anus in the atrium below the atrial aperture. Wastes are ejected through this aperture in a stream of water.

Metabolic wastes, such as the breakdown products of protein, are excreted at various parts of the body, including the surfaces of the gills and the intestine, and sometimes by a discrete kidney. In many cases wastes are stored as solid deposits rather than being excreted from the body as they are produced.

*Respiration.* Gas exchange occurs in the gill and also across various other body surfaces, such as the lining of the atrium.

*Water/vascular system.* Tunicates do not have the well-developed secondary body cavity (coelom) of other chordates, but traces of one perhaps are represented by cavities around the heart and by an extension of the gut called the epicardium around some of the internal organs. The body cavities are considered to be a part of the circulatory system. There are a heart and some large blood vessels but no tiny capillaries. The tunicate heart is unusual in that it periodically reverses the direction in which it pumps the blood, but the reasons for this are unknown. There are many different cell types in the blood.

*Hormones.* A variety of possible endocrine organs help to coordinate feeding and reproduction. Various chemical substances are known to act as hormones in vertebrates; however, their exact role in the tunicates themselves is uncertain.

*Features of defense and aggression.* The tunic provides ascidians with some defense. They also may be protected by chemicals that make them distasteful to predators. Appendicularians are small and therefore difficult to see. If attacked, they can escape from the house and form a new one. Thaliaceans are protected somewhat by transparency and can evade some predators by quickly ejecting a jet stream of water. They have well-developed light-producing organs, which may help to deter predators.

### EVOLUTION AND PALEONTOLOGY

Because they are soft-bodied animals, tunicates have left little fossil record apart from the hard mineral particles, called spicules, that are found in the tunics of some species. A single lineage within the class Ascidiacea, or perhaps a lineage of ascidian-like tunicates that branched off prior to the common ancestor of the Ascidiacea, probably gave rise to the other two classes. Embryonic thaliaceans show indications of having been derived from attached colonies. The pyrosomes, which resemble the colonies of some ascidians, evidently branched off first within the class Thaliacea and may not even be related to the doliloids and salps. Appendicularians may have evolved from a more typical tunicate after early attainment of sexual maturity and the loss of the adult stage (*i.e.,* by paedomorphosis, retention of some juvenile features in the adult). There is some evidence that appendicularians are an early branch whose features are ancestral rather than specialized. Within the Ascidiacea, the common ancestor is generally thought to have been a solitary animal that did not reproduce by budding. The basis for this theory is that many ascidians do not bud, and the different patterns of budding that characterize distinct groups suggest independent origins. Evolution within the group has involved considerable elaboration of complex colonies, with the zooids themselves tending to become smaller and simpler in structure. There is a distinct trend toward parental care, especially in the colonial forms.

### CLASSIFICATION

**Annotated classification.**

**SUBPHYLUM TUNICATA** (or **UROCHORDATA**)
Chordates with notochord restricted to the tail and, except in Appendicularia, only in tadpole larva; body covered with a tunic containing cellulose; atrium, except in Appendicularia, present and opening dorsally; heart present; coelom reduced; no clear traces of segmentation; about 1,300 species.

**Class Ascidiacea** (sea squirts)
Fixed as adults, solitary or colonial, oral and atrial apertures usually directed away from substrate.

*Subclass Enterogona*
Gonads unpaired, either within or behind intestinal loop; body may be divided into thorax and abdomen.

*Order Aplousobranchia.* Gills simple, unfolded and without longitudinal vessels or bars; digestive tract and genital organs in posterior part of body.

*Order Phlebobranchia.* Gills with longitudinal vessels and bars, without folds; gonads on one side, near digestive tract.

*Subclass Pleurogona*
Gonads and digestive tract by side of gill.

*Order Stolidobranchia.* Gill with longitudinal vessels, folded.

**Class Appendicularia** (or **Larvacea**)
Adult small, pelagic, retaining larval notochord and tail; pharynx simple with two gill openings; no distinct atrium.

**Class Thaliacea**
Pelagic forms; atrial aperture directed toward the rear of each zooid; asexual buds form from a ventral stolon.

*Order Pyrosomida.* Zooids embedded in a tube open at one end.

*Order Doliolida.* Complex alternation of generations between a solitary, asexually and sexually reproducing gonozooid and colonial, asexually reproducing oozooids; gill with several to many stigmata.

*Order Salpida.* Complex alternation of generations between solitary, asexually reproducing oozooids and aggregated, sexually reproducing gonozooids. Pharynx leads to atrium by a single pair of slitlike openings.

**Critical appraisal.** The above classification only approximates a natural, or genealogical, system. It is ambiguous with respect to the relationships of the three classes. Some authors put the class Appendicularia together with the class Thaliacea as Pelagotunicata, suggesting one possible relationship. Within the class Ascidiacea the system suggests that the original condition was that of the order Phelebobranchia and that the gill became simplified in a second group (order Aplousobranchia) and more complicated in a third (order Stolidobranchia). The Phlebobranchia, therefore, are not a single lineage but a grade, with some lineages close relatives of Aplousobranchia, others close relatives of Stolidobranchia, and others perhaps early branches. A change from simple to complex gills is also possible. Within the class Thaliacea the orders Doliolida and Salpida perhaps have a closer common ancestry with each other than with the order Pyrosomida, but this is not clear from the arrangement. Older systems were artificial rather than genealogical and often put all of the colonial forms together, even though it was known that colonial structure evolved more than once.

## Cephalochordates

The cephalochordates, or lancelets, are small, fishlike marine invertebrates that probably are the closest living relatives of the vertebrates. There are about 20 species in two families, each with a single genus. *Branchiostoma* was formerly called *Amphioxus,* a name that is retained as an informal term (Figure 2). The other genus is *Epigonichthys,* also called *Asymmetron.* The genus *Asymmetron* is sometimes retained for some species.

### GENERAL FEATURES

**Size and range of diversity of structure.** Adult lancelets reach a length of about six to seven centimetres (2.5 inches). There is little structural diversity within the group, the main difference between the two families being the restriction of the gonads to one side of the body in *Epigonichthys.*

**Importance.** Although edible, lancelets are never sufficiently abundant to constitute a significant source of food to humans or an important part of the food chain in nature. Rather, their significance has to do with their place in evolution, as invertebrates transitional to vertebrates providing clues for the history of human lineage. This connection was first shown by the Russian zoologist Aleksandr Kowalevsky in 1867 in embryological evidence that was influential in establishing that evolution has in fact occurred. Lancelets have a structure that illustrates the characteristic features of chordates in simple form.

Evolutionary significance

### NATURAL HISTORY

**Reproduction and life cycle.** Lancelet sexes are separate, and asexual reproduction does not occur. Eggs and sperm are shed directly into the water, where fertilization occurs. The early stages of development strikingly resemble those of both tunicates and vertebrates. A larva is produced that is similar in structure to the adult but is peculiarly asymmetrical (the gill slits on one side develop first), smaller, and simpler, with fewer gill slits and no atrium. The larvae spend much of their time feeding in the open

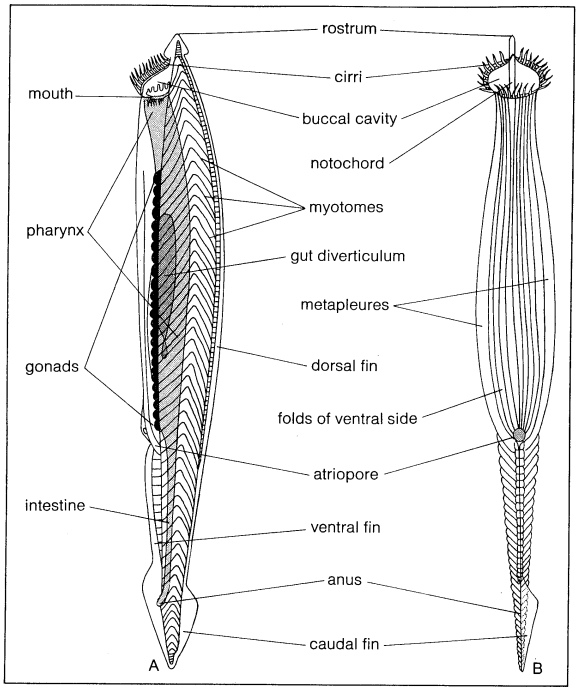



Figure 2: (A) Left lateral and (B) ventral views of the Amphioxus (*Branchiostoma lanceolatum*).
From P. Grasse, *Traite de Zoologie*, vol. 11 (1948); Masson & Cie, Paris

water but can be found on the bottom. After growing and developing, they metamorphose into the adult form and complete their life history in the substrate.

**Ecology and habitats.** Lancelets are distributed throughout the world along tropical and temperate coasts. They inhabit soft bottoms ranging from sand to coarse shelly sand or gravel in shallow coastal water. Lancelets lie buried beneath this substrate, often with their mouths protruding above the surface, allowing them to take in water laden with food. In China, lancelets are sometimes eaten and even support a small fishing industry.

**Food, feeding, and movement.** Lancelets can swim both forward and backward and can move rapidly through the gravel in which they live. Their behaviour is simple, largely being a matter of locating the proper habitat and escaping from predators. Larvae filter small organisms out of the water; at the time when they metamorphose into the adult, they also feed upon coarser materials deposited on the bottom. The adults filter small organisms from the overlying water by drawing a current into the mouth. The tentacle-like cirri around the mouth form a grid that keeps out sand and other large particles.

## FORM AND FUNCTION

**General features.** The lancelets are also called cephalochordates (Greek: *kephale,* "head") because the notochord extends from near the tip of the tail to well into the anterior of the body. Because they do not have the braincase, or cranium, of a vertebrate, lancelets are often called acraniates. The pharynx, with its many gill slits, is surrounded by the atrium, a large cavity with a single exit (the atriopore) on the lower surface of the body. The atrium protects the gills. Tunicates also have an atrium, but its evolution is probably independent of that of the cephalochordrate atrium.

Acraniates

The bodies of lancelets, like those of fishes and other vertebrates, are largely made up of serially repeated units (segments) that include blocks of muscles called metameres. This segmentation also extends to the nerves that supply the myotomes and to some body cavities, excretory structures, and other parts. Segmentation is thought to provide more effective body coordination during locomotion. The segments of vertebrates and cephalochordates are so similar that they were almost certainly present in the common ancestor of the two groups. Tunicates and hemichordates have no clear indications of ever having possessed segments. Segments occur in other animals, including annelid worms and arthropods, but these segments have a different composition and probably a separate evolutionary origin.

A distinct "secondary" body cavity (coelom), like that which contains the internal organs in vertebrates and many other animals, is well developed and forms a system of cavities and spaces. Like the coelom of hemichordates, echinoderms, and a few other animals, it develops as outpouchings in the gut of the embryo.

**External features.** Lancelets are steamlined animals. A dorsal fin extends along the upper surface of the body and continues as a caudal fin around a tail and as a ventral fin to an atrium on the lower surface. Paired fins are absent, but metapleural folds along the sides of the body suggest precursors of paired fins. The tip of the body projects slightly above and in front of the mouth, which is surrounded by a funnellike oral hood that bears the cirri. The anus opens well behind the atriopore, on the left side of the ventral fin. The general body surface is covered by a smooth cuticle layer.

Fins

**Internal features.** *Skeleton, tissues, and muscles.* The notochord extends virtually the entire length of the body and provides much of its support. It has a firm sheath and a core made up of a single series of cells that contains muscle fibres. These fibres probably maintain the stiffness of the notochord, the main role of which is keeping the body from shortening when the animal swims. The gills, fins, and cirri also contain stiff, supportive rods.

The main body musculature occurs in horizontal chevron-shaped blocks of muscle (myotomes) like those of fishes. This arrangement allows the muscles to pull more effectively in producing a side-to-side movement of the body in swimming. The remaining muscles are quite small and associated largely with feeding and the movement of internal organs.

*Nervous system and organs of sensation.* The cephalochordate nervous system is simple. The main nerve cord, which is single and hollow as in all chordates, has a slight swelling at the front that barely qualifies as a brain. Nerves from the main nerve cord occur in groups that roughly compare to those of vertebrates in arrangement and in the regions supplied. There are small eyelike organs in the nerve cord that can detect the direction of light and changes in its intensity. Various areas of the body surface, including some near the mouth, detect chemicals in the water and thereby aid in feeding.

*Digestion and excretion.* Lancelets are filter feeders that extract small particles suspended in the water. The mouth is covered by an oral hood, the edges of which form the buccal cirri. The cephalochordate commonly is buried in the substrate and positions its mouth above the surface of the sand. During feeding, the cirri form a kind of grid that keeps out large particles. Water is drawn into the mouth by the beating action of cilia on the gills. The pharynx is a large section of the gut just behind the mouth, extending about two-thirds the length of the body, with many narrow gill slits. The water current enters the pharynx, passes through the gill basket to the atrium, and leaves the body through the atriopore. On the floor of the pharynx, between the left and right series of gill slits, an endostyle secretes a sheet of mucus that moves upward along the gills and traps food particles suspended in the water current. The mucus is rolled up and transported to the intestine, where food is digested and absorbed. There is no distinct stomach. The intestine is straight, except for a blind outpouching called the caecum, which has, on the basis of position, been compared to the liver and pancreas of vertebrates. It extends forward along the right side of the pharynx.

Filter feeders

Lancelets have unique excretory structures called solenocytes, which occur only in some distantly related animals, such as annelids.

*Respiration.* The gill is largely a feeding organ, but it also serves for the exchange of gases in respiration. After the water has passed through the gill slits, it reaches the atrium and exits through the atriopore. Excretory products and eggs and sperm also exit the body through this opening.

Gill slits

*Circulatory system.* The general pattern of blood circu-

lation through vessels and tissues in cephalochordates is strikingly like that of vertebrates, although simpler. The most notable difference is that cephalochordates lack a heart. Blood is forced through the closed system by contractile blood vessels (especially one called the ventral aorta) and by blood vessels of the gills. Blood passes forward from the rear of the body to the ventral aorta, which is located beneath the endostyle, and then branches upward through vessels in the gills. Most of the blood then passes toward the rear of the animal, some of it moving through capillaries in the intestine and taking up food. From the posterior end of the body, blood passes forward and then makes a detour through capillaries in the caecum, much as it does through the liver of lower vertebrates, back to the ventral aorta. There are no corpuscles in the blood.

*Hormones.* The endostyle takes up iodine and forms thyroxine, an important hormone produced by the vertebrate thyroid gland. This homology is interpreted as a step in the evolution of the thyroid from the endostyle. It is not certain what role thyroxine plays in the physiology of the lancelets themselves, however.

### EVOLUTION AND PALEONTOLOGY

Soft-bodied animals such as lancelets rarely have a good fossil record. A few fossils have been interpreted as cephalochordates, but few of these determinations are well founded. A good possibility is *Pikaia,* a fossil discovered in the Burgess Shale (Middle Cambrian, about 530 million years old). *Pikaia* has myotomes and what looks like a notochord, indicating that it is a chordate, but only its shape suggests that it is a lancelet rather than a fish.

The cephalochordates make plausible models for the common ancestor of the chordates and for precursors to vertebrates. They evidently represent a collateral branch on the vertebrate lineage that has been somewhat modified since common ancestry, however, and should not be thought of as a human ancestor. Several features unique to cephalochordates and vertebrates suggest that they are "sister groups" more closely related to each other than either is to other chordates. These features include the segmented musculature and its innervation, the pattern of circulation, and several biochemical features. The atrium is thought to have evolved independently in cephalochordates and tunicates; hence there is little evidence for the two forming a single lineage.

Whether the ancestral chordate was more like a cephalochordate or a tunicate is debatable, because features absent in tunicates could mean that they never have been present or could mean that they have been lost. Some authors have argued that the simplicity of cephalochordates is due to degeneration, but there are no clear indications that this is true.

### CLASSIFICATION

**SUBPHYLUM CEPHALOCHORDATA**
Notochord extends anterior to dorsal nerve cord; atrium ventral; segments and coelom well developed; no heart; no paired fins.

**Family Branchiostomatidae**
Double row of gonads; *Branchiostoma.*

**Family Epigonichthyidae**
Gonads on right side of body only; *Epigonichthys.*

(M.T.G.)

## Vertebrates

The chordate subphylum Vertebrata is one of the best known of all groups of animals. Its members include the classes Agnatha, Chondrichthyes, and Osteichthyes (all fishes); Amphibia (amphibians); Reptilia (reptiles); Aves (birds); and Mammalia (mammals).

### GENERAL FEATURES

Characteristic features

Although the vertebral column is perhaps the most obvious vertebrate feature, it was not present in the first vertebrates, which probably had only a notochord. The vertebrate has a distinct head, with a differentiated tubular brain and three pairs of sense organs (nasal, optic, and otic). The body is divided into trunk and tail regions. The presence of pharyngeal slits with gills indicates a relatively high metabolic rate. A well-developed notochord enclosed in perichordal connective tissue, with a tubular spinal cord in a connective tissue canal above it, is flanked by a number of segmented muscle masses. A sensory ganglion develops on the dorsal root of the spinal nerve, and segmental autonomic ganglia grow below the notochord. The trunk region is filled with a large, bilateral body cavity (coelom) with contained viscera, and this coelom extends anteriorly into the visceral arches. A digestive system consists of an esophagus extending from the pharynx to the stomach and a gut from the stomach to the anus. A distinct heart, anteroventral to the liver, is enclosed in a pericardial sac. A basic pattern of closed circulatory vessels is largely preserved in most living forms. Unique, bilateral kidneys lie retroperitoneally (dorsal to the main body cavity) and serve blood maintenance and excretory functions. Reproductive organs are formed from tissue adjacent to the kidneys; this original close association is attested by the tubular connections seen in males of living forms. The ducts of the excretory organs open through the body wall into a cloacal chamber, as does the anus of the digestive tract. Reproductive cells are shed through nearby abdominal pores or through special ducts. A muscular tail continues the axial musculature of the trunk.

Distribution

Approximately 45,000 living species constitute the vertebrates. Species of several classes are found from the high Arctic or Antarctic to the tropics around the Earth; they are missing only from interior Antarctica and Greenland and from the North Polar ice pack. In size, vertebrates range from minute fishes to elephants and whales (of up to 100 tons), the largest animals ever to have existed. Vertebrates are adapted to life underground, on the surface, and in the air. They feed upon plants, invertebrate animals, and one another. Vertebrate faunas are important to humans for food and recreation.

### NATURAL HISTORY

In order to give a broad and comparative view of their life histories, the vertebrates are subdivided here into major groups based on morphology: the cyclostomes (jawless fishes), the chondrichthyes (cartilaginous fishes), the teleostomes (bony fishes), and the tetrapods.

Hagfishes and lampreys

**The cyclostomes.** The cyclostomes include two classes of living, jawless fishes (agnathous)—Petromyzontiformes (lamprey eels) and Myxiniformes (hagfishes). The hagfishes are totally marine, often living in deep waters associated with muddy bottoms. The lampreys may be marine as adults but spawn in fresh waters, where the larvae spend some time before metamorphosing to the adult. Some lampreys live entirely in fresh water and may change only slightly in habit as a result of metamorphosis. Without lateral fins, lampreys swim by undulations of the body and can control direction only for short distances.

The living agnaths are predatory, the lampreys being well known for attacking salmonoid fishes. The lamprey attaches to its prey using its round, suctorial mouth, and it rasps a hole through the outer tissues using a tongue armed with keratinized teeth. It suctions off bits of tissue, blood, and body fluids. The hagfishes feed somewhat similarly, but on a variety of prey—invertebrates (worms and soft-bodied forms) and dead fishes.

The lampreys produce small eggs, which develop directly into larvae that burrow into the muddy bottom of the stream. With its mouth at the surface of the mud, the larva filter feeds until large enough to metamorphose and swim off as a small adult. In contrast, the hagfishes produce relatively large encapsulated, yolky eggs up to two centimetres in length. When laid, these eggs attach to any available object by terminal hooks. The encased egg develops more or less directly into a miniature adult.

Cartilaginous fishes

**The chondrichthyes.** The sharks, rays, and chimaerids are usually marine, but some sharks have entered fresh waters (the Amazon) or even live there permanently (Lake Nicaragua). In size, sharks range from the whale shark, nearly 10 metres in length, to rather small species, three centimetres in length. They usually weigh 25 to 200 kilograms (55 to 440 pounds). Sharks are predatory animals. Some large shark species (basking and whale sharks) fil-

ter feed on small crustaceans. Herbivorus sharks are unknown. Sharks swim by undulations of the tail, but rays "fly" through the water by undulations of the pectoral fins. Most species occur in near-shore waters, but some range widely throughout the oceans. A few are found in deep water.

A few sharks produce live young (viviparous) after internal fertilization. The posterior angle of the male's pelvic fins are modified into a clasper, which acts as an intromittent organ in copulating with the female. Most sharks lay large yolky, encapsulated eggs with hooks for attachment. The young develop directly and begin life as miniature adults. The young that develop in the mother's uterus obtain nutrients from the large yolk sac until they are born alive. In a few cases, the uterine wall secretes nutrients.

The teleostome, or osteichthyian, fishes (those having an internal bony skeleton) can be divided into two groups: the subclasses Actinopterygii (ray-finned fishes) and Sarcopterygii (lobe-finned fishes). The latter group includes the lungfishes, which live in marshes, ponds, or streams, and are frequent air breathers. They lay fairly large eggs, with a limited amount of yolk, that are enclosed in jelly coats like those of an amphibian. The eggs develop into small fishes that feed on live prey. The larvae of the African lungfish have external gills to supplement oxygen intake.

Bony fishes

**The teleostomes.** Actinopterygian fishes are the common bony fishes of modern aquatic environments. They range in size from fishes that are only millimetres in size to those two or more metres (6.6 or more feet) in length, weighing 500 kilograms or more. Large species (sturgeons) are found in fresh waters (several other large species are found in the Amazon) as well as in marine environments. The diet may include plants, animals, and carrion. Most species are midwater swimmers, but many spend much time lying on the bottom. Tail, pectoral, and even dorsal fins are used in swimming. Reproduction in this group is by way of large numbers of small eggs, which produce small larvae or develop directly to the adult.

**The tetrapods.** The tetrapods live primarily on land and are rather similar in habit. Members include the amphibians, reptiles, birds, and mammals. Amphibians are widespread in the warmer parts of the continents, being absent only in the far north and in the Antarctic. Three orders are recognized: Candata (the salamanders), the frogs and toads (Anura, or Salientia), and the Apoda or Gymnophiona (caecilians). Modification takes many forms, from the moist glandular skin (some scale remnants persist in apodans) to the loss of many of the bones of the skull. Like their ancestors, amphibians are cold-blooded and tend to be aquatic or limited to moist surroundings. Salamanders are seemingly the least modified in body form. They do not actively pursue prey and at best are only marginal swimmers. In swimming or crawling, the salamander's body and tail undulate. Frogs and toads hop using hind-limb propulsion and the forelimbs as body props. This dominance of the hind limb in locomotion is best seen in swimming when the forelimbs are drawn back against the body. In contrast to the salamanders and frogs, the burrowing, wormlike apodans are without limbs.

Amphibians usually trap food using a tongue that can be shot out of the mouth, or they use the mouth itself to grasp and ingest food. There is great variation in foods; only the larvae of frogs and toads appear to be plant feeders, a specialization that is reflected in the highly modified jaws and guts of the tadpoles.

Amphibians have retained a simple egg cell with a gelatinous cover. The eggs are laid in ponds, streams, or even in damp places high in trees, usually in great numbers. Fertilized eggs develop into free-swimming larvae, which then metamorphose to adults, but in highly specialized forms.

Reptiles

The class Reptilia retains many of the structural characteristics of the ancestral amphibian. While most reptiles are carnivorous, feeding on other organisms, a few are herbivorous (*e.g.,* tortoises). As cold-blooded animals, reptiles tend to be limited to temperate and tropical areas, but, where found, they are relatively common, although not as large or conspicuous as birds or mammals. Most reptiles are terrestrial, but a few are aquatic. As basic tetrapods, reptiles move about by creeping or swimming in a fashion similar to amphibians. Some reptiles, however, can lift the body from the ground and run rapidly either in a quadrupedal or bipedal fashion. Reptiles lay relatively large, shelled eggs. In a few instances, the eggs and young are cared for by the female; in others, the young are born alive (ovovivipary).

Birds are warm-blooded, and, although most are capable of flight, others are sedentary and some are flightless. Like their relatives the reptiles, birds lay shelled eggs that differ largely in the amount of calcification (hardening) of the shell. The young are usually cared for in a nest until they are capable of flight and self-feeding, but some birds hatch in a well-developed state that allows them to begin feeding immediately or even take flight. The megapods lay their eggs in mounds of rotting vegetation, which supplies the heat for incubation. (Nesting activities similar to those of some birds are seen in the crocodilians.)

Mammals

The mammals range in size from tiny shrews or small bats weighing only a few grams to the largest known animals, the whales. Most mammals are terrestrial, feeding on both animal and vegetable matter, but a few are partially aquatic or entirely so, as in the case of the whales or porpoises. Mammals move about in a great variety of ways: burrowing, bipedal or tetrapedal running, flying, or swimming. Reproduction in mammals is usually viviparous, the young developing in the uterus, where nutritive materials are made available through an allantoic placenta or, in a few cases, a yolk sac. The fertilized egg develops directly into the adult. The monotremes (platypus and echidna) differ from other mammals in that they lay eggs which hatch, and the relatively undeveloped young are carried in a pouch or kept in a nest; the growing young lap up a milk nutrient fluid exuded from the belly of the mother.

## FORM AND FUNCTION

**External features.** The evolution of the notochord, dorsal nerve tube, and pharyngeal slits in chordate structure suggests improved swimming capability and probably greater ability to capture prey. Specialization in the vertebrate for the active capture of larger prey is evident both in the structure of the mouth and in the relatively simple structure of the pharynx, with its strong gill development. Specialization for feeding is again seen in the two basic groups of vertebrates, the agnathans and gnathostomes. Swimming adaptations are also numerous and involve variations both in body form and in medial fins and the two pairs of lateral fins.

**Internal features.** *The skeletal system.* Support and protection are provided by the exoskeletal and endoskeletal divisions of the skeletal system. The exoskeleton, when present, is basically protective but functions in tooth support in the mouth region. The endoskeleton protects the brain and spinal cord and assists primarily with locomotion in the trunk and tail regions. The endoskeleton begins as cartilage and may remain so or may develop into bone. The cartilaginous endoskeleton, found in the shark or chimaerid, is usually calcified so as to be stiffer and stronger. Bone is distinctive but highly variable; some types of bone contain cells, others do not, or the bone may be laminar, spongy, or arranged in sheathing layers around blood channels.

Exoskeleton and endoskeleton

*Tissues and muscles.* Tissue development in the vertebrate is unique in its complexity; tissues in the strict sense (defined as a mass or sheet of similar cells with a similar function), however, do not exist. The simplest situation is seen in the epidermis, but even here there is a layered system in which different cell types provide different functions (such as protection and secretion). The stratified epithelium of the vertebrate is highly characteristic of that group (a similar one is seen in only one invertebrate group, the class Chaetognatha).

Other tissues of the vertebrate are more complex than the epithelium. For example, skeletal muscle consists not only of striated muscle fibres but also of connective tissue, which binds it together and attaches it by way of tendons. This contractible tissue includes nerves and blood vessels and their contained blood. Skeletal muscles thus appear as simple organs, just as do the smooth muscles in the

wall of the gut or the iris muscles of the eye. Such unique histological complexity runs through the entire body of the vertebrate.

*Nervous system and organs of sensation.* The dorsal position, tubular structure, and epidermal origin of the central nervous system are definitive of the chordates, although some may see similarities with the hemichordates. The sensory structures are distinctive of the chordates and include the paired nasal, optic, and otic organs (along with the strongly differentiated head).

Sensory structures

The nasal vesicle is variously open to the environment, and its sensory cells, as chemical receptors, are not unlike those in the taste buds of the mouth. The eye is the most complex organ of the head and is a lateral outpocketing of the anterior end of the brain tube. Later it acquires a lens of epidermal origin. The act of focusing the eye (accommodation) shows extensive adaptive variation among the different groups of vertebrates.

The otic vesicle starts from a simple sac formed by the invagination of an ectodermal placode. These developmental changes also include the changes of innervation. Whereas the original structure was basically an equilibrium adaptation, other functions, such as an awareness of movement or the sensation of the proximity of prey, developed.

The lateral-line system of canals and sensory organs is a unique vertebrate feature. The elements of this system are found on the head as well as the body. This system is related to the ear and presumably at its origin served a similar function. This system is lost in terrestrial vertebrate forms.

*The digestive system.* The digestive system of the vertebrate is distinctive in its structure but not in its function. The mouth and pharynx can be considered as parts of this system; the latter as an expanded cavity in the head is unmatched in any other group. The stomach and gut have been discussed above.

Diverticulum

Presumably the original condition of the digestive glands was that of a ventral diverticulum which may have received the food mass into its cavity. This diverticulum, matched by the diverticulum seen in the amphioxus or the "intestine" of the tunicate, produced the secretions (bile-like and enzymes) of both liver and pancreas. Through time, the liver gradually differentiated from the pancreas. The size and separation of the liver from the gut suggest its separate blood and metabolic activities. The most obvious by-product of the liver, bile, necessitated the formation of a gall bladder and a duct connection with the gut. The pancreas, in contrast, continued to produce digestive enzymes, but its secretory cells were no longer in direct contact with the food mass. Because the pancreas was only a partial source of intestinal digestive enzymes, it was sometimes reduced in size and enclosed in the gut wall itself (agnaths) or dispersed as tiny bits of tissue in the mesentary supporting the gut (actinopterygians).

*The excretory system.* The excretory system is unique in its nephrons, which filter the blood in the glomeruli and remove a variety of wastes from the body through selective secretion and reabsorption. In the shark or the coelacanth *Latimeria,* urea is used to raise the osmotic pressure of the blood to that of the marine habitat, thus saving these organisms considerable metabolic energy. The large intestine (sometimes centred in a rectal gland) acts as an auxiliary excretory organ, as do also the gills of fishes or the sweat glands of mammals.

*Respiration and gas exchange.* Respiration, like excretion, involves specialized body structures, such as lungs or gills, but also can involve other areas, such as the skin itself. Respiration involves exchange of gases both between the body of the organism and the environment and between the blood system and the body tissues. It also involves cellular respiration where oxygen is used and carbon dioxide is produced. There is nothing characteristic of the vertebrate in this functional area; even the hemoglobin of the blood is suggested in the respiratory pigments of other animals.

*The circulatory system.* The circulatory system of vertebrates is closed in that fluids course through vessels, but there is free movement of cells in and out of blood. Some leukocyte (white blood cell) movement out of the capillaries and fluid leakage are observed in all tissues. Blood tissues are distinctive in the range of specialized cells, although these vary in detail among animals. The immune function of the blood is best developed in the vertebrate.

*The endocrine system.* The endocrine system is characterized by its separate organs. The occurrence of a pituitary or a thyroid gland is suggestive of the evolutionary change and specialization that took place within this group. The relatively unspecialized nature of some parts of this system is seen in certain scattered cells in the gut wall or even the clumps of islet cells of the pancreas.

## EVOLUTION AND PALEONTOLOGY

The knowledge of vertebrates as revealed by fossils has grown rapidly during the past few decades, but there is much still to be discovered. The ancestral vertebrate (protovertebrate) has been sought for more than 100 years, and the likelihood of finding it today is not much greater than in the past. It can be assumed that the protovertebrate was small and soft-bodied, two factors that suggest the improbability of finding a fossilized form in a recognizable condition. There are Cambrian fossils that have been suggested to be fossil cephalochordates and there are scales of agnath fishes, but the first type of fossil is too simple and the second already too complex to explain the transition.

(M.T.J.)

## CLASSIFICATION

**Annotated classification.**

SUBPHYLUM VERTEBRATA (or CRANIATA)
Bilaterally symmetrical; internal skeletal support with skull enclosing a highly developed brain and a vertebral column and nerve cord; paired, jointed appendages; skin; advanced organ systems; sense organs concentrated in head.

**Class Agnatha** (hagfishes, lampreys)
Primitive; jawless; paired fins are poorly developed or lacking; rasping tongue; notochord without bone; skin is soft, glandular, and slimy; true gill arches absent; marine habitat.

**Class Placodermi** (placoderms)
†Extinct; fishlike; jaws supported by both cranium and hyoid arch (amphistylic); partly ossified cranium; primitive; head and trunk have armour that is jointed at the neck; pelvic fins present or absent; pectoral fins or finlike structures often present; gill arches.

**Class Chondrichthyes** (sharks, rays, and skates)
Cartilaginous fishes; jaws; paired fins; no swim bladder; pelvic fins in males often modified to form claspers; gill arches internal to gills; reduced notochord; lateral-line system; paired nostrils; internal nares absent; separate sexes; internal fertilization and direct development; oviparous, ovoviviparous, or viviparous.

*Subclass Elasmobranchii* (sharks and rays)
Numerous teeth derived of placoid scales; 5 to 7 gill clefts; operculum absent; cloaca; upper jaw not fused with braincase; dorsal fin nonerectile; with spiracles; worldwide distribution.

*Subclass Holocephali* (chimeras)
Teeth fused to bony plates; no scales; 4 gill pairs under 1 gill opening on each side; no cloaca; no spiracles; operculum present; upper jaw fused to braincase; dorsal fin erectile; whiplike tail; claspers present in males; temperate marine freshwater.

**Class Osteichthyes** (bony fishes)
Jaws; partly or fully ossified skeleton; usually a swim bladder; paired fins; gills covered by a bony operculum; scales; paired nostrils with or without internal nares; lateral-line system; mostly oviparous with external fertilization; some ovoviviparous or viviparous.

*Subclass Actinopterygii* (ray-finned fishes)
Generally lack choanae; no fleshy base to paired fins; no internal nares; air sacs usually function as swim bladder; skeleton usually well ossified.

*Subclass Sarcopterygii* (lobe-finned fishes)
Usually possess a choana; paired fins with a fleshy base over a bony skeleton; persisting notochord; 2 dorsal fins; nares are internal.

**Class Amphibia**
Cold-blooded; respire by lungs, gills, skin, or mouth lining; larval stage in water or in egg; skin is usually moist with mucous glands and without scales; tetrapods; freshwater and terrestrial; paired appendages are legs; 10 pairs of cranial nerves; separate sexes; external fertilization with development into tadpole larvae; some have internal development, ovoviviparous or viviparous.

*Order Aponda* (or *Gymnophiona;* caecilians). Wormlike; no limbs or girdles; compact skull; lidless, minute eyes; persistent notochord; tail; scales present in some species.

*Order Anura* (or *Salientia;* frogs and toads). Tailless; elongated hind limbs modified for jumping; larvae lack true teeth and external gills.

*Order Caudata* (or *Urodela;* salamanders). Tail; limbs normal; many skeletal elements cartilaginous; larvae with true teeth and external gills.

**Class Reptilia**

Cold-blooded; no larval stage; breathing by lungs; well-ossified skull; dry skin; scales; no glands; 5-toed limbs; claws; 3- or 4-chambered heart with incomplete ventricle separation; 12 pairs of cranial nerves; internal fertilization, direct development; oviparous and ovoviviparous.

*Subclass Anapsida* (turtles, tortoises, terrapins)

No temporal skull openings; body encased in bony shell; no teeth in living members; oviparous.

*Subclass Lepidosauria*

No bipedal specializations; 2 complete temporal openings; complete palate; oviparous; male is without penis.

*Subclass Archosauria* (ruling reptiles)

Some ancient forms had bipedal locomotion; longer hind legs; semiaquatic; webbed feet; teeth in sockets; single penis; oviparous; includes extinct dinosaurs.

*Subclass Synaptosauria*

†Extinct; single temporal opening on area of cheek.

*Subclass Ichthyopterygia*

†Extinct; temporal openings high up on skull; fishlike; spindle-shaped body; high tail fin; triangular dorsal fin; paddlelike legs; marine.

*Subclass Synapsida*

†Extinct; mammallike; lateral temporal opening.

**Class Aves**

Warm-blooded; skull has only 1 condyle; front limbs primarily modified for flight; hind limbs are legs with 4 or fewer toes; body covered with feathers; scales on feet; 4-chambered heart; no teeth; horny beak; lungs with extended air sacs; 12 pairs of cranial nerves; internal fertilization; oviparous.

*Subclass Archaeornithes*

†Extinct; teeth in both jaws; long, feathered tail; less specialized for flight; body elongated and reptilelike; forelimb had 3 clawed digits; small brain and eyes; nonpneumatic bones.

*Subclass Neornithes* (true birds)

Well-developed sternum; tail is not long; no teeth; forelimbs modified to wings; teeth replaced by horny rhamphoteca over bill.

**Class Mammalia**

Warm-blooded; mammary glands; lower jaw is composed of 1 bone; hair; advanced brain; skin with different glands and hair; ears with 3 middle-ear bones; 12 pairs of cranial nerves; 4-chambered heart; young nourished by milk from mammary gland; internal fertilization; mostly viviparous, some oviparous.

*Subclass Prototheria*

Primitive; egg-laying; hair; mammary glands without nipples; pectoral girdle; separate oviducts that open into cloacal chamber that is shared with excretory ducts; oviparous.

*Subclass Theria*

Mammary glands with nipples; functional teeth; oviducts partly fused; with or without a cloaca; uterus and vagina; viviparous. (Ed.)

**Critical appraisal.** The classification of animals is presently in a state of flux. The classification presented here is traditional and conservative. Because traditional theories of taxonomy tend to be nonquantitative, various interpretations of relationships or patterns can be presented and defended.

The alternative cladistic style of taxonomy is an attempt to force taxonomy into a testable, highly objective operation. One tentative classification based in cladistics separates the vertebrates into two superclasses (Agnatha and Gnathostomata). Agnathans are jawless, while the gnathostomates encompass the remainder of the jawed vertebrates. Living agnathans are placed in the class Cyclostomata. Gnathostomates can be further divided into the epiclasses Elasmobranchiomorphi (sharks and rays) and Teleostomi (bony fishes and tetrapods). The former group are identified primarily by a cartilaginous skeleton, while the latter group have developed a bony skeleton. Two subepiclasses of the teleostomes are Ichthyopterygii (or Osteichthyes; bony fishes) and Cheiropterygii (tetrapods), the latter being further divided into the classes Amphibia, Reptilia, Aves, and Mammalia.

Although this classification includes and uses traditional taxonomic categories, their position in the hierarchy may be changed. Separation of agnath and gnathostome is opposed by those cladists who chart the origin of gnathostomes from the agnath, believing that the differences in mouth and tooth structure are a result of modification. The Gnathostomata is subdivided into the Elasmobranchiomorphi and the Teleostomi largely on the basis of mouth and tooth structure. The creation of epiclasses and subepiclasses in the alternative classification is not important in itself; the creation of a dichotomy between Ichthyopterygii and Cheiropterygii, however, is important, although from the evolutionary view it is evident that the one evolved from the other. (M.T.J.)

**BIBLIOGRAPHY**

*Chordates:* E.J.W. BARRINGTON, *The Biology of Hemichordata and Protochordata* (1965), is an account of the lower chordates and their evolution; N.J. BERRILL, *The Origin of Vertebrates* (1955), argues the thesis that the urochordate larva represents the prototype from which cephalochordates and vertebrates are derived; A. WILLEY, *Amphioxus and the Ancestry of the Vertebrates* (1894), an early but good comprehensive account, presents the orthodox theory of chordate relationships; R.P.S. JEFFERIES, *The Ancestry of the Vertebrates* (1986), expounds an alternate theory of chordate origin; and LIBBIE H. HYMAN, *The Invertebrates,* vol. 5, *Smaller Coelomate Groups* (1959), is a classic work treating the hemichordates in extensive detail. Later works include CHARLES K. WEICHERT and WILLIAM PRESCH, *Elements of Chordate Anatomy,* 4th ed. (1975); and R. MCNEILL ALEXANDER, *The Chordates,* 2nd ed. (1981); supplemented by BRIAN BRACEGIRDLE and PATRICIA H. MILES, *An Atlas of Chordate Structure* (1978).

*Protochordates:* N.J. BERRILL, *The Tunicata with an Account of the British Species* (1950, reprinted 1968), a taxonomic survey with a useful section on tunicate biology; PIERRE P. GRASSE (ed.), *Traité de zoologie: anatomie, systématique, biologie,* vol. 11, *Echinodermes, stomocordés, procordés* (1966), an advanced zoological treatise devoted to protochordates, with good illustrations; W.A. HERDMAN, "Tunicata (Ascidians and Their Allies)" and "Cephalochordata," in *The Cambridge Natural History,* vol. 7, pp. 35–138 (1904), an important general account; R.N. MILLAR, *The Marine Fauna of New Zealand: Ascidiacea* (1982), a morphological account of a single species of ascidian; WILLARD G. VAN NAME, "The North and South American Ascidians," *Bulletin of the American Museum of Natural History,* vol. 84 (1945). For a later treatment, see "Invertebrate Chordates: Tunicates and Lancelets," in VICKI PEARSE *et al., Living Invertebrates* (1987). (M.T.G.)

*Vertebrates:* CHARLES G. CRISPENS, JR., *The Vertebrates, Their Forms and Functions* (1978); J.Z. YOUNG, *The Life of Vertebrates,* 3rd ed. (1981); LIBBIE H. HYMAN, *Hyman's Comparative Vertebrate Anatomy,* 3rd ed., edited by MARVALEE H. WAKE (1979); EDWIN H. COLBERT, *Evolution of the Vertebrates: A History of the Backboned Animals Through Time,* 3rd ed. (1980); ALFRED SHERWOOD ROMER and THOMAS S. PARSONS, *The Vertebrate Body,* 6th ed. (1986); LEONARD B. RADINSKY, *The Evolution of Vertebrate Design* (1987); ROBERT L. CARROLL, *Vertebrate Paleontology and Evolution* (1988); and F. HARVEY POUGH, JOHN B. HEISER, and WILLIAM N. MCFARLAND, *Vertebrate Life,* 3rd ed. (1989). (M.T.J.)

# Christianity

Founded in the 1st century AD by Jesus of Nazareth (the Christ, or the Anointed One of God), Christianity has become the largest of the world's religions. Geographically the most widely diffused of all religions, it has a constituency of more than 1,000,000,000. Its largest groups are the Roman Catholic Church, the Eastern Orthodox churches, and the Protestant churches; in addition to these churches there are several independent churches of Eastern Christianity as well as numerous sects throughout the world. See also EASTERN ORTHODOXY; ROMAN CATHOLICISM; PROTESTANTISM.

This article first considers the nature and development of the Christian religion, its ideas, and its institutions. This is followed by an examination of several intellectual manifestations of Christianity. Finally, the position of Christianity in the world, the relations among its divisions and denominations, its missionary outreach to other peoples, and its relations with other world religions are discussed. For supporting material on various topics, see BIBLICAL LITERATURE; DOCTRINES AND DOGMAS, RELIGIOUS; JESUS; RITES AND CEREMONIES, SACRED; and SACRED OFFICES AND ORDERS.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 827.

The article is divided into the following sections:


The church and its history 251
- The essence and identity of Christianity 251
  - Historical views of the essence
  - The question of Christian identity
- The history of Christianity 256
  - The primitive church
  - The internal development of the early Christian church
  - Relations between Christianity and the Roman government and the Hellenistic culture
  - The early liturgy, the calendar, and the arts
  - The alliance between church and empire
  - Theological controversies of the 4th and 5th centuries
  - Popular Christianity in the late empire
  - Liturgy and the arts after Constantine
  - Political relations between East and West
  - Literature and art of the "Dark Ages"
  - Missions and monasticism
  - The Photian schism and the great East–West schism
  - From the Schism to the Reformation
  - Modern Christianity
  - The modern denominations
- Christian doctrine 272
  - The meaning of dogma
  - God the Father
  - God the Son
  - God the Holy Spirit
  - The Holy Trinity
  - The doctrine of man
  - The church
  - Last things
- Church year 302
  - Origins of the church year
  - The major church calendars
  - History of the church year
  - Liturgical colours
- Canon law 308
  - Nature and significance
  - History
  - The second Vatican Council and post-conciliar canon law
  - Anglican canon law

Aspects of the Christian religion 315
- Patristic literature 315
  - The ante-Nicene period
  - The post-Nicene period
  - The character of the heritage
- Christian philosophy 323
  - History of the interactions of philosophy and theology
  - Faith and reason
  - Christian philosophy as natural theology
  - Contemporary discussions
- Christian mysticism 330
  - History of Christian mysticism
  - Stages of Christian mysticism
  - Forms of Christian mysticism
  - Significance of Christian mysticism
- Christian myth and legend 335
  - Characteristics of Christian myth and legend
  - History of Christian myth and legend

The Christian community and the world 340
- The relationships of Christianity 340
  - Church and state
  - Church and society
  - Church and education
  - Church and the arts
  - Church and social welfare
  - Church and minorities
  - Church and family
  - Church and the individual
- Christian missions 353
  - Biblical foundations
  - The history of Christian missions
- Ecumenism 358
  - The biblical perspective
  - The history of ecumenism
- The Christian church and non-Christian religions 362
  - Conflicting Christian attitudes
  - Modern views

Bibliography 363


## THE CHURCH AND ITS HISTORY

### The essence and identity of Christianity

At the very least, Christianity is the faith tradition that focuses on the figure of Jesus Christ. In this context, faith refers both to the believers' act of trust and to the content of their faith. That tradition, viewed as a system of belief and behaviour, leads people to see Christianity as one of the world religions, alongside Hinduism, Buddhism, Islām, and others.

As a tradition, Christianity is more than a system of religious belief. It also has generated a culture, a set of ideas and ways of life, practices, and artifacts that have been handed down from generation to generation through the 20 centuries since Jesus first became the object of faith. Christianity is thus both a living tradition of faith and the culture that the faith leaves behind as a kind of deposit. The agent of Christianity is the church, the community of people who make up the body of believers. Christianity may incorporate, along with such believers, their doctrines, customs, and historical episodes.

To say that Christianity "focuses" on Jesus Christ is to say that whatever else it comprehends, somehow it brings these realities together in reference to an ancient historic figure. Few Christians would be content to keep this reference merely historical. Although their faith tradition is historical—*i.e.,* they believe that transactions with the di-

Jesus Christ as the focus of Christianity

vine do not occur in the realm of timeless ideas but among ordinary humans through the ages—the vast majority of Christians focus their faith in Jesus Christ as someone who is also a present reality. They may include many other references in their tradition and thus may speak of "God" and "human nature" or of "church" and "world," but they would not want to be nor would they be called Christian if they did not bring their eyes and attentions first and last to Jesus Christ.

While there is something simple about this focus on Jesus as the central figure, there is also something very complicated. That complexity is apparent when one tries to envision the more than 22,000 separate churches, sects, and denominations that make up the Christian faith tradition today. To project these separate bodies against the background of their development in the nations of the world is to suggest the bewildering variety. To picture people expressing their adherence to that tradition in their prayer life and cathedral-building, in their quiet worship or their strenuous efforts to change the world, is to suggest even more of the variety.

Given such complexity, it is natural that through the ages both those in the tradition and those surrounding it have made attempts at simplification. Two ways to do this have been to concentrate on the "essence" of the faith, and thus on the ideas that are integral to it, or to be concerned with the "identity" of the tradition, and thus on the boundaries of its historical experience.

Scholars in the modern world have tended to locate the focus of this faith tradition in the context of monotheistic religions. Christianity addresses the historical figure of Jesus Christ against the background of, and while seeking to remain faithful to, the experience of one God. It has consistently rejected polytheism, which allows for many gods, and atheism, which makes Jesus a purely and ordinarily human figure without divine or transcendent reference.

To monotheism as an element of the faith tradition of Christianity one may add that, with rare exceptions, Christianity refers to a plan of salvation or redemption. That is to say, the believers in the church picture themselves as in a plight from which they need rescue. For whatever reason, they have been distanced from their source in God and need to be saved. Christianity is based on a particular experience or scheme directed to the act of saving—that is, of bringing or "buying back," which is part of what redemption means, these creatures of God to their source in God. The agent of that redemption is Jesus Christ.

It is possible that through the centuries the vast majority of believers have not used the term essence to describe the central focus of their faith. The term is itself of Greek origin and thus represents only one part of the tradition, one element in the terms that have gone into making up Christianity. The search for an essence may be more urgent for philosophers, theologians (who interpret the language of the believing community), or historians than it is for the regular believers who do not share the burden of scholars. Essence refers to those qualities that give something its identity and are at the centre of what makes that thing different from everything else. To Greek philosophers it meant something intrinsic to and inherent in a thing or category of things, which gave it its character and thus separated it from everything of different character. Thus Jesus Christ belongs to the essential character of Christianity and gives it identity in the same way that Buddha does for Buddhism.

The definition of essence

If the mass of people do not have the scholar's problem of defining the essence of Christianity, in practice they must come to terms with what the word essence implies. Whether they are engaged in being saved or redeemed on the one hand, or thinking and speaking about that redemption, its agent, and its meaning on the other, they are concentrating on the essence of their experience. Those who have concentrated from within the faith tradition have also helped to give it its identity. It is not possible to speak of the essence of a historical tradition without referring to how its ideal qualities have been discussed through the ages. Yet one can take up the separate subjects of essence and identity in sequence, being always aware of how they interrelate.

## HISTORICAL VIEWS OF THE ESSENCE

**Early views.** The earliest members of the Christian faith tradition were Jews, as was Jesus himself, and thus they stood in the faith tradition inherited by Hebrew people in Israel (and lands to which they had been taken as captives in exile). They were monotheists, devoted to the God of Israel. When they made claims that Jesus was divine, it was part of their task to make their witness in ways that would not challenge monotheism.

Insofar as they began to separate or be separated from Judaism, which did not accept Jesus as Christ, these earliest Christians not only experienced salvation but also expressed certain ideas about the one on whom their faith focused. As with other religious people, they became involved in a search for truth. God, in the very nature of things, was necessarily the final Truth. But an early reference, preserved in the Gospel According to John, finds Jesus referring to himself not only as "the way" and "the life" but also as "the Truth." Roughly, this meant "all the reality there is" and was a reference to Jesus' participation in the reality of the one God.

From the beginning there were Christians who may not have seen Jesus as the Truth, or as a unique participant in the reality of God. There have been "humanist" devotees of Jesus, modernist adapters of the truth about the Christ; but even in the act of adapting him to humanist concepts in their day they have contributed to the debate of the essence of Christianity and brought it back to the issues of monotheism and a way of salvation.

Some believers and some scholars have always determined that the best way to preserve the essence of Christianity is to look at the earliest documents—the four Gospels and the letters that make up much of the New Testament—which tell whatever is believed to be known with any kind of assurance about what the earliest Christians remembered, taught, or believed about Jesus Christ. It is presumed that "the simple Jesus" and the "primitive faith" emerge from these documents as the core of the essence.

The New Testament as a source for the essence

Other believers and other scholars, however, have disturbed this simple notion of finding the essence by going back to the beginnings. The writings that make up the New Testament themselves reflect Jewish and Greek ways of thinking about Jesus and God. They are seen through the experience of different personalities, such as the Apostle Paul or the nameless composers of documents that came to be edited as the Gospels. Indeed, there are not only diverse ways of worship, of polity or governance of the Christian community, and of behaviour pictured or prescribed in the New Testament but also diverse theologies, or interpretations of the heart of the faith. Most believers see these diversities as complementing each other and leave to scholars the argument that the primal documents may compete with and contradict each other. Yet there is a core of ideas that all New Testament scholars and believers would agree are central to ancient Christian beliefs. One British scholar, James G. Dunn, for example, says they would all agree that "the Risen Jesus is the Ascended Lord." That is to say, there would have been no faith tradition and no scriptures had not the early believers thought that Jesus was "Risen," raised from the dead, and, as "Ascended," somehow above the ordinary plane of mortal and temporal experience. From that simple assertion early Christians could begin to complicate the search for essence.

An immediate question was how to combine the essential focus on Jesus with the essential monotheism. Some of the writers of the New Testament and more of the Apologists, late 1st- and 2nd-century reflectors on the meaning of this faith in both the Jewish and Greek contexts, saw Jesus as the "preexistent Logos." That is, before there was a historical Jesus born of Mary and accessible to the sight and touch of Jews and others in his own day, there was a Logos—a principle of reason, an element of ordering, a "word"—that participated in the Godhead and thus existed, but which only preexisted as far as the "incarnate" Logos, the word that took on flesh and humanity (John 1:1–14), was concerned.

Monotheism and the divinity of Jesus

In searching for an essence of truth and the way of salva-

tion, some primitive Jewish Christian groups, such as the Ebionites, and occasional theologians in later ages spoke in terms of what might be called a metaphor of adoption. These theologians used as their source certain biblical passages (*e.g.,* Acts 2:22). Much as an earthly parent might adopt a child, so the divine parent, the one Jesus called *abba,* "daddy" or father, had adopted him and taken him into the heart of the nature of what it is to be God. There were countless variations of themes such as the preexistent Logos or the concept of adoption, but they provide some sense of the ways the early Apologists carried out their task of contributing to the definition of the essence of their Jesus-focused yet monotheistic faith.

While it is easier to point to diversity than to simplicity or clarity among those who early expressed faith, it must also be said that from the beginning the believers insisted that they were—or were intended to be, or were commanded and were striving to be—united in their devotion to the essence of their faith tradition. There could not have been many final truths and there were not many legitimate ways of salvation. It was of the essence of their tradition to reject other gods and other ways, and most defining of essence and identity occurred as one set of Christians was concerned lest others might deviate from the essential faith and might, for example, be attracted to other gods or other ways.

While Jesus was among his disciples and those who ignored or rejected him, to make him the focus of faith or denial presented one type of issue. After the "Risen Jesus" had become the "Ascended Lord" and was no longer a visible physical presence, those at the head of the tradition had a different problem. Jesus remained, as was said, a present reality to them, and when they gathered to worship they believed that he was "in the midst of them." He was present in their minds and hearts, in the spoken word that testified to him, and also present in some form when they had their sacred meal and ingested bread and wine as his "body and blood." They created a reality around this experience; if once Judaism was that reality, now "Christianism," or Christianity, resulted.

The search for the essence of Christianity necessarily led people in the Greek world to concentrate on ideas. The focus on Jesus narrowed to ideas, to "beliefs about" and not only "belief in," and to doctrines. The essence began to be cognitive, referring to what was known, or substantive. This was most pronounced when people in the idea-centred Greek culture had to grasp through the mind the reality of someone who was not a visible presence.

As debates over the cognitive or substantive aspects of Jesus' participation in God became both intense and refined, the pursuit of essences became almost a matter of competition in the minds of the Apologists and the formulators of doctrines in the 3rd through the 6th centuries.

Confessions and creeds as the essence

During this time Christians met in council to develop statements of faith, confessions, and creeds. The claimed essence was used in conflict and rivalry with others. Christian Apologists began to speak, both to the Jews and to the other believers in the Greco-Roman world, in terms that unfavourably compared their religions to Christianity. The essence also came to be a definer of who had the best credentials and was most faithful. The claim that one had discerned the essence of Christianity could be used to rule out the faithless, the apostate, or the heretic. The believers in the essential truth and way of salvation saw themselves as insiders and others as outsiders. This concept became important—and, its victims would say, dangerous—after the Christian movement had triumphed in the Roman Empire, which became officially Christian. To fail to grasp or to misconceive what was believed to be the essence of faith might mean exile, harassment, or even death.

In the move from their exclusive roots in Judaism to their experience in Greek and Roman cultures, Christians did something rare if not unique in the history of religion: they adopted the entire scriptural canon of what they now saw to be another faith, Judaism, and embraced the Hebrew Scriptures as what they called the Old Testament. But while doing so, they also incorporated the insistent monotheism of Judaism as part of the essence of their truth and way of salvation, just as they incorporated the Hebrew Scriptures' story as part of their own identity-giving narrative and experience.

This narrowing of focus on Jesus Christ as truth meant also a complementary sharpening of focus on the way of salvation. There is no purpose in saving someone who does not need salvation. Christianity therefore began to make, through its councils and creeds, theologians and scholars, some attempts at definitive descriptions of what it is to be human. Later some of these descriptions were called "original sin," the idea that, inherited from Adam, the first-created human, all mortals carried a condition that made it impossible for them to be perfect or to please a personal God on their own. While Christians never agreed on a specific teaching on original sin, they did describe as the essence of Christianity the fact that something limited humans and led them to need redemption. Yet the concentration always returned to Jesus Christ as belonging more to the essence of Christianity than did any statements about the human condition.

The essence of Christianity eventually included statements about the reality to God. Christians inherited from the Jews a relatively intimate picture of a God who made their young and small universe, with its starry heavens, and then carried on discourse with humans, making covenants with them and rewarding or punishing them. But the Greek part of their tradition contributed the concept of a God who was greater than any ideas of God but who had to be addressed through ideas. Indeed, it was during this time that words such as essence, substance, and being—terms that did not belong to the Old or New Testament traditions—came to be wedded to biblical witness in the creeds. Christians, it might be said, used the vocabulary and repertory of options then available to them in speaking of the all-encompassing and the ineffable and grafted these onto the witness to God that was essential to their faith. Modern Christians, including many who reject the notion of creeds or any non-biblical language, are still left with the problems and intentions of the ancients: how to think of Jesus in such a way that they are devoted to him not in isolation, as an end in himself—for that would be idolatry of a human—but in the context of the total divine reality.

It is impossible to chronicle the efforts at expressing essence without pointing to diversity within the unity. Yet the belief in final unity belongs to any claims of finding an essence. Thus it was both a typical and a decisive moment when in the 5th century Vincent of Lérins, a Gallic theologian, provided a formula according to which Christianity expressed a faith that "has been believed everywhere, always, and by all" (*quod ubique, quod semper, quod ab omnibus creditum est*). Even if not all Christians could agree on all formulations, it was widely held that there was some fundamental "thing" that had thus been believed.

The formula of Vincent of Lérins

**Medieval and Reformation views.** For a thousand years, a period that began with what some historians called "Dark Ages" in the Christian West and that endured through both the Eastern and Western extensions of the Roman Empire, the essence of Christian faith was guarded differently than it had been in the first three centuries, before Christianity became official. By the 5th century, and with rare challenges until the 15th century, in the West the bishop of Rome, as pope and vicar of Christ, became the final custodian of the doctrinal essence of Christianity and could back support of his doctrine with the legal arm of the empire, the sword. In the Eastern churches no single pontiff ruled over the bishops, but they saw themselves just as surely and energetically in command of the doctrines that made up the essence of Christianity.

The Western drama was more fateful for Christianity in the modern world. The pope and the bishops of Roman Catholicism progressively saw themselves as determiners of the essence through doctrines and canons that enhanced the ancient grasp of faith. As they came to dominate in Europe, they allowed little room for those who did not agree with them. Jews were confined to ghettos, segregated and self-segregated enclaves where they did not and could not share the full prerogatives of Christendom, the "dominion" that now housed Christianity. When sects

Official doctrines as the essence

that were defined as heretical in their dissent—Waldenses, Albigenses, Cathari, and others—emerged to counter or contradict Roman Catholic concepts of Christian essence, they had to go into hiding or were pushed into enclaves beyond the enforcing reach of the custodians of official teaching. The essence of Christianity had become a set of doctrines and laws articulated and controlled by a hierarchy that saw those doctrines as a divine deposit of truth. Theologians might argue about the articulations with great subtlety and intensity, but in that millennium few would have chosen to engage in basic disagreement over the official teachings, all of which were seen to be corollaries of the basic faith in Jesus Christ as participating in the truth of God and providing the way of salvation.

When speaking of popes, bishops, councils, and theologians, it is necessary to make another distinction along with that which saw a narrowing of focus. Through these centuries there was also increasing differentiation between the official clergy and the wider body of believers. Most of what was debated centuries later about the essence of medieval Christianity came from the records of these authorities. As more is learned about the faith of the ordinary believers, it becomes more evident in the records of social history that people offered countless variations on the essence of the faith. Some historians point to survivals of European pagan customs and interpretations, which the later Protestant reformers saw to be threats to divine truth and the Jesus-centred way of salvation. Many people used the church's officially legitimated faith in the power of saints' relics to develop patterns of dealing with God that, according to the Reformers, detracted from the uniqueness of Jesus Christ as the only agent of salvation.

During this thousand years in both Western and Eastern Christianity, when the faith had a cultural monopoly, there was an outburst of creativity and a fashioning of a Christian culture that greatly enhanced and complicated any once-simple notions of an essence. Christianity was as much a cultural tradition as it was a faith tradition, an assertion that the leadership of the medieval church would not have regarded as diminishing or insulting.

As Christian culture grew ever more complex, however, there arose a constant stream of dissenters and individual reformers who tried to get back to what they thought was the original essence. Typical among these was St. Francis of Assisi, who in his personal style of devotion and simple way of life was often seen as capturing in his person and teachings more of the original essence of Jesus' truth and way of salvation than did the ordained authorities in the church and empires.

Out of this cluster of dissenters came late medieval reformers such as Jan Hus in Bohemia, John Wycliffe in England, and Girolamo Savonarola in Florence. For all their differences, they were united in their critique of what they thought complicated the essence of Christianity. On biblical prophetic grounds they sought simplicity in the cognitive, moral, and devotional life of Christianity. They may have disagreed over the essence of the faith, but they were united in what they thought were accretions that obscured the essence.

The Reformation as debates over essence

When the Protestant Reformation divided Western Christianity—as Eastern Christians, already separated since the 11th century, looked on—the 16th-century European world experienced a foretaste of the infinite Christian variety to come. The reforms that gave rise to the many Protestant bodies—Lutheran, Anglican, Presbyterian, Reformed, Anabaptist, Quaker, and others—were themselves debates over the essence of Christianity. Taken together, they made it increasingly difficult for any one to claim a monopoly on the custodianship of that essence. Each new sect offered a partial discernment of a different essence or way of speaking of it, even if the vast majority of Protestants agreed that the essence could be retrieved best, or, indeed uniquely, through recovery of the central message of the Holy Scriptures.

After the ferment of Reformation, most of the dissenting groups, as they established themselves in various nations, found it necessary to engage in their own narrowing of focus, rendering of precise doctrines, and understanding of divine truth and the way of salvation. Within a century theologians at many Protestant universities were adopting systems that paralleled the old scholasticisms against which some reformers had railed. Those who had once thought that definition of doctrine failed to capture the essence of Christianity were now defining their concept of the essence in doctrinal terms, but were doing so for Lutherans, Reformed, Presbyterians, and even for more radical dissenters and resistors of creeds, such as the Anabaptists.

The belief of Vincent of Lérins that there is a faith that has been held by everyone, always, and everywhere, lived on through the proliferation of Protestant denominations and Roman Catholic movements and, in sophisticated ways, has helped animate the modern ecumenical movement. Thus some have spoken of that movement as a reunion of churches, an idea that carries an implication that they had once been "one," and a further hint that that one included an essence on which people agreed. Reunion, then, would mean a stripping away of accretions, a reducing of the number of arguments, and a refocusing on essentials.

**Modern views.** The modern church and world brought new difficulties to the quest for defining an essence of Christianity. Both as a result of Renaissance humanism, which gloried in human achievement and encouraged human autonomy, and of Reformation ideas that believers were responsible in conscience and reason for their faith, an autonomy in expressing faith developed. Some spoke of Protestantism as being devoted to the right of private judgment. The danger, said Roman Catholic custodians of the essence, was that proud believers who did not submit to church authority would issue as many concepts of essence as there were believers to make the claims.

Effect of the Enlightenment

In the 18th century the Western philosophical movement called the Enlightenment further obscured searches for the essence of Christianity. The Enlightenment proclaimed optimistic views of human reach and perfectibility that challenged formerly essential Christian views of human limits. The deity became a benevolent if impersonal force, not an agent that arranged a way of salvation to people in need of rescue. The Enlightenment also clearly urged a view of human autonomy and of the use of reason in a search for truth. But this reason did not need to be responsive to supernatural revelation, as contained in the Old and New Testaments. Indeed, it called the integrity of those scriptures themselves into question through methods of historical and literary criticism. No longer should one rely on the word of priests who passed on notions of essential Christianity through systems of authority and force.

While many Westerners moved out of the orbit of faith as a result of the Enlightenment and the rise of criticism, many others—in Germany, France, England, Scotland, and, eventually, the Americas—chose to remain Christians, people of faith if now of faith differently expressed. Some, in groups called Arminian, professed such high views of human potential that their essence of Christianity prescribed no need for salvation. They thus constituted a real challenge to the profession of Vincent of Lérins. Another camp of Christians, the Unitarians, rejected the ideas of both a preexistent Logos made incarnate in Christ and a Jesus adopted into godhead. Jesus was seen as the great teacher or exemplar. They thus also tested the boundaries of essential teaching about a way of salvation. And at the heart of Deist Christianity was a view of God that remained "mono-" in that it was devoted to a single principle, but as "deist" instead of "theist" it departed from the ancient picture of a personal God engaged in human affairs. These were further blows to the integrity of Vincent of Lérins' concept and more reasons for the orthodox to use Vincent's concept to exclude Arminians, Unitarians, Deists, and other innovators from the circle of Christianity.

In the 19th century philosophical and historical criticism did inspire some Christians to renew the search for essences. For example, in the wake of the German Idealist philosopher G.W.F. Hegel, Hegelian scholars tried to rescue Christianity by viewing it as an unfolding of "absolute spirit." They followed Christian history through a constant dialectic, a series of forces and counterforces producing new syntheses. A problem with this Hegelian

approach arose as the historical Jesus came to be seen merely as one stage in the unfolding of absolute spirit; he was not a decisive agent of the way of salvation "once for all," as the biblical Letter to the Hebrews had claimed him to be. Soon biblical scholars such as David Friedrich Strauss were speaking of the historical Jesus as a myth of a certain set of people in one moment of the dialectical unfolding. The Christian faith itself began to dissolve, and many Hegelians began to reject the God of the Christian faith along with the historical Jesus.

Another group of 19th-century theologians took the opposite course. In the spirit of the 18th-century German philosopher Immanuel Kant, these neo-Kantians spoke not of the noumenal world, the unseen realm of essences beyond visible reality, but of the phenomenal realm, the world of history in which things happened. Theologians in this school engaged in a century-long "quest for the historical Jesus," in which they sought the simple essence of Christianity. Significantly, it was the greatest exemplar of this historical tradition, the German theologian Adolf von Harnack, who wrote the best-known modern book on the essence of Christianity, *Das Wesen des Christentums* (1900; *What Is Christianity?*).

The quest for the historical Jesus

The call had come to purge Christianity of what Harnack called traces of "acute Hellenization," the Greek ideas of essence, substance, and being that were introduced into the Christian tradition in the creed-making period. Instead, the focus was shifted to the Fatherhood of God and the announcement of the Kingdom, as the rabbi Jesus of Nazareth had proclaimed in the Gospels. While this approach did match the thirst for simplification in the minds of many of the Christian faithful, it also so diminished the concept of God that it often became a form of Christian humanism and was seen by the orthodox to be another departure from the essence of Christianity even as it claimed to find this in the historical Jesus. And scholars could not agree on the details of that historical Jesus after historical and literary critics had analyzed the Gospels.

Throughout the modern period some thinkers took another route toward expressing the essence of Christianity. Whether among German Pietists, the followers of John Wesley into Methodism, or any number of Roman Catholic or Protestant movements of devotion, there grew the notion that the theologians would never find the essence of Christianity. Instead, one would discern this essence in acts of piety, closeness to the fatherly heart of God as shown in the life of Jesus, or intimate communion with God on emotional or affective, and not cognitive, rational, or substantial (*i.e.,* doctrinal), grounds. These pietisms have been immensely satisfying to millions of modern believers, though they have often been handicapped in the intellectual arena when pressed for the definitions people need in a world of choice.

There have been some modern Christians who have shifted the topic from the essence of Christianity to its absoluteness among the religions. They have been moved by what the Germans called *Religionswissenschaft,* the study of world religions. In that school, the focus fell on the sacred, what the German theologian Rudolf Otto called "the idea of the Holy." There could be any number of expressions of this idea. On those terms, as the German scholar Ernst Troeltsch showed, it was more difficult to speak of the "absoluteness" of Christianity and its truth; one had to speak of it on comparative terms. Yet some early 20th-century comparativists, such as the Swedish Lutheran archbishop Nathan Söderblom, applied their understanding of the study of religion to help animate the movement for Christian reunion.

Ecumenism and essence

The modern ecumenical movement is made up of people who believe the church has different cultural expressions that must be honoured and differing confessional or doctrinal traditions designed to express the essential faith. These traditions demand criticism, comparison, and perhaps revision, with some possible blending toward greater consensus in the future. At the same time, years of serious ecumenical endeavour have shown that, among Christians of great intelligence and good will, elaborations of what constitutes the essence of Christianity are as confusing as they are inevitable and necessary.

In the Protestant and Orthodox side of the ecumenical movement, which took institutional form in the World Council of Churches in 1948, there were two main strands. Both of these included advocates of what were types of essential concepts. One set was devoted first to "Life and Work," a view that the essentials of Christianity could be best found and expressed when people followed the way or did the works of Christ, since this constituted his essence. The other set, concerned with "Faith and Order," stressed the need for comparative study of doctrine, with critical devotion to the search for what was central. By no means did these groups cling any longer to the notion that when they found unity they would have found a simple essence of Christianity. Yet they believed that they could find compatible elements that would help to sustain them on the never-ending search for what was central to the faith tradition.

Some modern scholars—for example, the British theologian John Hick—viewing the chaos of languages dealing with the essentials of the faith and the complex of historical arguments, pose the understanding of the essence in the future. They speak of "eschatological verification," referring to the end, the time beyond history, or the time of fulfillment. In that future, one might say, it will have become possible to assess the claims of faith. Theologians of these schools argue that such futuristic notions motivate Christians and the scholars among them to clarify their language, refine their historical understandings, and, some would say, focus their devotion and spirituality.

## THE QUESTION OF CHRISTIAN IDENTITY

From these comments on the search for the essence of Christianity, the task of defining the core of the faith tradition, it can be seen that at all times the question of Christian identity is at stake. What the psychologist Erik Erikson said of the individual, that a sense of identity meant "the accrued confidence that one's ability to maintain inner sameness and continuity . . . is *matched* by the sameness and continuity of one's meaning for others," is thus translated to the concerns of the group. This means that Christians strive, in the midst of change, to have some "inner sameness and continuity" through the focus on Jesus Christ and the way of salvation.

At the same time, Christians posit that this identity will be discoverable by and useful to those who are not part of the tradition: secularists, Buddhists, Communists, or other people who parallel or rival Christian claims about truth and salvation.

The phenomenological approach to identity

On these terms, the writers of Christian history normally begin phenomenologically when discussing Christian identity; that is, they do not bring norms or standards by which they have determined the truth of this or that branch of Christianity or even of the faith tradition as a whole but begin by identifying everyone as Christian who call themselves Christian. Thus, from one point of view, the Church of Jesus Christ of Latter-day Saints, or the Mormons, is, as scholar Jan Shipps calls it, "a new religious tradition." The followers of the *Book of Mormon* incorporated the Old and New Testaments into their canon—just as the New Testament Christians incorporated the entire scripture of a previous tradition—and then supplied reinterpretations. As a new religious tradition, Mormonism would not be Christian. But because Mormons use Christian terminology and call themselves Christian, they might also, from some points of view and in some areas, belong to a discussion of Christianity. They may be perceived as departing from the essence of Christianity because other Christians regard their progressive doctrine of God as heretical. Yet Mormons in turn point to perfectionist views of humanity and progressive views of God among more conventionally accepted Christian groups. In areas where the Mormons want to be seen as "latter-day" restorers, basing their essential faith on scriptures not previously accessible to Christians, they would be ruled out of conventional Christian discussion and treatment. Yet they share much of Christian culture, focus their faith in Jesus, proclaim a way of salvation, and want to be included for other purposes, and thus fall into the context of a Christian identity at such times.

This phenomenological approach, one that accents historical and contemporary description and resists prescription, does not allow the historian to state the essence of Christianity as a simple guide for all discussion. It is necessary for the scholar to put his own truth claims in a kind of suspension and to record faithfully, sorting out large schools of coherence and pointing to major strains. It is not difficult to state that something was a majority view if the supporting data are present. For example, it is not difficult to say what Roman Catholics at particular times have regarded as the essence of Christianity or what the various Orthodox and Protestant confessions regard as the true way of salvation. Someone using the phenomenological method, however, would stand back and refuse to be the arbiter when these confessional traditions disagree over truth.

Vincent of Lérins, then, speaks more for the hunger of the Christian heart or the dream of Christian union than for the researcher, who finds it more difficult to see a moment when everyone agreed on everything everywhere. Yet provisionally it remains safe to say that Christian identity begins and ends with a reference to Jesus in relation to God's truth and a way of salvation. The rest is a corollary of this central claim, an infinite set of variations and elaborations that are of great importance to the separated Christians who hold to them in various times and places. (M.E.M.)

## The history of Christianity

### THE PRIMITIVE CHURCH

**The relation of the early church to late Judaism.** Christianity began as a movement within Judaism at a period when the Jews had long been under foreign influence and rule and had found in their religion (rather than in their politics or cultural achievements) the linchpin of their community. From Amos (8th century BC) onward the religion of Israel was marked by tension between the concept of monotheism, with its universal ideal of salvation (for all nations), and the notion of God's special choice of Israel. In the age after Alexander the Great (*i.e.*, the Hellenistic period, 3rd century BC–3rd century AD), the dispersion of the Jews throughout the Hellenistic kingdoms and the Roman Empire gave some impetus to the universalistic tendency. But the attempts of foreign rulers, especially the Syrian king Antiochus Epiphanes (in 168–165 BC), to impose Greek culture and religious syncretism in Palestine provoked zealous resistance on the part of many Jews. In Palestinian Judaism the predominant note was separation and exclusiveness. Jewish missionaries to other areas were strictly expected to impose the distinctive Jewish customs of circumcision, kosher food, and sabbaths and other festivals.

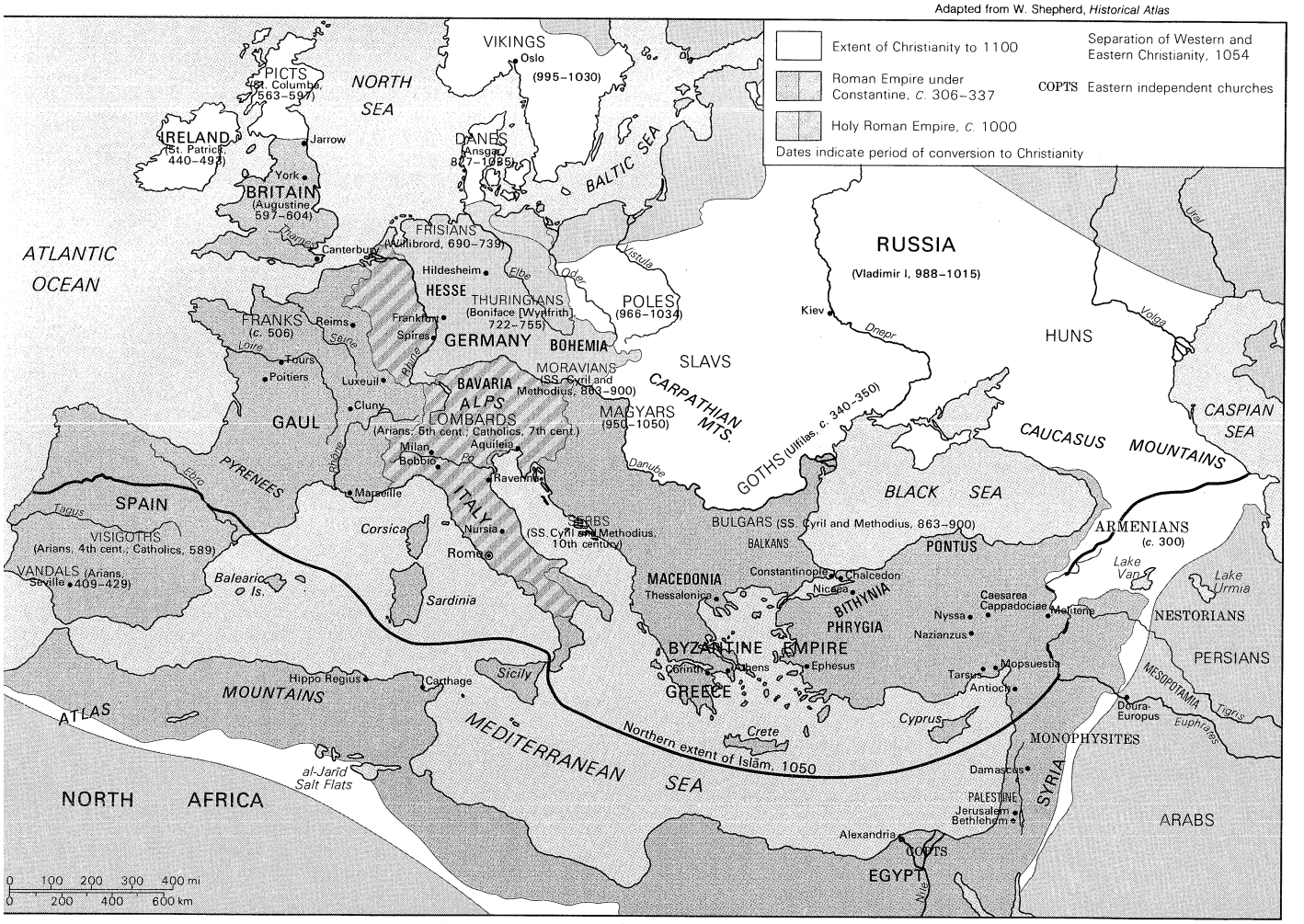
Universalistic and particularistic tendencies in Judaism

The relationship of the earliest Christian Church to Judaism turned principally on two questions: (1) the messianic role of Jesus of Nazareth and (2) the permanent validity of the Mosaic Law for all.

The Hebrew Scriptures viewed history as the stage of a providential drama eventually ending in a triumph of God over all present sources of frustration (*e.g.*, foreign domination or the sins of Israel). God's rule would be established by an anointed prince (the Messiah) of the line of David, king of Israel in the 10th century BC. The proper course of action leading to the consummation of the drama, however, was the subject of some disagreement. Among the diverse groups were the aristocratic and conservative Sadducees, who accepted only the five books of Moses (the Pentateuch), and the more popular and strict Pharisees. The Pharisees not only accepted biblical books outside the Pentateuch but also embraced doctrines—such

Jewish sectarianism

Adapted from W. Shepherd, *Historical Atlas*




Spread of Christianity through the 11th century in Europe, North Africa, and the Middle East.

as those on resurrection and the existence of angels—of recent acceptance in Judaism, many of which were derived from apocalyptic expectations that the consummation of history would be heralded by God's intervention in the affairs of men in dramatic, cataclysmic terms. The Sanhedrin (central council) at Jerusalem was made up of both Pharisees and Sadducees. The Zealots were aggressive revolutionaries seeking independence from Rome. Other groups were the Herodians, supporters of the client kingdom of the Herods (a dynasty that supported Rome) and abhorrent to the Zealots, and the Essenes, a quasi-monastic dissident group, probably including the sect that preserved the Dead Sea Scrolls. This latter sect did not participate in the Temple worship at Jerusalem and observed another religious calendar; from their desert retreat they awaited divine intervention and searched prophetic writings for signs indicating the consummation.

What relation the followers of Jesus had to some of these groups is not clear. In the canonical Gospels (those accepted as authentic by the church) the main targets of criticism are the scribes and Pharisees, whose attachment to the tradition of Judaism is presented as legalistic and pettifogging. The Sadducees and Herodians likewise receive an unfriendly portrait. The Essenes are never mentioned. Simon, one of Jesus' 12 disciples, was or had once been a Zealot. Jesus probably stood close to the Pharisees.

Under the social and political conditions of the time, there could be no long future either for the Sadducees or for the Zealots—whose attempts to make apocalyptic dreams effective led to the destruction of Judaea after the two major Jewish revolts of 66–70 and 132–135 against the Romans. The choice for many Jews thus lay between the Pharisees and Christianity, the former dedicated to the meticulous preservation of the Mosaic Law and the latter to the universal propagation of the biblical faith as a religion for all mankind. Pharisaism as enshrined in the Mishna (Oral Law) and the Talmud (commentary on and addition to the Oral Law) became normative Judaism. By looking to the Gentile (non-Jewish) world and carefully dissociating itself from the Zealot revolutionaries, Christianity made possible its ideal of a world religion, at the price of sacrificing Jewish particularity and exclusiveness. The fact that Christianity has never succeeded in gaining the open allegiance of more than a minority of Jews is more a mystery to theologians than to historians.

**The relation of the early church to the career and intentions of Jesus.** The prime sources for knowledge of Jesus of Nazareth are the four canonical Gospels in the New Testament. Only a few probably authentic sayings of Jesus became preserved in oral tradition independent of these documents, though many sayings came to be put into his mouth. These noncanonical sayings are called agrapha (not in Scripture). The *Gospel of Thomas,* preserved in a Coptic Gnostic library found about 1945 in Egypt, contains several such sayings, besides some independent versions of canonical sayings. At certain points the Gospel tradition finds independent confirmation in the letters of the Apostle Paul. The allusions in non-Christian sources (the Jewish historian Josephus, the Roman historians Tacitus and Suetonius, and Talmudic texts) are almost negligible, except as refuting the unsubstantiated notion that Jesus might never have existed.

The Gospels as sources for knowledge about the early church

The first three Gospels, Matthew, Mark, and Luke, have a literary relation to one another and are hence called Synoptic. Mark was probably used by Matthew and Luke. John, differing in both pattern and content, appears richer in theological interpretation but in detail may preserve good historical information. As their titles imply, the Gospels are not detached reports but were written to serve religious needs; they resemble oil paintings rather than photographs. Legendary and apologetic (defensive) motifs, and the various preoccupations of the communities for which they were first produced, can readily be discerned as influences upon their narratives. Historical scholarship at present has insufficient tools to eliminate subjective judgments about the probability of many details (upon which there will always be disagreement), but the most persuasive scholarly consensus accepts the substance of the Gospel tradition as a veracious picture.

A prominent uncertainty is the matter of chronology. Matthew places the birth of Jesus at least two years before Herod the Great's death late in 5 BC or early in 4 BC. Luke connects Jesus' birth with a Roman census that, according to Josephus, occurred in AD 6–7 and caused a revolt against the governor Quirinius. Luke could be right about the census and wrong about the governor. The crucifixion under Pontius Pilate, prefect of Judaea (AD 26–36), was probably about the year 29–30, but again certainty is impossible.

Encounter with John the Baptist, the ascetic in the Judaean Desert who preached repentance and baptism in view of God's coming Kingdom, marked a decisive moment for Jesus' career. He recognized in John the forerunner of the kingdom that his own ministry was inaugurating. The first preaching of Jesus, in his home region of Galilee, took the form of vivid parables and was accompanied by miraculous healings. The Synoptic writers give a single climactic visit of Jesus to Jerusalem at the end of his career; but John may be right (implicitly supported by Luke 13:7) in representing his visits as more frequent and the period of ministry as lasting more than a single year. Jesus' attitude to the observance of the law generated conflict with the Pharisees, and, though the people protected him, he also aroused the fear and hostility of the ruling Jewish authorities. A triumphal entry to Jerusalem at Passover time (the period celebrating the Exodus of the Hebrews from Egypt in the 13th century BC) was the prelude to a final crisis. After a last supper with his disciples he was betrayed by one of them, Judas. Arrest and trial followed, first before the Sanhedrin and then before Pilate, who condemned him to crucifixion. The accusation before Pilate was sedition, in which the Evangelists saw a framed charge. It was universal Christian belief that three days after his death he was raised from the dead by divine power.

Jesus preached the imminent presence of God's Kingdom, in some texts as future consummation, in others as already present. The words and acts of Jesus were believed to be the inauguration of a process that was to culminate in a final triumph of God. His disciples recognized him as the Messiah, the Anointed One. He is not recorded to have used the word of himself. The titles Prophet and Rabbi also were applied to Jesus. His own enigmatic self-designation was "Son of man," sometimes in allusion to his suffering, sometimes to his future role as judge. This title is derived from the version of the Book of Daniel (7:13), where "one like a son of man," contrasted with beast figures, represents the humiliated people of God, ascending to be vindicated by the divine Judge. In the developed Gospel tradition the theme of the transcendent judge seems to be most prominent.

The Kingdom of God and apocalyptic hopes

Apocalyptic hope could easily merge into messianic zealotry. Moreover, Jesus' teaching was critical of the established order and encouraged the poor and oppressed, even though it contained an implicit rejection of revolution. Violence was viewed as incompatible with the ethic of the Kingdom of God. Whatever contacts there may have been with the Zealot movement (as the narrative of feeding 5,000 people in the desert may hint), the Gospels assume the widest distance between Jesus' understanding of his role and the Zealot revolution.

With this distance from revolutionary idealism goes a sombre estimate of human perfectibility. The gospel of repentance presupposes deep defilement in individuals and in society. The sufferings and pains of humanity under the power of evil spirits calls out for compassion and an urgent mission. All the acts of a disciple must express love and forgiveness, even to enemies, and also detachment from property and worldly wealth. To Jesus, the outcasts of society (prostitutes, the hated and oppressive tax agents, and others) were objects of special care, and censoriousness was no virtue. Though the state is regarded as a distant entity in certain respects, it yet has the right to require taxes and civic obligations: Caesar has rights that must be respected and are not incompatible with the fulfillment of God's demands.

Some of the futurist sayings, if taken by themselves, raise the question whether Jesus intended to found a church.

A negative answer emerges only if the authentic Jesus is assumed to have expected an immediate catastrophic intervention by God. There is no doubt that he gathered and intended to gather around him a community of followers. This community continued after his time, regarding itself as the specially called congregation of God's people, possessing as covenant signs the rites of baptism and Eucharist (Lord's Supper) with which Jesus was particularly associated—baptism because of his example, Eucharist because the Last Supper on the night before the crucifixion was marked as an anticipation of the messianic feast of the coming age.

Expansion of Christianity to include Gentiles

A closely related question is whether Jesus intended his gospel to be addressed to Jews only or if the Gentiles were also to be included. In the Gospels Gentiles appear as isolated exceptions, and the choice of 12 Apostles has an evident symbolic relation to the 12 tribes of Israel. The fact that the extension of Christian preaching to the Gentiles caused intense debate in the 40s of the 1st century is decisive proof that Jesus had given no unambiguous directive on the matter. Gospel sayings that make the Jews' refusal to recognize Jesus' authority as the ground for extending the Kingdom of God to the Gentiles must, therefore, have been cast by the early community.

**The Gentile mission and St. Paul.** Saul, or Paul (as he was later called), was a zealous Pharisee who persecuted the primitive church. Born at Tarsus (Asia Minor), he had come to Jerusalem as a student of the famous Rabbi Gamaliel and had harried a Christian group called by Luke the "Hellenists," who were led by Stephen (the first Christian martyr) and who regarded Jesus as a spiritual reformer sent to purge the corrupt worship of Jerusalem. While on a mission to persecute the followers of Jesus, Paul was suddenly converted to faith in Christ and, simultaneously, to a conviction that the Gospel must pass to the non-Jewish world under conditions that dispensed with exclusively and distinctively Jewish ceremonies. Paul was disapproved by Christian Jews who were of conservative opinions and remained throughout his career a controversial figure. He gained recognition for the converts of the Gentile mission by the Christian community in Jerusalem; but his work was considered an affront to Jewish traditionalism, and his program of being "all things to all men" led to bitter charges that he was an unprincipled trimmer. He saw clearly and correctly that the universal mission of the church to all humanity, implicit in the coming of the Messiah, or Christ, meant a radical break with rabbinical conservatism.

Gentile–Jewish debates within Christianity

Owing to the preservation of some weighty letters, Paul is the only vivid figure of the apostolic age (1st century AD). Like his elder contemporary Philo of Alexandria, also a Hellenized Jew of the dispersion, he interpreted the Old Testament allegorically (symbolically) and affirmed the primacy of spirit over letter in a manner that was in line with Jesus' freedom with regard to the sabbath. The crucifixion of Jesus he viewed as the supreme redemptive act and also as the means of expiation for the sin of mankind. Salvation is, in Paul's thought, therefore, not found by a conscientious moralism but rather is a gift of grace, a doctrine in which Paul was anticipated by Philo. But Paul linked this doctrine with his theme that the Gospel represents liberation from the Mosaic Law. The latter thesis created difficulties at Jerusalem, where the church was under the presidency of James, the brother of Jesus, and the circle of the intimate disciples of Jesus. James, martyred at Jerusalem in 62, was the primary authority for the Christian Jews, especially those made anxious by Paul; the canonical letter ascribed to James opposes the antinomian (anti-law) interpretations of the doctrine of justification by faith. A middle position seems to have been occupied by Peter. All the Gospels record a special commission of Jesus to Peter as the leader among the 12 Apostles. But Peter's biography can only be dimly constructed; he died in Rome (according to early tradition) in Nero's persecution (64) about the same time as Paul.

Apart from its success, the supremacy of the Gentile mission within the church was ensured by the effects on Jewish Christianity of the fall of Jerusalem (70) and Hadrian's exclusion of all Jews from the city (135). Jewish Christianity declined and became the faith of a very small group without links to either synagogue or Gentile church. Some bore the title Ebionites, "the poor" (compare Matthew 5:3). Among them some did not accept the tradition that Jesus was born of a virgin.

In the theology of Paul, the human achievement of Jesus was important because his obedient fidelity to his vocation gave moral and redemptive value to his self-sacrifice. A different emphasis appears in the Gospel According to John, written (according to 2nd-century tradition) at Ephesus. John's Gospel partly reflects local disputes, not only between the church and the Hellenized synagogue but also between orthodox Christianity and deviationist Gnostic groups in Asia Minor. John's special individuality lies in his view of the relation between the historical events of the tradition and the Christian community's present experience of redemption. The history is treated symbolically to provide a vehicle for faith. Because it is less attached to the contingent events of a particular man's life, John's conception of the preexistent Logos becoming incarnate (made flesh) in Jesus made intelligible to the Hellenistic world the universal significance of Jesus. In antiquity, divine presence had to be understood as either inspiration or incarnation. If the Synoptic Gospels suggest inspiration, the Gospel According to John chooses incarnation. The tension between these two types of Christology (doctrines of Christ) first became acute in the debate between the schools of Antioch and Alexandria in the late 4th century.

**The contemporary social, religious, and intellectual world.** Many Palestinian Jews appreciated the benefits of Roman rule in guaranteeing order and peace. The Roman government could tolerate regional and local religious groups and found it convenient to control Palestine through client kings like the Herods. The demand that divine honours be paid not only to the traditional Roman or similar gods but also to the emperors was not extended to Judaea except under the emperor Caligula (reigned 37–41). It was enough that the Jews dedicated temple sacrifices and synagogues in the emperor's honour. The privileges of Roman citizenship were possessed by some Jewish families, including that of the Apostle Paul.

Relationship of early Christianity to state and society

In his letter to the Romans, Paul affirmed the providential role of government in restraining evil. Christians did not need to be disaffected from the empire, though the deification of the emperor was offensive to them. Moreover, although as an agency of social welfare the church offered much to the downtrodden elements in society, the Christians did not at any stage represent a social and political threat. The ancient world did not possess a working-class movement in the modern sense, and Christianity did not create or foster one. After the example of their master, the Christians encouraged humility and patience before wicked men. Even the institution of slavery was not the subject of fundamental Christian criticism before the 4th century. The church, however, was not lost in pious mysticism. It provided for far more than the cultic (liturgical) needs of its members. Inheriting a Jewish moral ideal, its activities included food for the poor, orphans, and foundlings; care for prisoners; and a community funeral service.

The church inherited from Judaism also a strong sense of being holy, separate from idolatry and pagan eroticism. As polytheism with its attendant permissiveness permeated ancient society, a moral rigorism severely limited Christian participation in some trades and professions. At baptism a Christian was expected to renounce his occupation if that necessarily implicated him in public or private compromise with polytheism, superstition, dishonesty, or vice. About military service there was disagreement. The majority held that a soldier, if converted and baptized, was not required to leave the army, but there was hesitation about whether an already baptized Christian might properly enlist. Strict Christians also thought poorly of the teaching profession because it involved instructing the young in literature replete with pagan ideals and what was viewed as indecency. Acting and dancing were similarly suspect occupations. Any involvement in magic was completely forbidden.

The Christian ethic therefore demanded some detachment from society. In some cases this made for economic

difficulties. The structure of ancient society was dominated not by class but by the relationship of patron and client. A slave or freedman depended for his livelihood and prospects upon his patron. In antiquity a strong patron was indispensable if one was negotiating with police or tax authorities or lawcourts or if one had ambitions in the imperial service. The authority of the father of the family was considerable. Conversely, a man's power in society depended on the extent of his dependents and supporters. Often, Christianity penetrated the social strata first through women and children, especially in the upper classes. But once the householder was a Christian, his dependents tended to follow. The Christian community itself was close-knit. Third-century evidence portrays Christians banking their money with fellow believers; and widely separated groups helped one another with trade and mutual assistance.

Women in ancient society—Greek, Roman, or Jewish—had a domestic, not a public, role; feminine subordination was self-evident. To Paul, however, Christian faith transcends barriers to make all free and equal (Galatians 3:28). Of all ancient writers Paul was the most powerful spokesman for equality. Nevertheless, just as he refused to harbour a runaway slave, so he opposed any practice that would identify the church with social radicalism (a principal pagan charge against it). Paul did not avoid self-contradiction (1 Corinthians 11:5, 14:34–35). His opposition to a public liturgical role for women decided subsequent Catholic tradition in the East and West. Yet in the Greek churches (though not often in the Latin) women were ordained as deacons—in the 4th century by prayer and imposition of hands with the same rite as male deacons—and had a special responsibility at women's baptism. Widows and orphans were the neediest in antiquity, and the church provided them substantial relief. It also encouraged vows of virginity, and by AD 400 women from wealthy or politically powerful families acquired power as superiors of religious communities. It seemed natural to elect as abbess a woman whose family connections might bring benefactions.

Gentile religious environment

The religious environment of the Gentile mission was a tolerant, syncretistic blend of many cults and myths. Paganism was concerned with success; the gods gave victory in war, good harvests, success in love and marriage, and sons and daughters. Defeat, famine, civil disorder, and infertility were probable signs of cultic pollution and disfavour. People looked to religion for help in mastering the forces of nature rather than to achieve moral improvement. Individual gods cared either for specific human needs or for specific places and groups. The transcendent God of biblical religion was, therefore, very different from the numerous gods of limited power and local significance. In Asia Minor Paul and his coworker Barnabas were taken to be gods in mortal form because of their miracles. To offer sacrifice on an altar seemed a natural expression of gratitude to any dead, or even living, benefactor. Popular enthusiasm could bestow divine honours on such heroes as dead pugilists and athletes. In the Roman Empire it seemed natural to offer sacrifice and burn incense to the divine emperor as a symbol of loyalty, much like standing for a national anthem today.

Traditional Roman religion was a public cult, not private mysticism, and was upheld because it was the received way of keeping heaven friendly. To refuse participation appeared to be disloyal. The Jews could gain acceptance for their refusal by virtue of the undoubted fact that their monotheism was an ancestral national tradition. The Christians, however, did everything in their power to dissuade people from following the customs of their fathers, whether Gentiles or Jews, and thereby seemed to threaten the cohesion of society and the principle that each racial group was entitled to follow its national customs in religion.

If ancient religion was tolerant, the philosophical schools were seldom so. Platonists, Aristotelians, Stoics, Epicureans, and Skeptics tended to be very critical of one another. By the 1st century BC, an eclecticism emerged; and by the 2nd century AD, there developed a common stock of philosophy shared by most educated people and by some professional philosophers, which derived metaphysics involving theories on the nature of Being from Plato, ethics from the Stoics, and logic from Aristotle. This eclectic Platonism provided an important background and springboard for early Christian apologetics. Its main outlines appear already in Philo of Alexandria, whose thought influenced not only perhaps the writer of the anonymous letter to the Hebrews in the New Testament but also the great Christian thinkers Clement of Alexandria, Origen, and Ambrose of Milan. Because of this widespread philosophical tendency, the Christian could generally assume some belief in Providence and assent to high moral imperatives. Platonism in particular provided a metaphysical framework within which the Christians could interpret the entire pattern of creation, the Fall of humanity, the incarnation, redemption, the church, sacraments, and last things.

## THE INTERNAL DEVELOPMENT OF THE EARLY CHRISTIAN CHURCH

**The problem of jurisdictional authority.** In the first Christian generation, authority in the church lay either in the kinsmen of Jesus or in those whom he had commissioned as Apostles and missionaries. The Jerusalem church under James, the brother of Jesus, was the mother church. Paul admitted that if they had refused to grant recognition to his Gentile converts he would have laboured in vain. If there was an attempt to establish a hereditary family overlordship in the church, it did not succeed. Among the Gentile congregations, the Apostles sent by Jesus enjoyed supreme authority. As long as the Apostles lived, there existed a living authoritative voice to which appeal could be made. But once they all had died, there was an acute question regarding the locus of authority. The earliest documents of the 3rd and 4th Christian generations are mainly concerned with this issue: what is the authority of the ministerial hierarchy? The apostolic congregations had normally been served by elders (Greek *presbyteroi,* "priests") or overseers (*episkopoi,* "bishops"), assisted by attendants (*diakonoi,* "deacons"). The clergy were responsible for preaching, for administering baptism and Eucharist, and for distributing aid to the poor. In each city the president or senior member of the college (assembly) of presbyters naturally had some special authority; he corresponded with other churches and, when they were ordaining a new president, would go as the representative of his own community and as a symbol of the catholicity—the universality and unity—of the church of Christ.

Bishops, presbyters, and deacons

Ignatius, bishop of Antioch early in the 2nd century, wrote seven letters on his way to martyrdom at Rome that indicate how critical the centrifugal forces in the church had made the problem of authority. The bishop, he insisted, is the unique focus of unity without whose authority there is no sacrament and no church. A few years earlier the letter of Bishop Clement of Rome (*c.* AD 95) to the church at Corinth based the hierarchy's authority on the concept of a historical succession of duly authorized teachers. Clement understood the clergy and laity to be essentially distinct orders within the one community, just as in the Old Testament there were high priests, priests, Levites (Temple functionaries), and laymen. The principles of Clement and Ignatius became important when the church was faced by people claiming recognition for their special charismatic (spiritual) gifts and especially by Gnostic heretics claiming to possess secret oral traditions whispered by Jesus to his disciples and not contained in publicly accessible records such as the Gospels.

The authority of the duly authorized hierarchy became enhanced by the outcome of another 2nd-century debate, about the possibility of absolution for sins committed after baptism. The *Shepherd of Hermas,* a book that enjoyed canonical status in some areas of the early church, enforced the point that excessive rigorism produces hypocrisies. By the 3rd century the old notion of the church as a society of holy people was being replaced by the conception that it was a school for frail sinners. In spite of protests, especially that of the schism led by the theologian and schismatic pope Novatian at Rome in 251, the final consensus held that the power to bind and loose (compare

Matthew 16:18–19), to excommunicate and absolve, was vested in bishops and presbyters by their ordination.

Early Christianity was predominantly urban; peasants on farms were deeply attached to old ways and followed the paganism favoured by most aristocratic landowners. By AD 400 some landowners had converted and built churches on their property, providing a "benefice" for the priest, who might often be one of the magnate's servants. In the East and in North Africa each township normally had its own bishop. In the Western provinces bishops were fewer and were responsible for larger areas, which, from the 4th century onward, were called by the secular term dioceses (administrative districts). In the 4th century pressure to bring Western custom into line with Eastern and to multiply bishops was resisted on the ground that it would diminish the bishops' social status. By the end of the 3rd century the bishop of the provincial capital was acquiring authority over his colleagues: the metropolitan (from the 4th century on, often entitled archbishop) was chief consecrator of his episcopal colleagues. The bishops of Rome, Alexandria, and Antioch in the 3rd century were accorded some authority beyond their own provinces. Along with Jerusalem and Constantinople (founded in 330), these three sees (seats of episcopal authority) became, for the Greeks, the five patriarchates. The title *papa* ("father") was for 600 years an affectionate term applied to any bishop to whom one's relation was intimate; it began to be specially used of bishops of Rome from the 6th century and by the 9th century was almost exclusively applied to them.

Expansion of administrative authority

From the beginning, the Christians in Rome were aware of special responsibilities for them to lead the church. About AD 165, memorials were erected at Rome to the Apostles Peter and Paul, to Peter in a necropolis on the Vatican Hill, and to Paul on the road to Ostia. The construction reflects a sense of being guardians of an apostolic tradition, a self-consciousness expressed in another form when, about 190, Bishop Victor of Rome threatened with excommunication Christians in Asia Minor who, following immemorial custom, observed Easter on the day of the Jewish Passover rather than (as at Rome) on the Sunday after the first full moon after the spring equinox. Stephen of Rome (256) is the first known pope to base claims to authority on Jesus' commission to Peter (Matthew 16:18–19).

Bishops were elected by their congregations—*i.e.*, by the clergy and laity assembled together. But the consent of the laity decreased in importance as recognition by other churches increased. The metropolitan and other provincial bishops soon became just as important as the congregation as a whole; and, though they could never successfully impose a man on a solidly hostile community, they could often prevent the appointment falling under the control of one powerful lay family or faction. From the 4th century on, the emperors occasionally intervened to fill important sees, but such occurrences were not a regular phenomenon (until the 6th century in Merovingian Gaul).

**The problem of scriptural authority.** After the initial problems regarding the continuity and authority of the hierarchy, the greatest guarantee of true continuity and authenticity was found in the Scriptures. Christians inherited (without debate at first) the Hebrew Bible as the Word of God to the people of God at a now superseded stage of their pilgrimage through history. If St. Paul's Gentile mission was valid, then the Old Testament Law was viewed as no longer God's final word to his people. Thus, the Hebrew Bible began to be called the *old* covenant. There was some hesitation in the church about the exact books included. The Greek version of the Old Testament (Septuagint) included books (such as the Wisdom of Solomon, Ecclesiasticus, and others) that were not accepted in the Hebrew canon. Most, but not all, Gentile Christian communities accepted the Septuagintal canon. The 3rd-century Alexandrian theologian Origen and especially the Latin biblical scholar Jerome (4th–5th century) believed it imprudent to base theological affirmations on books enjoying less than universal recognition. The fact that in many English Bibles the parts of the Old Testament accepted in the Septuagint but not in the Hebrew canon are often printed separately under the (misleading) title Apocrypha is a tribute to these ancient hesitations.

The emergence of a canon of Christian Scripture

The growth of the New Testament is more complex and controversial. First-century Christians used oral tradition more than writing to pass on the story of Jesus' acts and words, often told in the context of preaching and teaching. No one thought they needed to be in writing to bear authority. Mark first conceived the plan of composing a connected narrative. Nevertheless, even after the Gospels were in common circulation, oral tradition was still current and could even be preferred. A carefully copied document, however, provided greater security against contamination of the tradition. The Synoptic Gospels seem to have been used by the Apologist Justin Martyr at Rome about AD 150 in the form of an early harmony (or synthesis of the Gospels); to this, Justin's Syrian pupil Tatian added the Gospel According to John to make his *Diatessaron* (according to the four), a harmony of all four Gospels so successful that in Mesopotamia (Tatian's homeland) it virtually ousted the separate Gospels for 250 years.

On a second grade of authority stood the apostolic letters, especially those of Paul. The main body of his correspondence was circulating as a corpus (body of writings) well before AD 90.

Significance of Marcion

Paul's antitheses of law and grace, justice and goodness, and the letter and the spirit were extended further than Paul intended by the radical semi-Gnostic heretic Marcion of Pontus (*c.* 140–150), who taught that the Old Testament came from the inferior vengeful Jewish God of justice and that the New Testament told of the kindly universal Father. As the current texts of Gospels and letters presupposed some divine revelation through the Old Testament, Marcion concluded that they had been corrupted and interpolated by Judaizers. Marcion therefore established a fixed canon of an edited version of Luke's Gospel and some of the Pauline Letters (expurgated), and no Old Testament at all.

The orthodox reaction (by such theologians as Justin, Irenaeus, and Tertullian in the 2nd century) was to insist on the Gospel as the fulfillment of prophecy and on creation as the ground of redemption. Reasons were found for accepting the four already current Gospels, the full corpus of Pauline Letters, Acts, John's Revelation (Apocalypse), and some of the Catholic Letters (these last—I, II, and III John, James, and Jude—were the subject of hesitations). On the authorship of the Letter to the Hebrews there were doubts: Rome rejected it as non-Pauline and Alexandria accepted it as Pauline. The list once established was a criterion (the meaning of "canon") for the authentic Gospel of the new covenant and soon (by transference from the old) became entitled the New Testament. (The Greek word *diathēkē* means both covenant and testament.) The formation of the canon meant that special revelation ended with the death of the Apostles and that no authority could be attached to the apocryphal gospels, acts, and apocalypses proliferating in the 2nd century.

**The problem of theological authority.** Third, a check was found in the creed. At baptism, after renouncing "the devil and his pomps," initiates declared their faith in response to three questions of the form:

> Do you believe in God the Father almighty? Do you believe in Jesus Christ his Son our Lord. . . ? Do you believe in the Holy Spirit in the church and in the Resurrection?

In time, these interrogations became the basis of declaratory creeds, adapted for use by the clergy who felt themselves required to reassure colleagues who were not especially confident of their orthodoxy. The so-called Apostles' Creed is a direct descendant of the baptismal creed used at Rome by AD 200. Each church (or region) might have its own variant form, but all had the threefold structure.

The internal coherence given by creed, canon, and hierarchy was necessary both in the defense of authentic Christianity against Gnostic theosophical speculations and also in confronting pagan society. The strong coherence of the scattered congregations was remarkable to pagan observers.

Gnosticism and Montanism

**Early heretical movements.** Gnosticism was the greatest threat to Christianity before 150 and somewhat thereafter. Gnostics taught that there is total opposition between this evil world and God. Redemption was viewed as libera-

tion from the chaos of a creation derived from either incompetent or malevolent powers, a world in which the elect are alien prisoners. The method of salvation was to discover the Kingdom of God within one's elect soul and to learn how to pass the hostile powers barring the soul's ascent to bliss. Gnosticism destroyed the notion of a historical disclosure of God. Its pessimism and dualism (in which matter was viewed as evil and spirit good) had disturbing moral consequences, involving both asceticism and libertinism. Its claims to a totally transcendent revelation were antirational, allowed for no natural goodness in the created order, and eliminated individual responsibility. Both the orthodox theologians and the pagan 3rd-century philosopher Plotinus dismissed Gnosticism as a pretentious but dangerous mumbo jumbo for misleading the half-educated.

The orthodox stressed the need to adhere to tradition, which was attested by the churches of apostolic foundation. A more hazardous reply was to appeal to ecstatic prophecy. About AD 172 a quasi-pentecostal movement in Phrygia was led by Montanus with two prophetesses, reasserting the imminence of the end of the world. He taught that there was an age of the Father (Old Testament), an age of the Son (New Testament), and an age of the Spirit (heralded by the prophet Montanus). Montanism won its chief convert in the Latin theologian Tertullian of Carthage. Its claim to supplement the New Testament was generally rejected.

## RELATIONS BETWEEN CHRISTIANITY AND THE ROMAN GOVERNMENT AND THE HELLENISTIC CULTURE

**Church–state relations.** The Christians were not respectful toward ancestral pagan customs. Their preaching of a new king sounded like revolution. The opposition of the Jews to them led to breaches of the peace. Thus the Christians could very well be unpopular, and they often were. Paul's success at Ephesus provoked a riot to defend the cult of the goddess Artemis. In AD 64 a fire destroyed much of Rome; the emperor Nero killed a "vast multitude" of Christians as scapegoats. For the first time, Rome was conscious that Christians were distinct from Jews. But there probably was no formal senatorial enactment proscribing Christianity at this time. Nero's persecution was local and short. Soon thereafter, however, the profession of Christianity was defined as a capital crime, though of a special kind because one gained pardon by apostasy (rejection of a faith once confessed) demonstrated by offering sacrifice to the pagan gods or the emperor. Popular gossip soon accused the Christians of secret vices, such as eating murdered infants (due to the secrecy surrounding the Lord's Supper and the use of the words body and blood) and sexual promiscuity (due to the practice of Christians calling each other "brother" or "sister" while living as husband and wife). The governor of Bithynia in AD 111, the younger Pliny, told the emperor Trajan that to his surprise he discovered the Christians to be guilty of no vice, only of obstinacy and superstition. Nevertheless, he executed without a qualm those who refused to apostatize.

Causes of persecutions

Early persecutions were sporadic, caused by local conditions and depending on the attitude of the governor. The fundamental cause of persecution was that the Christians conscientiously rejected the gods whose favour was believed to have brought success to the empire. But distrust was increased by Christian detachment and reluctance to serve in the imperial service and in the army. At any time in the 2nd or 3rd centuries, Christians could find themselves the object of unpleasant attention. A pogrom could be precipitated by a bad harvest, a barbarian attack, or a public festival of the emperor cult. Yet, long periods of peace occurred. In 248–250, when Germanic tribes threatened the empire, popular hostility culminated in the persecution under the emperor Decius (reigned 249–251): by edict all citizens were required to offer sacrifice and to obtain from commissioners a certificate witnessing to the act. Many of these certificates have survived. The requirement created an issue of conscience, especially because certificates could be bought by bribes. Under renewed attack (257–259), the great bishop-theologian Cyprian of Carthage was martyred. The persecuting emperor Valerian, however, became a Persian prisoner of war, and his son Gallienus issued an edict of toleration restoring confiscated churches and cemeteries. The church prospered from 261 to 303, but the empire suffered external attack, internal sedition, and rampant inflation. In February 303 the worst of all persecutions erupted under the co-emperors Diocletian and Galerius. The persecutions ended and peace was reached with the Edict of Milan, a manifesto of toleration issued in 313 by the joint emperors Licinius and his Christian colleague Constantine. Disagreements about the point at which the state must be resisted led to long lasting schisms in Egypt (Melitianism) and North Africa (Donatism).

**Christianity and classical culture.** St. Paul could quote such pagan poets as Aratus, Menander, and Epimenides. Clement of Rome cited the dramatists Sophocles and Euripides. Educated Christians shared this literary tradition with educated pagans. The defenders of Christianity against pagan attack (especially Justin Martyr and Clement of Alexandria in the 2nd century) welcomed classical philosophy and literature; they wished only to reject all polytheistic myth and cult and all metaphysical and ethical doctrines irreconcilable with Christian belief (*e.g.*, Stoic materialism and Platonic doctrines of the transmigration of souls and the eternity of the world). Clement of Alexandria, second known head of the catechetical school at Alexandria, possessed a wide erudition in the main classics and knew the works of Plato and Homer intimately. His successor at Alexandria, Origen, showed less interest in literary and aesthetic matters but was a greater scholar and thinker; he first applied the methods of Alexandrian philology to the text of the Bible.

Nevertheless, both pagans and Christians instinctively assumed the unity of ancient culture and pagan religion; it was hard for Christians to attack paganism without seeming negative toward the totality of classical culture as well as disaffected toward the imperial government. The urgent eschatological hope of the earliest church had built into its ethic a deep detachment from this world's goods, however beautiful they might be esteemed. This detachment emerged in one form in the evaluation of celibacy as superior to marriage, in another in a conscious renunciation of pretensions to high culture (in the manner of the not always popular or socially accepted pre-Christian Cynic philosophers with whom pagans found it natural to compare the Christians). The passionate urgency of the Christian mission admitted no distraction, an attitude that stamped any serious interest in science, history, or belles lettres with the stigma of worldliness.

Attitude of Christians toward other religions

The attitude of the Christians toward other religions (except Judaism) was generally very negative. All forms of paganism—the Oriental mystery (salvational) religions of Isis, Attis, Adonis, and Mithra as well as the ancient cults of classical Greece and Rome—were regarded as the cults of evil spirits. Like the Jews, the Christians (unless Gnostic) were opposed to syncretism. With the exception of the notion of baptism as a rebirth, Christians generally and significantly avoided the characteristic vocabularies of the mystery religions. The mysteries of Isis, Attis, Adonis, and (to some extent, perhaps) Mithra were basically fertility rites to ensure good crops. They answered to needs different from those addressed by the Christian gospel. A Mithraic rite with bread and water was noted by Justin Martyr as a counterpart of the bread and wine of the Christian Eucharist. The spring rites mourning Attis' death and then celebrating his revival at the festival known as the Hilaria offered a parallel to the ceremonies of Holy Week and Easter as developed in the 4th century. The point where parallel can be treated as influence, however, is a delicate matter to determine. Between Christianity and the pagan cults the most prominent difference consisted in the syncretistic tolerance of the latter; initiation into the mysteries of Isis did not mean renouncing allegiance to Apollo or Attis, whereas the Christian baptism required exclusive devotion.

Many converts naturally brought old attitudes with them into the church. Amulets and peasant superstitions were long the object of critical attention by the clergy. The Christians tried to provide counter-attractions by placing

Christian festivals on the same days of the year as pagan feasts. Solar monotheism was popular in late 3rd-century paganism, and soon the Western churches were keeping the winter solstice (December 25) as Christ's Nativity—the East kept January 6. Midsummer Day was replaced by the feast of John the Baptist. The church fought hard against the heathen celebration of January 1, but with little success. Only Easter (celebrating Christ's Resurrection) and Pentecost (celebrating the advent of the Holy Spirit) were feasts owing nothing to Gentile analogies for their origin; they both were derived from Jewish feasts. From the 5th century AD on, great pagan temples, such as the Parthenon in Athens, were gradually transformed into churches.

**Apologetics.** The Christian Apologists of the 2nd century sought to drive a wedge between the pagan religion that they abhorred and the Greek philosophy that, with occasional reservations, they welcomed. Second-century Platonism found it easy to think of Mind (nous) or Reason (Logos) as divine power immanent within the world. Philo of Alexandria had spoken of the Logos as mediating between the transcendent God and this created order. The Logos theology was developed by Justin Martyr both to make a positive evaluation of the best elements in the Greek philosophical tradition and to make the incarnation of Christ intelligible to the Greek mind. But the Apologists upset some of their fellow Christians by talking of the divine Logos as virtually a second God beside the Father and thus compromising the monotheism that the orthodox were defending against Gnostic dualism. The critics of the Logos theology, labeled Monarchians, affirmed that Father, Son, and Spirit were one God; the three names were epithets, not substantives. In the 3rd century a Roman presbyter, Sabellius, was excommunicated for this opinion, and the defenders of the Logos theology ousted the opponents of speculative apologetics. Clement of Alexandria and Origen provided the Greek churches with a Platonizing theology that was strongly opposed to the Monarchian position.

## THE EARLY LITURGY, THE CALENDAR, AND THE ARTS

Christian worship and practices

Paul's letters mention worship on the first day of the week. In John's Apocalypse, Sunday is called "the Lord's day." The weekly commemoration of the Resurrection replaced for Christians the synagogue meetings on Saturdays; the practice of circumcision was dropped, and initiation was by baptism; continuing membership of the church was signified by weekly participation in the Eucharist. Baptism in water in the name of Father, Son, and Holy Spirit was preceded by instruction (catechesis) and fasting. Persons about to be baptized renounced evil and, as they made the declaration of faith, were dipped in water; they then received by anointing and by the laying on of hands (confirmation) the gift of the Holy Spirit and incorporation within the body of Christ, thus concluding the entire rite. Only the baptized were allowed to be admitted to the Eucharist, when the words of Jesus at the Last Supper were recalled; the Holy Spirit was invoked upon the people of God making the offering, and the consecrated bread and wine were distributed to the faithful. Accounts of these rites are given in the works of Justin (*c.* 150) and especially in the *Apostolic Tradition* of Hippolytus of Rome (*c.* 220).

To fall into a grave fault after baptism entailed exclusion. Excommunicated persons would continue to attend for the first part of the service, which included psalms, readings, and a sermon. Montanists, such as Tertullian, and the Roman schismatic Novatian denied the church's power to grant absolution. Even Cyprian of Carthage found Novatian easier to criticize for schism than for rigorism. But in the 3rd century a system of public penance emerged; it was allowed once a lifetime under condition of ascetic discipline. Penitents were restored to fellowship with church members by the laying on of hands. In time, less arduous and less public severities came to be required.

Before the 4th century, worship was in private houses. A house church of AD 232 has been excavated at Doura-Europus on the Euphrates. Whereas pagan temples were intended as the residence of the god, churches were designed for the community. The rectangular basilica with an apse (semicircular projection to house the altar) was found especially suitable. The Doura-Europus church has Gospel scenes on the walls. But many Old Testament heroes also appear in the earliest Christian art; Jewish models probably were followed. The artists also adapted conventional pagan forms (good shepherd; praying persons with hands uplifted). Fishing scenes, doves, and lyres also were popular. In themselves neutral, they carried special meaning to the Christians.

Vincenzo Biolghini



The Good Shepherd (centre), Orantes with uplifted hands, and (in the four quarters) the story of Jonah with figures derived from Greek mythological prototypes. Ceiling painting in the catacombs of SS. Peter and Marcellinus, Rome, 3rd century.

The words of several pre-Constantinian hymns survive (*e.g.*, "Shepherd of tender youth," by Clement of Alexandria), but only one with musical notation (Oxyrhynchus papyrus 1786 of the 3rd century).

The earliest Christians wrote to convert or to edify, not to please. Their literature was not produced with aesthetic intentions. Nevertheless, the pulpit offered scope for oratory (as in Melito of Sardis' *Homily on the Pascha, c.* 170). Desire for romance and adventure was satisfied by apocryphal Acts of the Apostles, recounting their travels, with continence replacing love. Justin and Irenaeus did not write for high style but simply to convey information. Apologists hoping for well-educated readers, however, could not be indifferent to literary tastes. By AD 200 the most graceful living writer of Greek literature was Clement of Alexandria, the liveliest writer of Latin, Tertullian. Wholly different in temperament (Clement urbane and allusive, Tertullian vigorous and vulgar), both men wrote distinguished prose with regard to form and rhetorical convention.

By the 3rd century the Bible needed explanation. Origen of Alexandria set out to provide commentaries and undertook for the Old Testament a collation of the various Greek versions with the original Hebrew. Many of his sermons and commentaries were translated into Latin by Tyrannius Rufinus and Jerome (*c.* 385–400); their learning and passionate mystical aspiration shaped Western medieval exegesis (critical interpretive methods).

## THE ALLIANCE BETWEEN CHURCH AND EMPIRE

Constantine the Great, declared emperor at York, Britain (306), was converted to Christianity (312), became sole emperor (324), virtually presided over the ecumenical Council of Nicaea (325), founded the city of Constantinople (330), and died in 337. In the 4th century he was regarded as the great revolutionary, especially in religion. He did not make Christianity the religion of the empire, but his foundation of Constantinople (conceived to be

Constantinople as the new Rome

the new Rome) as a Christian city profoundly affected the future political and ecclesiastical structure. Relations with old Rome were not to be cordial either in matters of church or state. Despite massive legislation (some attempting to express Christian ideals—*e.g.,* making Sunday a rest day), he failed to check the drastic inflation that began about 250 and that soon created deep unrest and weakened the empire before the barbarian invasions of the 5th century.

Constantine brought the church out of its withdrawal from the world to accept social responsibility and helped pagan society to be won for the church. On both sides, the alliance of the church and emperor evoked opposition, which among the Christians emerged in the monks' retirement to the desert. Except for the brief reign of Julian the Apostate (361–363), pagans relapsed into passive resistance. The quietly mounting pressure against paganism in the 4th century culminated in the decrees of Emperor Theodosius I (reigned 379–395), who made orthodox Christianity an ingredient of good citizenship. Under Theodosius many pagan temples were closed or even destroyed (*e.g.,* the Alexandrian Serapeum). But until the time of Justinian (reigned 527–565), pagans were largely unmolested by the government. Heretics were more harshly treated. Ecclesiastical censures (from 314 on) were often enforced by the civil penalty of exile. One heretic, Priscillian, was even executed for witchcraft (385), but in the face of vehement church protests.

The link between church and state was expressed in the civil dignity and insignia granted to bishops, who also began to be entrusted with ambassadorial roles. By 400 the patriarch of Constantinople (to his avowed embarrassment) enjoyed precedence at court before all civil officials. In the writings of Ambrose (bishop of Milan, 374–397), "Roman" and "Christian" are almost synonyms. The Arian controversy (involving a denial of the divinity of Jesus) developed into a conflict between church and state when the emperor Constantius was supporting Arianism; and Ambrose enforced upon Theodosius submission to the church as its son, not its master. With an orthodox emperor, however, most Christians thought of church and empire as virtually coterminous.

Relationships with the barbarians

The church was significantly slow to undertake missionary work beyond the frontiers of the empire. The Goth Ulfilas converted the Goths to Arianism (*c.* 340–350) and translated the Bible, omitting, as unsuitable, warlike passages of the Old Testament. The Goths passed their Arian faith on to other Germanic tribes, such as the Vandals. (The first tribe to become Catholic was the Franks, in about 506, soon to be followed by the Visigoths.) In the 5th century the Western provinces were overrun by the barbarian Goths, Vandals, and Huns. The Roman Army had long drawn its recruits from the barbarian tribesmen and was itself now under barbarian generals. Theodosius I's will placed his two sons under the guardianship of the barbarian general Stilicho, who effectively ruled until they were able to assume responsibility. In the 5th century, Western emperors exercised less power than generals, and the imperial succession ended when a German leader, Odoacer, decided (476) to rule without an emperor. The end of the line of Western emperors made little difference to either church or state. In the West the position of the papacy was enhanced by the decline of state power, and this prepared the way for the popes' temporal sovereignty over parts of Italy (which they retained from the 7th to the 19th century).

The barbarian invasions destroyed Western schools. Specifically church schools were first created in late antiquity. The main preservers and transmitters of ancient culture, however, were the monks. Monasticism had begun in the Egyptian desert in the 4th century with Anthony the Hermit and with Pachomius, the first organizer of an ascetic community under a rule of obedience. Basil, bishop of Caesarea Cappadociae (370–379), rejected the hermit ideal and insisted on communities with a rule safeguarding the bishop's authority and with concrete acts of service to perform (*e.g.,* hospital work and teaching). The monastic ideal quickly spread to the West but owed its decisive shape there to John Cassian of Marseille (*c.* 360–435) and Benedict of Nursia (*c.* 480–*c.* 547). The manual work of monks often was the copying of manuscripts. Benedict's contemporary Cassiodorus (*c.* 490–*c.* 585) had the works of classical authors copied (*e.g.,* Cicero and Quintilian) as well as Bibles and the works of the early Church Fathers.

## THEOLOGICAL CONTROVERSIES OF THE 4TH AND 5TH CENTURIES

**Western controversies.** Until about 250 most Western Christian leaders spoke Greek, not Latin (*e.g.,* Irenaeus and Hippolytus). The main Latin theology came not from Rome but from North Africa (*e.g.,* Tertullian and Cyprian). Tertullian wrote *Against Praxeas,* in which he discussed the doctrines of the Trinity and the Person of Christ. But in 251 Novatian's schism at Rome diverted interest away from speculative theology to juridical questions about the membership of the church and the validity of sacraments. These questions led to a schism between Rome and the churches of North Africa, which centred on a controversy at Carthage over ideas espoused by Donatus (313). The Donatist issue, which raised questions about the validity of the sacraments, involved the theology of Cyprian (bishop of Carthage, 248–258) and dominated all North African church life. Cyprian and the Donatists said that the validity of the sacraments depended on the worthiness of the minister; Rome and North African Christians in communion with Rome said that it did not because the sacraments received their validity from Christ, not man. Thus, even if inefficacious, baptism could be validly administered by a schismatic. Much of the great theologian Augustine's energies as bishop of Hippo (from 396 to 430) went into trying to settle the Donatist issue, in which he finally despaired of rational argument and reluctantly came to justify the use of limited coercion.

Donatism, Pelagianism, Arianism, and Monophysitism

The other major controversy of the Western Church was a more confused issue, namely, whether faith is caused by divine grace or human freedom. Augustine ascribed all credit to God. The British monk Pelagius protested that Augustine was destroying responsibility and denying the capacity of man to do what God commands. Both men applied inappropriate, impersonal categories of thought to the problem; and though Pelagianism was condemned, several of the extreme positions of Augustine (especially on predestination and the transmission of original sin) failed to receive the church's cordial endorsement.

**Eastern controversies.** In the Greek East, the 4th century was dominated by controversy about the propositions of Arius, an Alexandrian presbyter (*c.* 250–336), that the incarnate Lord who was born, wept, suffered, and died could not be one with the transcendent first cause of creation who is beyond all suffering. The Council of Nicaea (325) condemned Arianism and affirmed the Son of God to be identical in essence with the Father. As this formula included no safeguard against Monarchianism, a long controversy followed, especially after Constantine's death (337). Athanasius, bishop of Alexandria (reigned 328–373), fought zealously against Arianism in the East and owed much to Rome's support, which made the controversy add to the tensions between East and West. These tensions survived the settlement of the Arian dispute when the Council of Constantinople (381) eliminated Arianism in the East but also asserted Constantinople to be the second see of Christendom, as the new Rome. This assertion was unwelcome to Alexandria, traditionally second city of the empire, and to Rome because it implied that the dignity of a bishop depended on the secular standing of his city. Rivalry between Alexandria and Constantinople led to the fall of John Chrysostom, patriarch of Constantinople (reigned 398–404), when he appeared to support Egyptian monks who admired the controversial theology of Origen. It became a major feature of the emerging Christological debate (the controversy over the nature of Christ).

The Christological controversy stemmed from the rival doctrines of Apollinaris of Laodicea (flourished 360–380) and Theodore of Mopsuestia (*c.* 350–428), representatives of the rival schools of Alexandria and Antioch, respectively. At the Council of Ephesus (431), led by Cyril, patriarch of Alexandria (reigned 412–444), an extreme Antiochene Christology—taught by Nestorius, patriarch

of Constantinople—was condemned for saying that the man Jesus is an independent person beside the divine Word and that therefore Mary, the mother of Jesus, may not properly be called mother of God (Greek *theotokos,* or "God-bearer"). Cyril's formula was "one nature of the Word incarnate." A reaction led by Pope Leo I (reigned 440–461) against this one-nature (Monophysite) doctrine culminated in the Council of Chalcedon (451), which affirmed Christ to be two natures in one person (hypostasis). Thus, the Council of Chalcedon alienated Monophysite believers in Egypt and Syria.

During the next 250 years the Byzantine emperors and patriarchs desperately sought to reconcile the Monophysites. Three successive attempts failed: (1) under the emperor Zeno (482) the *Henotikon* (union formula) offended Rome by suggesting that Monophysite criticism of Chalcedon might be justified; (2) under the emperor Justinian the Chalcedonian definition was glossed by condemning the "Three Chapters," which includes the writings of Theodore of Mopsuestia, Theodoret, and Ibas, all strong critics of Cyril of Alexandria's theology and of Monophysitism; the Syrian Monophysite Jacob Baradaeus reacted to this by creating a rival Monophysite episcopate and permanent schism; (3) under the emperor Heraclius (reigned 610–641) the Chalcedonians invited the Monophysites to reunite under the formula that Christ had two natures but only one will (Monothelitism), but this reconciled almost no Monophysites and created divisions among the Chalcedonians themselves. Chalcedon's "two natures" continues to be rejected by the Armenian Apostolic Church, Coptic Orthodox Church, Ethiopian Orthodox Church, and Syrian Orthodox Patriarchate of Antioch (Syrian Jacobites).

### POPULAR CHRISTIANITY IN THE LATE EMPIRE

The continuity of pre-Christian antiquity and Christian society is nowhere more apparent than in popular religious practice. Pagans were normally devoted to local shrines of particular gods. The church tried to meet this psychological need by establishing shrines of martyrs. The martyr cult, a matter of private devotion from 150 until 250, became so popular after the Decian persecution that official control was required. Invocation of Mary as "mother of God" is first attested in a 3rd-century papyrus. At Rome the shrines of Peter and Paul, where Constantine built basilicas, attracted many pilgrims. The holy places in Jerusalem and Bethlehem, however, were preeminent.

Emergence of the cult of martyrs and saints

Preachers might warn that pilgrimage did not necessarily bring one nearer to God and that one must not worship the martyrs being venerated, but at the popular level such exhortations seemed sophisticated. The bones of martyrs and other holy persons were so treasured that a traffic in bogus relics was created. By 400, particular saints were being invoked for particular needs (one for health, another for fertility, travel, prediction, or the detection of perjury). When the barbarian leader Alaric's Goths sacked Rome (410), citizens asked why Peter and Paul had failed to protect their city.

Pagan critics said that the old gods, true givers of success and miracle, were offended by neglect. To meet such criticisms the churches found it necessary to provide similar assurances of success, miraculous cures, and patron saints. By the 6th century, wonder-working shrines, cloths that had touched holy relics, and pictures (icons) were invested with numinous (spiritual) power. Because of the antielitist ideology of the Christian tradition, even highly educated figures such as Augustine and Pope Gregory I the Great (reigned 590–604) were sympathetic to this popular movement. It became a means of winning the barbarian tribesmen.

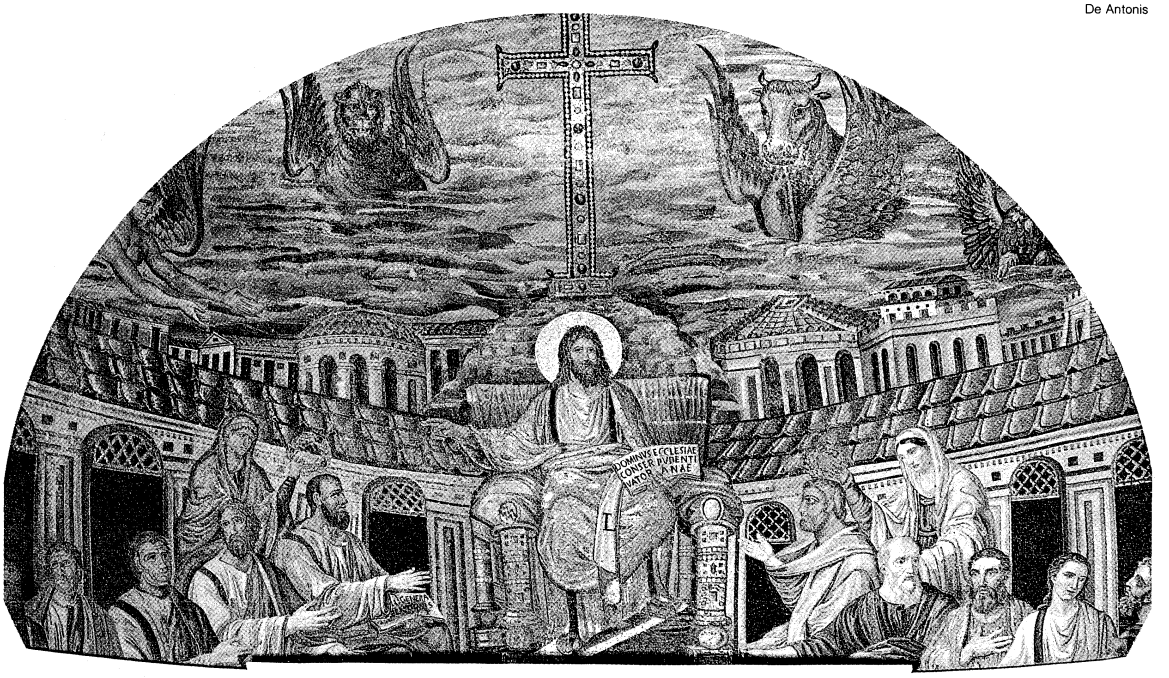
### LITURGY AND THE ARTS AFTER CONSTANTINE

The veneration of martyrs and the growth of pilgrimages stimulated liturgical elaboration. Great centres (Jerusalem and Rome, in particular) became models for others, which encouraged regional standardization and cross-fertilization. Though the pattern of the eucharistic liturgy was settled by the 4th century, there were many variant forms, especially of the central prayer called by the Greeks *anaphora* ("offering") and by the Latins *canon* ("prescribed form"). Liturgical prayers of Basil of Caesarea became widely influential in the East. Later, liturgies were ascribed to local saints and heroes: Jerusalem's to St. James, Alexandria's to St. Mark, and Constantinople's to John Chrysostom. The spirit of Greek liturgies encouraged rich and imaginative prose. Latin style was restrained, with epigrammatic antitheses; and the Roman Church changed from Greek to Latin about AD 370. The Canon of the Latin mass as used in the 6th century was already close to the form it has since retained.

Development of new forms of worship

Music also became elaborate, with antiphonal psalm chanting. Some reaction came from those who believed that the music was obscuring the words. Both Athanasius of Alexandria and Augustine defended music on the

De Antonis




Christ as Ruler, with the Apostles and Evangelists (represented by the beasts). The female figures are believed to be either Santa Pudenziana and Santa Praxedes or symbols of the Jewish and Gentile churches. Mosaic in the apse of Santa Pudenziana, Rome, AD 401–417.

condition that the sense of the words remained primary in importance. The Latin theologians Ambrose of Milan, Prudentius, and Venantius Fortunatus provided Latin hymns of distinction. The ascription of the Roman chants (Gregorian) to Pope Gregory I the Great was first made in the 9th century. In the Greek East in the time of Justinian, Romanos Melodos created the kontakion, a long poetic homily.

Architecture was stimulated by Constantine's great buildings at Jerusalem and Rome. The exteriors of these churches remained simple, but inside they were richly ornamented with marble and mosaic, the decoration being arranged on a coherent plan to represent the angels and saints in heaven with whom the church on earth was joining for worship. An enormous number of churches built in and after the 4th century have been excavated. The outstanding buildings that survive largely intact belong to the age of Justinian (6th century) and are at Constantinople and Ravenna.

The veneration of saints led to the production of a specific category of literature known as hagiography. If available, authentic tradition would be used; but if there was none, the writers felt quite free to create a biography from conventional materials and elements of folklore. The lives of saints belong to the poetry of the Middle Ages but are important to the historian as documents of social history.

Historical and polemical writings

The first church historian was Eusebius, bishop of Caesarea in the 4th century, who collected records up to the advent of Constantine. His work was translated and continued in Latin by Tyrannius Rufinus of Aquileia. The history of the church from Constantine to about 430 was continued by three Greek historians: Socrates Scholasticus, Sozomen, and Theodoret (whose works were adapted for the Latin world by Cassiodorus). Ecclesiastical history from 431 to 594 was chronicled by Evagrius Scholasticus. The consequences of Chalcedon as interpreted by Monophysite historians were recorded by Timothy Aelurus, Zacharias Scholasticus, and John of Nikiu.

The monastic movement produced its own special literature, especially the classic *Life of St. Antony* by Athanasius, the collections of sayings of the Desert Fathers, John Climacus' *Heavenly Ladder,* and Moschus' *Spiritual Meadow.*

The Arian and Christological controversies produced important polemical writers—Athanasius, the three Cappadocian Fathers (Basil, Gregory of Nazianzus, and Gregory of Nyssa), Cyril of Alexandria, and Theodoret. After 500, Monophysite theology had eminent figures—Severus of Antioch and the Alexandrian grammarian John Philoponus, who was also a scientist and a commentator on Aristotle. But much theology was non-polemical—*e.g.,* catechesis and biblical commentaries. In the 6th century, "chains" (*catenae*) began to be produced in which the reader was given a summary of the exegesis of a succession of commentators on each verse.

In the West, Hilary of Poitiers, Ambrose of Milan, and, above all, the incomparable scholar Jerome (translator of the standard Latin Bible, or Vulgate) gave Latin theology confidence and so made possible the massive achievement of Augustine—the exquisite prose of his *Confessions* and his majestic treatises *On the Trinity* and *The City of God.*

## POLITICAL RELATIONS BETWEEN EAST AND WEST

Reciprocal influences between theology and political thought and action

The old tensions between East and West were sharpened by the quarrels about Chalcedon. In Rome every concession made by Constantinople toward the Monophysites increased the distrust. Justinian's condemnation of the Three Chapters (Fifth Council, Constantinople, 553) was forced on a reluctant West, where it created temporary schisms but was eventually accepted. From the time of Pope Gregory I the Great the papacy—encouraged by the successful mission to the Anglo-Saxons—was looking as much to the Western barbarian kingdoms as to Byzantium.

In the 7th century the Eastern Empire was fighting for its life, first against the Persians and then the Arabs, and the Balkans were occupied by the Slavs. The submergence of Alexandria, Antioch, and Jerusalem under Muslim rule left the patriarch of Constantinople with enhanced authority, whereas the Slav invasions drove a wedge between East and West that encouraged separate developments. The attempts of the Byzantine emperors to force the papacy to accept the Monothelite (one-will) compromise produced a martyr pope, Martin (reigned 649–655); the story of his tortures did nothing to make Rome love the Byzantines. When the Monothelite heresy was finally rejected at the Sixth Council (Constantinople, 680–681), the imprudent pope Honorius (reigned 625–638), who had supported Monothelitism, was expressly condemned, which distressed Roman defenders of papal prerogatives. Greek hostility to the West became explicit in the canons of a council held at Constantinople (Quinisext, 692) that claimed to have ecumenical status but was not recognized in Rome.

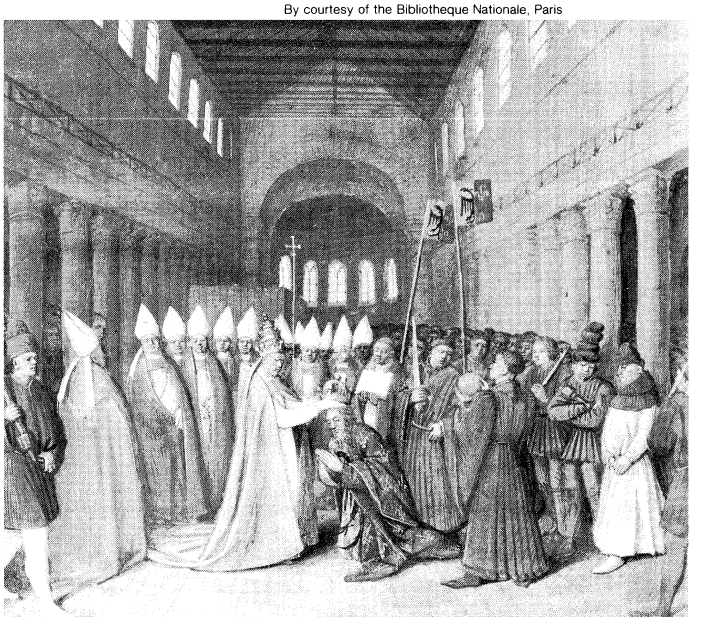
From 726 on, Byzantium was absorbed in the iconoclastic (destruction of images) debate, which became a struggle not only to keep icons but also to combat the subjection of the church to the will of the emperor. The imperial attack on images was severely criticized in the West. Yet, after the Greek iconoclasts were condemned at the Seventh Council (Nicaea, 787), the bishops of the Frankish king Charlemagne at the synod of Frankfurt in Germany (794)—with the reluctant consent of Pope Adrian I (reigned 772–795)—censured the decision. A renewed upsurge of iconoclasm in the East (815–843) produced a counterreaction in the West, and ultimately the West accepted the decisions of the Seventh Council. Icons were differently evaluated in the Western churches, where holy pictures were viewed as devotional aids, not, as was the case in the East, virtually sacramental media of salvation.

The greatest protagonist of icons was John of Damascus, an Arab monk in Muslim Palestine, who was the author of an encyclopaedic compendium of logic and theology. Within the empire, Theodore Studites, abbot of the Studium (monastery) near Constantinople, vigorously attacked iconoclasm; he also led a revival of monasticism and stressed the importance of copying manuscripts. His ideals passed to the monastic houses that began to appear on Mount Athos from 963 onward.

The hostility between the iconoclast emperors and the popes encouraged the 8th-century popes to seek a protector. This was provided by the rise of Charles Martel (reigned 719–741) and the Franks. The Frankish kings guarded Western Church interests, and the papal–Frankish alliance reached its climax in the papal coronation of Charlemagne as the first Holy Roman emperor at Rome on Christmas Day, 800—the Holy Roman Empire lasted until 1806. Charlemagne exercised immense authority

The Holy Roman Empire

By courtesy of the Bibliotheque Nationale, Paris



Pope Leo III crowning Charlemagne as emperor, AD 800, miniature in the *Grandes Chroniques de France,* manuscript illuminated by Jean Fouquet, *c.* 1460. In the Bibliothèque Nationale, Paris (MS. fr. 6465).

over the Western Church, and the revival of church life produced controversies about predestination (Gottschalk, Erigena, Hincmar of Reims) and the Eucharist (Paschasius Radbertus, Ratramnus, Rabanus Maurus). The Christological controversy was revived with a Spanish dispute as to whether Christ was adopted to be Son of God.

In the chaos of the rapid Frankish decline, the papacy was glad to look again to Constantinople for protection. The emperor Basil I (reigned 867–886), founder of the Macedonian dynasty, could not prevent the Arabs from taking Sicily, but he was able to reestablish Byzantine control in southern Italy.

In the 10th century, however, the West passed under the control of the Ottonian dynasty in Germany. The Ottos, accustomed to the system in which great landowners built and owned the churches on their estates as private property, treated Rome and all important sees in this spirit. Bishops were appointed on royal nomination and forbidden to appeal to Rome.

The rise of Islām and the Arab campaign to subjugate unbelievers by military conquest broke upon the Byzantine Empire in 634, just as it was exhausted after defeating Persia. The will to resist was wholly absent. Moreover, the provinces initially overrun, Syria (636) and Egypt (641), were already alienated from the Byzantine government that was persecuting Monophysites in those areas. In 678 and again in 718, the Arabs were at the walls of Constantinople. In the West their defeat by Charles Martel at Poitiers, Fr. (732), limited their advance to the Pyrenees. The Monophysite Copts in Egypt and Syrian Jacobites (followers of Jacob Baradaeus) soon found that they enjoyed greater toleration under Muslim Arabs than under Chalcedonian Byzantines, just as in later times the Greeks were to discover more religious freedom under Turkish than under Latin rule. In the 8th century the Muslims were more a military than a theological threat, and a considerable time passed before Christian and Muslim theologians engaged in serious dialogue.

### LITERATURE AND ART OF THE "DARK AGES"

The Monothelite and iconoclastic controversies produced herculean theological endeavours: the criticism of Monothelitism by the monk Maximus the Confessor (580–662) was based upon subtle and very careful considerations of the implications of Chalcedon. The great opponents of iconoclasm, John of Damascus and Theodore Studites, also composed hymns and other theological treatises. Greek mystical theology had an outstanding representative in Symeon the New Theologian (949–1022), abbot of St. Mamas at Constantinople, whose doctrines about light visions anticipated the hesychasm (quietistic prayer methods) of Gregory Palamas in the 14th century. But the most learned theologian of the age was beyond doubt the patriarch Photius (see below *The Photian schism*).

The ending of iconoclasm

Iconoclasm was not an anti-intellectual, anti-art movement. The iconoclasts everywhere replaced figures with the cross or with exquisite patterns. The ending of iconoclasm in 843 (the restoration of orthodoxy), however, liberated the artists adept in mosaic and fresco to portray figures once again, spurring a new revival of decoration. Music also became more elaborate; the kontakion was replaced by the *kanon,* a cycle of nine odes, each of six to nine stanzas and with a different melody. The *kanon* gave more scope to the musicians by providing greater variety. Byzantine hymns were classified according to their mode, and the mode changed each week. Besides John of Damascus and Theodore Studites, the great hymn writers of this period were Cosmas of Jerusalem and Joseph of Studium.

The so-called Dark Ages in the West produced virtually no sculpture or painting—other than illuminated manuscripts, of which marvelous specimens were made (*e.g.,* the Book of Kells and the Lindisfarne Gospels). The Irish and Anglo-Saxon monks did not construct noble buildings but knew how to write and to illuminate a book. In the age of Charlemagne exquisite calligraphy was continued (*e.g.,* the Utrecht Psalter), with intricate ivory and metalwork of superb finesse. Great buildings also began to emerge, partly based on Byzantine models, such as the churches at Ravenna. The Ottonian renaissance in Germany encouraged even more confidently the erection of church buildings, producing such masterpieces as the surviving cathedrals at Hildesheim and Spires and setting out a characteristically German style of architecture.

The barbarian kingdoms soon produced their own Christian literature: Gregory of Tours wrote the history of the Franks, Isidore of Seville that of the Visigoths, and Cassiodorus that of the Ostrogoths. Isidore, utilizing his vast reading, compiled encyclopaedias on everything from liturgical ceremonies to the natural sciences. The outstanding figure of this incipient "nationalist" movement was the English monk Bede, whose *Ecclesiastical History of the English People* was completed in 731 and whose exegetical works came to stand beside Augustine and Gregory I the Great as indispensable for the medieval student.

### MISSIONS AND MONASTICISM

Conversion of northern and eastern tribes and peoples

The Arian barbarians soon became Catholics, including, by 700, even the Lombards in northern Italy. There remained immense areas of Europe, however, to which the Gospel had not yet been brought. Gregory I the Great evangelized the Anglo-Saxons, who in turn sent missionaries to northwestern Europe—Wilfrid and Willibrord to what is now The Netherlands, and Boniface to Hesse, Thuringia, and Bavaria. In consequence of Boniface's work in Germany, a mission to Scandinavia was initiated by Ansgar (801–865), and the mission reached Iceland by 996. In the 10th century the mission from Germany moved eastward to Bohemia, to the Magyars, and (from 966) to the Poles. By 1050 most of Europe was under Christian influence with the exception of Muslim Spain.

In the Byzantine sphere, early missions went to the Hunnish tribesmen north of the Caucasus. The Nestorians, entrenched in Persia, carried the Gospel to the Turkmen and across Central Asia to China. In the 9th century the mission to the Slavs began with the work of Cyril and Methodius, who created a Slavonic alphabet and translated the Bible into the Slavonic language. Although their labours in Moravia were undermined by Frankish clergy, it was their achievement that made possible the faith and medieval culture of both Russia and Serbia.

The Benedictine Rule—initiated by Benedict of Nursia—succeeded in the West because of its simplicity and restraint; more formidable alternatives were available in the 6th century. By 800, abbeys existed throughout western Europe, and the observance of Benedict's Rule was fostered by Charlemagne and his son Louis the Pious. These houses, such as Bede's monastery at Jarrow (England) or the foundations of Columban (*c.* 543–615) at Luxeuil (France) and Bobbio (Italy), became centres of study and made possible the Carolingian renaissance of learning. In this renaissance the 8th-century English scholar Alcuin and his monastery at Tours occupy the chief place. Around monasteries and cathedrals, schools were created to teach acceptable Latin, to write careful manuscripts, and to study not only the Bible and writings of the Church Fathers but also science. Scribes developed the beautiful script that was known as Carolingian minuscule. The Carolingian renaissance was short-lived, however, and decay began to set in (850–950) and was not checked until the foundation of the monastery at Cluny (France) in 909.

Monasticism in 9th-century Byzantium was centred upon the Studites, who came to be a faction against the court. A remoter and otherworldly asceticism developed with the foundation of monasteries on Mount Athos (Greece) from 963 onward. A distinctive feature of Athonite monasticism was that nothing female was to be allowed on the peninsula.

### THE PHOTIAN SCHISM AND THE GREAT EAST–WEST SCHISM

**The Photian schism.** The end of iconoclasm (843) left a legacy of faction. Ignatius, patriarch of Constantinople intermittently from 847 to 877, was exiled by the government in 858 and replaced by Photius, a scholarly layman who was head of the imperial chancery—he was elected patriarch and ordained within six days. Ignatius' supporters dissuaded Pope Nicholas I (reigned 858–867) from recognizing Photius. Nicholas was angered by Byzantine missions among the Bulgars, whom he regarded as be-



Virgin Mary with (left) Justinian, holding a model of Hagia Sophia, and (right) Constantine, holding a model of the city of Constantinople. Mosaic from Hagia Sophia, 9th century.
By courtesy of the Dumbarton Oaks Center for Byzantine Studies, Washington, D.C.

longing to his sphere. When Nicholas wrote to the Bulgars attacking Greek practices, Photius replied by accusing the West of heretically altering the creed in saying that the Holy Spirit proceeds from the Father and from the Son (*Filioque*). He declared Pope Nicholas deposed (867), but his position was not strong enough for such imprudence.

A new emperor, Basil the Macedonian, reinstated Ignatius; and in 869 Nicholas' successor, Adrian II (reigned 867–872), condemned Photius and sent legates to Constantinople to extort submission to papal supremacy from the Greeks. The Greeks resented the papal demands, and when Ignatius died in 877 Photius quietly became patriarch again. Rome (at that moment needing Byzantine military support against Muslims in Sicily and southern Italy) reluctantly agreed to recognize Photius, but on the condition of an apology and of the withdrawal of Greek missions to the Bulgars. Photius acknowledged Rome as the first see of Christendom, discreetly said nothing explicitly against the *Filioque* clause, and agreed to the provision that the Bulgars could be put under Roman jurisdiction providing that Greek missions were allowed to continue.

The main issue in the Photian schism was whether Rome possessed monarchical power of jurisdiction over all churches (as Nicholas and Adrian held), or whether Rome was the senior of five semi-independent patriarchates (as Photius and the Greeks thought) and therefore could not canonically interfere with the internal affairs of another patriarchate.

**The great East–West schism.** The mutual distrust shown in the time of Photius erupted again in the middle of the 11th century after papal enforcement of Latin customs upon Greeks in southern Italy. The patriarch of Constantinople, Michael Cerularius, closed Latin churches in Constantinople as a reprisal. Cardinal Humbert came from Italy to protest, was accorded an icy reception, and left a bull of excommunication (July 16, 1054) on the altar of the great church of Hagia Sophia. The bull anathematized (condemned) Michael Cerularius, the Greek doctrine of the Holy Spirit, the marriage of Greek priests, and the Greek use of leavened bread for the Eucharist.

The significance of the date 1054

At the time, the breach was treated as a minor storm in which both sides had behaved with some arrogance. As Greeks and Latins became more estranged, however, people looked back on the events of 1054 as the moment of the final breach between East and West. (Not until Dec. 7, 1965, were the mutual excommunications of 1054 abolished, by Pope Paul VI and the ecumenical patriarch Athenagoras I.) (H.Cha.)

## FROM THE SCHISM TO THE REFORMATION

**Differences between the Eastern and Western churches.** A major factor in the consolidation and expansion of Christianity in the West was the growth in the prestige and power of the bishop of Rome. Pope Leo I the Great made the primacy of the Roman bishop explicit both in theory and in practice and must be counted as one of the most important figures in the history of the centralization of authority in the church. The next such figure was Gregory I the Great, whose work shaped the worship, the thought, and the structure of the church as well as its temporal wealth and power.

Even while still a part of the universal church, Byzantine Christianity had become increasingly isolated from the West by difference of language, culture, politics, and religion and followed its own course in shaping its heritage. The Eastern churches never had so centralized a polity as did the church in the West but developed the principle of the administrative independence or "autocephaly" of each national church. During the centuries when Western culture was striving to domesticate the German tribes, Constantinople, probably the most civilized city in Christendom, blended classical and Christian elements with a refinement that expressed itself in philosophy, the arts, statecraft, jurisprudence, and scholarship. A thinker such as Michael Psellus in the 11th century, who worked in several of these fields, epitomizes this synthesis. It was from Byzantine rather than from Roman missionaries that Christianity came to most of the Slavic tribes, including some who eventually sided with Rome rather than Constantinople; Byzantium was also the victim of Muslim aggressions throughout the period known in the West as the Middle Ages. Following the pattern established by the

emperors Constantine and Justinian, the relation between church and state in the Byzantine empire coordinated the two in such a way as to sometimes subject the life and even the teaching of the church to the decisions of the temporal ruler—the phenomenon often, though imprecisely, termed caesaropapism.

Caesaropapism

All these differences between the Eastern and Western parts of the church, both the religious differences and those that were largely cultural or political, came together to cause the schism between the two. The break in 1054 was followed by further evidences of alienation—in the 13th century, in the sack of Constantinople by Western Christians in 1204 and the establishment of the Latin patriarchate there; and in the 15th century, after the failure of the union of Florence and after the fall of Constantinople to the Turks in 1453.

**Papacy and empire.** Conflict with the East was both a cause and an effect of the distinctive development of Western Christianity during the Middle Ages. If popes Leo I and Gregory I may be styled the architects of the medieval papacy, popes Gregory VII (reigned 1073–85) and Innocent III (reigned 1198–1216) should be called its master builders. Gregory VII reformed both the church and the papacy from within, establishing the canonical and moral authority of the papal office when it was threatened by corruption and attack; in the pontificate of Innocent III the papal claims to universality reached their zenith at all levels of the life of the church. Significantly, both these popes were obliged to defend the papacy against the Holy Roman emperor and other temporal rulers. The battle between the church and the empire is a persistent theme in the history of medieval Christianity. Both the involvement of the church in feudalism and the participation of temporal rulers in the Crusades can be read as variations on this theme. Preoccupied as they usually are with the history of the church as an institution and with the life and thought of the leaders of the church, the documentary sources of knowledge about medieval Christianity make it difficult for the social historian to descry "the religion of the common man" during this period. Both the "age of faith" depicted by neo-Gothic Romanticism and the "Dark Ages" depicted by secularist and Protestant polemics are gross oversimplifications of history. Faith there was during the Middle Ages, and intellectual darkness and superstition too; but only that historical judgment of medieval Christianity is valid that discerns how subtly faith and superstition can be blended in human piety and thought.

**Medieval thought.** No product of medieval Christianity has been more influential in the centuries since the Middle Ages than medieval thought, particularly the philosophy and theology of Scholasticism, whose outstanding exponent was Thomas Aquinas (1224/25–1274). The theology of Scholasticism was an effort to harmonize the doctrinal traditions inherited from the Fathers of the early church and to relate these traditions to the intellectual achievements of classical antiquity. Because many of the early Fathers both in the East and in the West had developed their theologies under the influence of Platonic modes of thought, the reinterpretation of these theologies by Scholasticism required that the doctrinal content of the tradition be disengaged from the metaphysical assumptions of Platonism. For this purpose the recovery of Aristotle—first through the influence of Aristotelian philosophers and theologians among the Muslims, and eventually, with help from Byzantium, through translation and study of the authentic texts of Aristotle himself—seemed providential to the Scholastic theologians. Because it managed to combine a fidelity to Scripture and tradition with a positive, though critical, attitude toward the "natural" mind, Scholasticism is a landmark both in the history of Christianity and in the history of Western culture, as a symbol (depending upon one's own position) either of the Christianization of society and culture or of the betrayal of Christianity to the society and culture of the Middle Ages.

Scholasticism

**Reformation.** The latter interpretation of Scholasticism and of the medieval church itself animated the Protestant Reformation. Protestantism differed from the various protest movements during the later Middle Ages by the thoroughness of its polemic against the ecclesiastical, theological, and sacramental developments of Western Catholicism. Initially the Protestant Reformers maintained the hope that they could accomplish the reformation of the doctrine and life of the church from within, but this proved impossible either because of the intransigency of the church, the extremism of the Protestant movements, or the political and cultural situation—or because of all of these factors. The several parties of the Reformation may be conveniently classified according to the radicalism of their protest against medieval theology, piety, and polity. The Anglican Reformers, as well as Martin Luther and his movement, were, in general, the most conservative in their treatment of the Roman Catholic tradition; John Calvin and his followers were less conservative; the Anabaptists and other groups in the left wing of the Reformation were least conservative of all. Despite their deep differences, almost all the various Reformation movements were characterized by an emphasis upon the Bible, as distinguished from the church or its tradition, as the authority in religion; by an insistence upon the sovereignty of free grace in the forgiveness of sins; by a stress upon faith alone, without works, as the preconditions of acceptance with God; and by the demand that the laity assume a more significant place in both the work and the worship of the church.

Characterization of Reformation movements

The Reformation was launched as a movement within the established Christianity that had prevailed since Constantine. It envisaged neither schism within the church nor the dissolution of the Christian culture that had developed for more than a millennium. But when the Reformation was over, both the church and the culture had been radically transformed. In part this transformation was the effect of the Reformation; in part it was the cause of the Reformation. The voyages of discovery, the beginnings of a capitalist economy, the rise of modern nationalism, the dawn of the scientific age, the culture of the Renaissance—all these factors, and others besides, helped to break up the "medieval synthesis." Among these factors, however, the Reformation was one of the most important and, certainly for the history of Christianity, the most significant. For the consequences of the Reformation, not in intention but in fact, were a divided Christendom and a secularized West. Roman Catholicism, no less than Protestantism, has developed historically in the modern world as an effort to adapt historic forms to the implications of these consequences. Established Christianity, as it had been known in the West since the 4th century, ended after the Reformation, though not everywhere at once.

## MODERN CHRISTIANITY

Paradoxically, the end of "established Christianity" in the old sense resulted in the most rapid and most widespread expansion in the history of the church. The Christianization of the Americas and the evangelization of Asia, Africa, and Australasia for the first time gave geographic substance to the Christian title "ecumenical." Growth in areas and in numbers, however, need not be equivalent to growth in influence. Despite its continuing strength throughout the modern period, Christianity retreated on many fronts and lost much of its prestige and authority both politically and intellectually.

During the formative period of modern Western history, roughly from the beginning of the 16th to the middle of the 18th century, Christianity participated in many of the movements of cultural and political expansion. The explorers of the New World were followed closely by missionaries—that is, when the two were not in fact identical. Protestant and Roman Catholic clergymen were prominent in politics, letters, and science. Although the rationalism of the Enlightenment alienated many people from the Christian faith, especially among the intellectuals of the 17th and 18th centuries, those who were alienated often kept a loyalty to the figure of Jesus or to the teachings of the Bible even when they broke with traditional forms of Christian doctrine and life. Citing the theological conflicts of the Reformation and the political conflicts that followed upon these as evidence of the dangers of religious intolerance, representatives of the Enlightenment gradually introduced disestablishment, toleration, and religious liberty into most Western countries; in this movement

they have been joined by Christian individuals and groups that advocated religious freedom not out of indifference to dogmatic truth but out of a concern for the free decision of personal faith.

The Counter-Reformation

The state of Christian faith and life within the churches during the 17th and 18th centuries both reflected and resisted the spirit of the time. Even though the Protestant Reformation had absorbed some of the reformatory energy within Roman Catholicism, the theology and morals of the church underwent serious revision in the Roman Catholic Counter-Reformation. Fighting off the attempts by various countries to establish national Catholic churches, the papacy sought to learn from the history of the Reformation and to avoid the mistakes that had been made then. Protestantism in turn discovered that separation from Rome did not necessarily inoculate it against many of the trends that it had denounced in Roman Catholicism. The orthodox dogmatics of the 17th century both in Lutheranism and in the Reformed churches displayed many features of medieval Scholasticism, despite the attacks of the Reformers upon the latter. Partly as a compensation for the overemphasis of orthodoxy upon doctrine at the expense of morals, Pietism summoned Protestant believers to greater seriousness of faith and purpose. Valid though its summons was, the subjectivity of Pietism unwittingly played into the hands of its enemies, helping to make it possible for the rationalism of the Enlightenment to undermine traditional Christian belief.

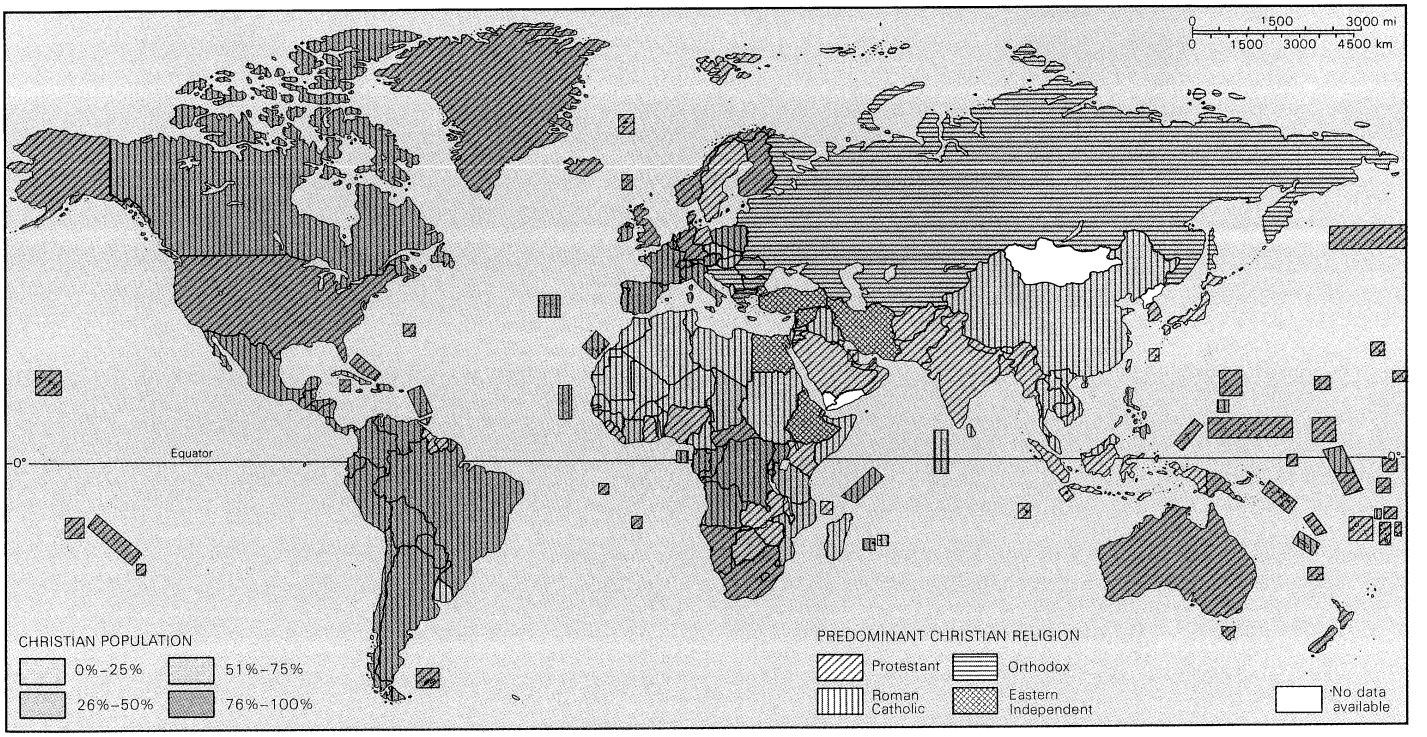
In alliance with the spirit of the Enlightenment, the revolutions of the 18th, 19th, and 20th centuries aided this process of undermining Christianity. Roman Catholicism in France, Eastern Orthodoxy in Russia, and Protestantism in former European colonies in Africa and Asia were identified—by their enemies if not also by themselves—as part of the ancien régime and were nearly swept away with it. As the discoveries of science proceeded, they clashed with old and cherished notions about the doctrine of creation, many of which were passionately supported by various leaders of organized Christianity. The age of the revolutions—political, economic, technological, intellectual—was an age of crisis for Christianity. It was also an age of opportunity. The critical methods of modern scholarship, despite their frequent attacks upon traditional Christian ideas, helped to produce editions of the chief documents of the Christian faith—the Bible and the writings of the Fathers and Reformers—and to arouse an unprecedented interest in the history of the church. The 19th century has been called the great century in the history of Christian missions, both Roman Catholic and Protestant. By the very force of their attacks upon Christianity, the critics of the church helped to arouse within the church new apologists for the faith, who creatively reinterpreted it in relation to the new philosophy and science of the modern period. The 20th century saw additional challenges to the Christian cause in the form of Communism, of resurgent world religions, and of indifference. Both the relation of church and state and the missionary program of the churches thus demanded reconsideration. But the 20th century also saw renewed efforts to heal the schisms within Christendom. The ecumenical movement began within Protestantism and Anglicanism, eventually included some of the Eastern Orthodox churches, and, especially since the second Vatican Council (1962–65), has engaged the sympathetic attention of Roman Catholicism as well.

## THE MODERN DENOMINATIONS

By the late 20th century Christianity had become the most widely disseminated faith on Earth. Virtually no nation has remained unaffected by the activities of Christian missionaries, although in many countries Christians are only a small fraction of the total population. Most of the countries of Asia and of Africa have Christian minorities, some of which, as in India and even in China, number several million members. The concentration of Christians, however, remains in the domain of Western culture. Each major division of Christianity—Eastern Orthodoxy, Roman Catholicism, and Protestantism—is treated in the *Macropædia* in a separate article where its history, tenets, and practices receive a fuller exposition than this article can give them and where a bibliography on the denominations of the division is supplied. The purpose here is to provide an overview of the principal divisions and thus to set the articles about the individual traditions into their proper context.

**Roman Catholicism.** The Roman Catholics in the world outnumber all other Christians combined. They are organized in an intricate system that spans the structure of the church from the local parish to the papacy. Under the central authority of the papacy, the church is divided into dioceses, whose bishops act in the name and by the authority of the pope but retain considerable administrative freedom within their individual jurisdictions; the principle

Church organization




World distribution of Christianity by country.

of "collegiality" articulated by the second Vatican Council has expanded that freedom. Similarly, the parish priest stands as the executor of papal and diocesan directives. Alongside the diocesan organization and interacting with it is a chain of orders, congregations, and societies; all of them are, of course, subject to the pope, but they are not directly responsible to the bishop as are the local parishes. It would, however, be a mistake to interpret the polity of the Roman Catholic Church in so purely an organizational manner as this. For Roman Catholic polity rests upon a mandate that is traced to the action of Jesus Christ himself, when he invested Peter and, through Peter, his successors with the power of the keys in the church. Christ is the invisible head of his church, and by his authority the pope is the visible head.

This interpretation of the origin and authority of the church determines both the attitude of Roman Catholicism to the rest of Christendom and its relation to the social order. Believing itself to be the true church of Jesus Christ on earth, it cannot deal with other Christian traditions as equals without betraying its very identity. This does not mean, however, that anyone outside the visible fellowship of the Roman Catholic Church cannot be saved; nor does it preclude the presence of "vestiges of the church" in the other Christian bodies. At the second Vatican Council the Roman Catholic Church strongly affirmed its ties with its "separated brethren" both in Eastern Orthodoxy and in the several Protestant churches. As the true church of Christ on earth, the Roman Catholic Church also believes itself responsible for the proclamation of the will of God to organized society and to the state. This role brought the church into conflict with the state throughout church history. Yet the political activities of individual churchmen must not be confused with the fundamental obligation of the church, as the "light of the world" to which the revelation of God has been entrusted, to address the meaning of that revelation and of the moral law to the nations, and to work for a social and political order in which both revelation and the moral law can function.

Both in democratic and in totalitarian societies, whether Fascist or Communist, during the 20th century, the relation of the Roman Catholic Church to the state continued to engage the attention of political leaders and of prelates and theologians.

*Doctrine.* The understanding that Roman Catholicism has of itself, its interpretation of the proper relation between the church and the state, and its attitude toward other Christian traditions are all based upon Roman Catholic doctrine. In great measure this doctrine is identical with that confessed by orthodox Christians of every label and consists of the Bible, the dogmatic heritage of the ancient church as laid down in the historic creeds and in the decrees of the ecumenical councils, and the theological work of the great doctors of the faith in the East and West. If, therefore, the presentation of the other Christian traditions in this article compares them with Roman Catholicism, this comparison has a descriptive rather than a normative function; for, to a considerable degree, Protestantism and Eastern Orthodoxy have often defined themselves in relation to Roman Catholicism. In addition, most Christians past and present do have a shared body of beliefs about God, Christ, and the way of salvation.

Roman Catholic doctrine is more than this shared body of beliefs, as is the doctrine of each of the Christian groups. It is necessary here to mention only the three distinctive doctrines that have achieved definitive formulation during the 19th and 20th centuries: the infallibility of the pope, the immaculate conception, and bodily assumption of the Virgin Mary. On most other major issues of Christian doctrine, Roman Catholicism and Eastern Orthodoxy are largely in agreement, while Protestantism differs from both Eastern Orthodoxy and Roman Catholicism on several issues. For example, Roman Catholic theology defines and numbers the sacraments differently from Orthodox theology; but, over against Protestantism, Roman Catholic doctrine insists, as does Eastern Orthodoxy, upon the centrality of the seven sacraments—baptism, confirmation, Eucharist, extreme unction, penance, matrimony, and holy orders—as channels of divine grace.

Three distinctive doctrines of Roman Catholicism

*Liturgy.* The Roman Catholic doctrine of the sacraments is a summary, in liturgical form, of that which is affirmed by Roman Catholic liturgy. The church is not primarily an organization, nor is it a school of doctrine. It is the place where God approaches humanity through grace and where humanity approaches God through worship. Hence the focus of Roman Catholic piety is the Eucharist, which is both a sacrament and a sacrifice. Other forms of corporate worship and of private devotion radiate from this point of central focus. The obligations of church membership are also derived from the sacramental system, either as preparations for worthy participation in it or as expressions of the obedience sustained by it. Instruction in these obligations and in the implication of the faith for the moral and intellectual life is the responsibility of Roman Catholic educational institutions all over the world. The missions of the church and its institutions of mercy, like the schools, are largely in the hands of religious orders.

**The Eastern churches.** Separated from the West, the Orthodox churches of the East have developed their own way for half of Christian history. Each national church is autonomous. The "ecumenical patriarch" of Constantinople is not the Eastern pope but merely the first in honour among equals in jurisdiction. Eastern Orthodoxy interprets the primacy of Peter and therefore that of the pope similarly, denying the right of the pope to speak and act for the entire church by himself, without a church council and without his episcopal colleagues. Because of this polity Eastern Orthodoxy has identified itself more intimately with national cultures and with national regimes than has Roman Catholicism. Therefore the history of church–state relations in the East has been very different from the Western development, because the church in the East has sometimes tended toward the extreme of becoming a mere instrument of national policy while the church in the West has sometimes tended toward the extreme of attempting to dominate the state. The history of ecumenical relations between Eastern Orthodoxy and Protestantism during the 20th century was also different from the history of Protestant–Roman Catholic relations. While keeping alive their prayer for an eventual healing of the schism with the Latin Church, some of the Orthodox churches have established communion with Anglicanism and with the Old Catholic Church and have participated in the conferences and organizations of the World Council of Churches.

Doctrinal authority

Doctrinal authority for Eastern Orthodoxy resides in the Scriptures, the ancient creeds, the decrees of the first seven ecumenical councils, and the tradition of the church. In addition to the issues mentioned in the discussion of Roman Catholicism above, the chief dogmatic difference between Roman Catholic and Eastern Orthodox thought is on the question of the procession of the Holy Spirit from the Father and from the Son, or the *Filioque.*

But "orthodoxy," in the Eastern use of the term, means primarily not a species of doctrine but a species of worship. The Feast of Orthodoxy on the first Sunday of Lent celebrates the end of the iconoclastic controversies and the restoration to the churches of the icons, which are basic to Orthodox piety. In Orthodox churches (as well as in those Eastern churches that have reestablished communion with Rome), the most obvious points of divergence from general Western practice are the Byzantine liturgy, the right of the clergy to marry before ordination, though bishops may not be married, and the administration to the laity of both species in the Eucharist at the same time by the method of intinction.

The rediscovery of Eastern Orthodox liturgy and piety by Western Christians is an interesting by-product of the ecumenical contacts of the 19th century and of the Russian Revolution. Russian Orthodox scholars and theologians emigrated to the West, especially to France and the United States, where they became active participants in the dialogue among the separated churches.

**Protestantism.** Formulating a definition of Protestantism that would include all its varieties has long been the despair of Protestant historians and theologians, for there is greater diversity within Protestantism than there is between some forms of Protestantism and some non-Protestant Christianity. For example, a high-church An-

glican or Lutheran has more in common with an Orthodox theologian than with a Baptist theologian. Amid this diversity, however, it is possible to define Protestantism formally as non-Roman Western Christianity and to divide most of Protestantism into four major confessions or confessional families—Lutheran, Anglican, Reformed, and Free Church.

Four major Protestant confessions

*Lutheranism.* The largest of these non-Roman Catholic denominations in the West is the Lutheran Church. The Lutheran churches in Germany, in Scandinavian countries, and in the Americas are distinct from one another in polity, but almost all of them are related through various national and international councils, of which the Lutheran World Federation is the most comprehensive. Doctrinally, Lutheranism sets forth its distinctive position in the Book of Concord, especially in the Augsburg Confession. A long tradition of theological scholarship has been responsible for the development of this position into many and varied doctrinal systems. Martin Luther moved conservatively in this reformation of the Roman Catholic liturgy, and the Lutheran Church, though it has altered many of his liturgical forms, has remained a liturgically traditional church. Most of the Lutheran churches of the world have participated in the ecumenical movement and are members of the World Council of Churches, but Lutheranism has not moved very often across its denominational boundaries to establish full communion with other bodies. The prominence of Lutheran societies in the history of missions during the 18th and 19th centuries, after the relative inactivity following the Reformation, gave an international character to the Lutheran Church; so did the development of strong Lutheran churches in North America, where the traditionally German and Scandinavian membership of the church was gradually replaced by a more cosmopolitan constituency.

*Anglicanism.* The Anglican Communion is not only the established Church of England but also the Christian denomination of many believers throughout the world. Like Lutheranism, Anglicanism has striven to retain whatever it could of the Roman Catholic tradition of liturgy and piety, but after the middle of the 19th century the Oxford Movement in Anglicanism went much further in the restoration of ancient liturgical usage and doctrinal belief. Although the Catholic revival also served to rehabilitate the authority of tradition in Anglican theology generally, great variety continued to characterize the theologians of the Anglican Communion. Anglicanism is set off from most other non-Roman churches in the West by its retention of and its insistence upon the apostolic succession of ordaining bishops. The Anglican claim to this apostolic succession, despite its repudiation by Pope Leo XIII in 1896, has largely determined the role of the Church of England in the discussions among the churches. Anglicanism has often taken the lead in inaugurating such discussions, but in such statements as the Lambeth Quadrilateral it has demanded the presence of the historic episcopate as a prerequisite to the establishment of full communion. During the 19th and 20th centuries many leaders of Anglican thought were engaged in finding new avenues of communication with industrial society and with the modern intellectual. The strength of Anglicanism in the New World and in the younger churches of Asia and Africa confronted this communion with the problem of deciding its relation to new forms of Christian life in these new cultures. As its centuries-old reliance upon the establishment in England was compelled to retrench, Anglicanism discovered new ways of exerting its influence and of expressing its message.

*Presbyterian and Reformed Churches.* Protestant bodies that owe their origins to the reformatory work of John Calvin and his associates in various parts of Europe are often termed Reformed, particularly in Germany, France, and Switzerland. In Britain and in the United States they have usually taken their name from their distinctive polity and have been called Presbyterian. They are distinguished from both Lutheranism and Anglicanism by the thoroughness of their separation from Roman Catholic patterns of liturgy, piety, and even doctrine. Reformed theology has tended to emphasize the sole authority of the Bible with more rigour than has characterized the practice of Anglican or Lutheran thought, and it has looked with deeper suspicion upon the symbolic and sacramental traditions of the Catholic centuries. Perhaps because of its stress upon biblical authority, Reformed Protestantism has sometimes tended to produce a separation of churches along the lines of divergent doctrine or polity, by contrast with the inclusive or even latitudinarian churchmanship of the more traditionalistic Protestant communions. This understanding of the authority of the Bible has also led Reformed Protestantism to its characteristic interpretation of the relation between church and state, sometimes labeled theocratic, according to which those charged with the proclàmation of the revealed will of God in the Scriptures (*i.e.,* the ministers) are to address this will also to civil magistrates; Puritanism in England and America gave classic expression to this view. As the church is "reformed according to the Word of God," so the lives of the individuals in the church are to conform to the Word of God; hence the Reformed tradition has assigned great prominence to the cultivation of moral uprightness among its members. During the 20th century most of the Reformed churches of the world took an active part in the ecumenical movement.

Emphasis on the Bible as sole authority

*Free churches.* In the 19th century the term Free churches was applied in Great Britain to those Protestant bodies that did not conform to the establishment, such as Congregationalists, Methodists, and Baptists (and Presbyterians in England); but since that time it has come into usage among the counterparts to these churches in the United States, where each of them has grown larger than its British parent body. Just as the Reformed denominations go beyond both Anglicanism and Lutheranism in their independence of Roman Catholic traditions and usages, so the Free churches have tended to reject some of the Roman Catholic remnants also in classical Presbyterian worship and theology. Baptists and Congregationalists see the local congregation of gathered believers as the most nearly adequate visible representation of Christ's people on earth. The Baptist requirement of free personal decision as a prerequisite of membership in the congregation leads to the restriction of baptism to believers (*i.e.,* those who have made and confessed such a decision of faith) and therefore to the repudiation of infant baptism; this in turn leads to the restriction of communion at the Eucharist to those who have been properly baptized. In Methodism the Free-church emphasis upon the place of religious experience and upon personal commitment leads to a deep concern for moral perfection in the individual and for moral purity in the community. The Disciples of Christ, a Free church that originated in the United States, makes the New Testament the sole authority of doctrine and practice in the church, requiring no creedal subscription at all; a distinctive feature of their worship is their weekly celebration of Communion. Emphasizing as they do the need for the continuing reformation of the church, the Free churches have, in most (though not all) cases, entered into the activities of interchurch cooperation and have provided leadership and support for the ecumenical movement. This cooperation—as well as the course of their own historical development from spontaneous movements to ecclesiastical institutions possessing many of the features that the founders of the Free churches had originally found objectionable in the establishment—has made the question of their future role in Christendom a central concern of Free churches on both sides of the Atlantic.

*Other churches and movements.* In addition to these major divisions of Protestantism, there are other churches and movements not so readily classifiable; some of them are quite small, but others number millions of members. These churches and movements would include, for example, the Society of Friends (or Quakers), known both for their cultivation of the "Inward Light" and for their pacifism; the Unitarian Universalist body, which does not consistently identify itself as Christian; Christian Science; Unity and other theosophic movements, which blend elements from the Christian tradition with practices and teachings from other religions; Pentecostal churches and churches of divine healing, which profess to return to

primitive Christianity; and many independent churches and groups, most of them characterized by a free liturgy and a fundamentalist theology. Separately and together, these groups illustrate how persistent has been the tendency of Christianity since its beginnings to proliferate parties, sects, heresies, and movements. They illustrate also how elusive is the precise demarcation of Christendom, even for those observers whose definition of normative Christianity is quite exact. (J.J.Pe.)

## Christian doctrine

### THE MEANING OF DOGMA

Christian doctrine has often taken a specific form called "dogma." Although the term originally referred to "that which is regarded as good," the early Christians employed it to mean a teaching that came from divine revelation and was authoritatively defined as true by the church. The collection of such teachings is also referred to as dogma. Taken together, these teachings were seen to be vital for the salvation of Christians. To deviate from any of the dogmas was considered dangerous to salvation and to the life of the church.

Whereas in some periods, including early Christian times and whenever Christianity was dominant in the culture, those in authority and most of the faithful thought of dogma in positive terms, it has often suffered criticism in modern times. The concept seems to militate against the freedom of inquiry by which moderns seek truth, and it can be easily transformed into an instrument of state in officially Christian societies, where deviations can be punished as treason. In addition, to some Christians, the great German scholar Adolf von Harnack among them, dogma appeared to be a corruption of the essentially simple faith of Jesus and the early Christians. It was seen as an intrusion of Greek philosophy, an "acute Hellenization" that obscured the Christian truth it had set out to guard.

The Hellenization of the Christian message

Without question, some Hellenization did take place, and dogma was often misused. Yet it is difficult to picture a rapidly growing body such as the early church existing permanently without efforts to provide authoritative definitions of the ideas its leadership regarded as true and for which they would be responsible. Therefore, any attempt to discern the meaning of Christian faith through the ages necessarily involves an understanding of the reasons for the existence of dogma and a comprehension of what the main dogmatic teachings in the various times and places of the church's life have been.

The early Christian definitions of dogma drew on Greek thought in general and, in some cases, on Neoplatonic philosophy in particular. Out of this borrowing of categories not derived from the Bible came basic, indeed crucial, definitions concerning both the Trinity and Christology. Historians, particularly in the modern West, have appraised and criticized these applications of Greek thought.

The Eastern Orthodox Church, however, views the formation of dogma not as a purely human process (which would have to be judged a falsification of divine truths) but as a divine–human process in which the Holy Spirit, proceeding from God, and the human spirit, proceeding from history, participate. According to this view, the origin of revelation in God means that the truths of dogma are divine, eternal, and immutable.

Concepts of dogma in the various churches

Dogma has varying positions in the different churches. Whereas in Protestant churches doctrines and creeds generally are associated primarily with theology and preaching, in the Eastern Orthodox Church dogma is directly related to the liturgical life of the church. The confessions of faith of the Orthodox Church are not to be understood as abstract formulations of a pure doctrine but as hymns of worship incorporated appropriately in the liturgy.

In contrast to the rather definitive dogmatic development in the Roman Catholic Church and Reformation Christianity, a much greater freedom in the interpretation of dogma is guaranteed in the Eastern Church. Even the formulation of a dogma by an ecumenical council in Eastern Orthodoxy does not have a binding legal character until it is received within the total ecumenical church consciousness.

**Orientation of the Western churches.** From their inception, the Western churches have viewed the fundamental relationship between God and humanity primarily in judicial terms. Characteristically, the Apostle Paul in his letter to the Romans depicted the experience of salvation as justification. To be justified means to receive the verdict "just" or "righteous," whether from a judge—the law court was the source of the Pauline theological image—or from God, as witnessed to in the Christian scheme of salvation. The Roman Church was strongly Jewish–Christian in character and combined this character without difficulties with the basic orientation of the Roman view of religion—*i.e.*, the relationship between God and humanity was primarily a judicial one. The legal character of Roman religion was expressed in the fact that the efficacy of the state cult ceremonies was dependent upon the strictest observation of a wide variety of regulations. Later developments of Roman Catholic Christianity depended largely upon the basis of this legal thinking. In Rome the specifically Western sacrament of penance developed in the context of legal terminology and was dominated by the idea of justification.

The relationship between God and humanity as judicial

In the judicial foundation of the sacrament of penance (in which an offender might regain a right relationship with the church through the performance of certain works), later possibilities of corruption are built in (*e.g.*, the indulgence, or remission of temporal punishment upon the granting of absolution by a priest and, perhaps, the payment of a fee or performance of certain works). The indulgence resulted from a fusion of Roman and German legal thought.

Just as judicial notions dominated divine–human personal relationships, so they helped shape questions of authority. Thus the Western notions of both church and priesthood acquired a legal cast. The church understands itself as a spiritual–judicial institution, founded by Christ but perpetuated by structures that often resemble those in human government. The priest is the legitimate bearer of this legal order. In the sphere of this kind of legal thinking the papacy and the doctrine of papal primacy developed. The idea of a jurisdictional primacy played a prominent role in the formation of the doctrine of the papacy. Kingly authority passed over to the priestly, the emperor's crown to the episcopal tiara. At the high point of this development, Pope Boniface VIII in 1302 proclaimed himself the highest ruler of the world, to whom Christ has committed both swords, the spiritual and the temporal. This judicial idea is also reflected in the individual priest's consciousness of his office. Ordination by the bishop confers upon the priest a legal authority to administer the sacraments and to exercise the power of the keys. In consciousness of this legal process the priest absolves the sinner from his sins in confession with the words: "I absolve you."

On the basis of this legal consciousness the Western church also developed its own canon law. Canon law in the West penetrated, indeed dominated, the societal sphere of life much more strongly than it did in the Orthodox Church (see below *Canon law*).

Influence of judicial thinking in Western theology and monasticism

Judicial thinking was similarly significant in the theology of the West. Whereas the East never assigned a decisive significance to the justification doctrine of the Apostle Paul, in the West the theologian Tertullian introduced a series of fundamental juridical concepts into theology (*e.g.*, persona—a legal person). For Augustine, the doctrine of justification was the foundation of his view of the relationship of humans to God as well as of his view of sin, guilt, and grace. For Anselm, archbishop of Canterbury from 1093 to 1109, the existing, valid judicial relationship between God and humans had to be the basic presupposition of all theological thinking. Anselm believed that he could cogently derive—even for unbelievers—the truth of Christian faith and the necessity of the incarnation of God from the idea of satisfaction (*i.e.*, that one could make satisfaction for a crime against another).

Western monasticism also received its special imprint from judicial thought. Sanctification is believed to be accomplished by practicing good works and, particularly, surplus works or "works of supererogation"—*i.e.*, those that the saint performs over and above those necessary

for the satisfaction of his own sins. Alexander of Hales (*c.* 1170/85–1245), an English theologian and philosopher, advanced the doctrine that out of the works of satisfaction of Christ, the saints, and the martyrs, the church has collected a "treasury of good works," which the pope properly has at his disposal.

This judicial thinking is even extended to the eschatological expectation and the view of the last things, for many of the biblical notions about God acting out of love in the final dealings with humans tend to be challenged or obscured by overpowering pictures of divine righteousness. At the end of the world stands the radical separation of humanity into the redeemed, who enter eternal blessedness, and the damned, who are delivered over to eternal punishment. According to the Roman Catholic view, sinners improve their prospects by acceptance of an interim state, in which they can ameliorate their positions in respect to God before the Last Judgment by making amends for their sins. Through indulgences, masses for the dead, and other acts, the church expanded its spiritual–judicial authority even to this realm of the departed souls of purgatory (the state of existence after death in which temporal punishment is meted out). Although the doctrine of purgatory remains among the official teachings of contemporary Roman Catholicism, it has been downplayed in much popular expression and belief. The pictures of God acting beneficently toward sinners have progressively challenged the medieval and early modern accent on God acting punitively.

**Orientation of the Eastern churches.** In the New Testament, ideas later regarded as mystical are associated with the writings and experiences of Paul and in the Gospel and Letters associated with the name John. The Orthodox churches have encouraged a piety that elevates these mystical themes and makes them prominent. In these Eastern settings justification is less frequently considered. Instead, the idea that humans are to partake in the nature of the Godhead—to be holy, reborn, newly created, resurrected, and transfigured—comes to the fore. Indeed, not only humanity but also the whole of creation partakes of this potentially transforming character. Orthodox mystical piety thus has a bearing on the entire cosmos. The central concept is not the righteousness but the love of God. Thus, a different overall development of religious perception took place in the East, especially noticeable in the conception and development of the sacrament of penance. Since, in the East, the idea of educating the Christian to a life of sanctification is decisive, the juridical conception of penance never gained wide acceptance, and neither the doctrine nor the practice of indulgences was developed in the Eastern Church.

Motifs of the Eastern churches

Also, the Eastern Church did not claim power to intervene in the realm of the dead, to loosen or to bind. Eastern Christianity is familiar only with intercession for the dead, because the bond of the faithful who are joined together into the body of Christ is not extinguished even with death.

The legal idea is, however, present in the Eastern Church's view of the ecclesiastical office, especially in its conception of the episcopal office and apostolic succession. But the Eastern view is embedded in the conception of the church as the mystical body of Christ and of the Holy Spirit as the stream of life of the church. The bishops of the Eastern Orthodox Church have always remained primarily bishops (and not also temporal rulers) of their church and have always preserved the spiritual character of their office—even at the times when the churches were under Islāmic rule and the bishops were assigned the function of ethnarchs (official quasi-political representatives of the Christian portions of the population). The conception that the Orthodox priest has of the essence of his priesthood is, then, not determined through a judicial idea, and in the Orthodox sacrament of penance the formula of absolution has the form not of a declaration but of a prayer for divine forgiveness.

Judicial features are also absent from the Eastern Church's conception of human sanctification and, thereby, of the task of monasticism. In Eastern monasticism, a doctrine of good works or of the treasury of the church was never able to arise. Saints were venerated as spiritually gifted personalities who realized in this earthly life the angelic life of the heavenly church.

In Orthodox theology the schema of justification has scarcely played a role. Its chief motif is the incarnation of God and the idea that humans partake of aspects of deity. Though the term must be used cautiously, to avoid idolatry of the created human, it is not entirely inappropriate to speak of the "deification" of the human. Thus, the emphasis of Christian proclamation lay upon rebirth, the new creation of the human being, the process of transformation into a new creature, resurrection with Christ, and the ascension of humans to God, along with other changes properly called transfigurations. Only the penetration of Reformation ideas in the 16th and 17th centuries compelled Orthodox theologians to take a position on the doctrine of justification.

The incarnation of God and the deification of humans

In Roman thought sin is a violation of the legal relationship fixed by God between himself and man. Sin for the Eastern Church, however, is viewed as a narrowing of possibilities, or a contraction of essence, a sickness or infection of the original being of the image of God. Accordingly, redemption is not primarily the restoration of a judicial relationship disturbed by sin but rather a renewal of being, transfiguration of being, completion of being, and deification.

Thus, the idea of love is dominant in Eastern piety. Characteristic of this is a catechetical sermon by John Chrysostom, patriarch of Constantinople, about Christ's parable of the workers in the vineyard as told in Matthew, chapter 20. Still read at Easter in the 20th century from all pulpits of the Orthodox Church, the sermon is a song of triumph about the victory of the boundless love of God:

> You who have fasted and you who have not fasted, rejoice today! The table is laden, enjoy it everyone! The calf is fattened, may no one leave hungry! Everyone partake of the banquet of faith! Everyone partake of the riches of goodness! May no one complain of poverty, for the common Kingdom has appeared.

The conception of the Last Judgment in the Eastern Church did not develop according to the strictly juridical interpretation that was customary in the West. On the contrary, trust in grace and in the "love of man" by the divine Logos, as well as supplication for divine compassion, are dominant. Hence, the Orthodox Church understood only with difficulty the theological concern of the Western Reformation. It comprehended the Western Reformation only to the extent that it concurred with the latter in its rejection of certain Roman Catholic doctrines and practices—*e.g.*, the doctrine of papal primacy and the demand of priestly celibacy. The central argument about justification was reflected upon by only a few Orthodox theologians educated in the West, such as Cyril Lucaris, and in general the Eastern Church has withstood efforts to graft justification onto Orthodoxy.

### GOD THE FATHER

On the basis of their religious experiences, the mystics of Christianity of all eras have concurred in the belief that one can make no assertions about God, because God is beyond all concepts and images. Inasmuch as human beings are gifted with reason, however, the religious experience of transcendence demands historical clarification. Thus, in Christian theology two tendencies stand in constant tension with each other. On the one hand, there is the tendency to systematize the idea of God as far as possible. On the other, there is the tendency to eliminate the accumulated collection of current conceptions of God and to return to the understanding of the utter transcendence of God. Theologians, by and large, have had to acknowledge the limits of human reason and language to address the "character" of God, who is beyond normal human experience but who impinges on it. But because of the divine–human contact, it became necessary and possible for them to make some assertions about the experience, the disclosure, and the character of God.

Tendencies within Christian theology to make assertions about God

All great epochs of the history of Christianity are defined by new forms of the experience of God and of Christ. Rudolf Otto, a 20th-century German theologian, attempted to describe to some extent the basic ways of

experiencing the transcendence of the "holy." He called these the experience of the "numinous" (the spiritual dimension), the utterly ineffable, the holy, and the overwhelming. The "holy" is manifested in a double form: as the *mysterium tremendum* ("mystery that repels"), in which the dreadful, fearful, and overwhelming aspect of the numinous appears, and as the *mysterium fascinosum* ("mystery that attracts"), by which humans are irresistibly drawn to the glory, beauty, adorable quality, and the blessing, redeeming, and salvation-bringing power of transcendence. All of these features are present in the Christian concepts of God as explicated in the ever new experiences of the charismatic leaders.

**Characteristic features of the Christian concept of God.** Within the Christian perception and experience of God, definite characteristic features stand out: (1) the personality of God, (2) God as the Creator, (3) God as the Lord of history, and (4) God as Judge. (1) God, as person, is the "I am who I am" designated in Exodus 3:14. The personal consciousness of human beings awakens in the encounter with God understood as a person: "The Lord used to speak to Moses face to face, as a man speaks to his friend" (Exodus 33:11). (2) God is also viewed as the Creator of heaven and Earth. The believer thus maintains, on the one hand, acknowledgement of divine omnipotence as the creative power of God, which also operates in the preservation of the world, and, on the other hand, trusts in the world, which—despite all its contradictions—is understood as one world created by God according to definite laws and principles and according to an inner plan. The decisive aspect of creation, however, is that God fashioned humans according to the divine image and made the creation subject to them. This special position of humans in the creation, which makes them coworkers of God in the preservation and consummation of the creation, brings a decisively new characteristic into the understanding of God. (3) This new characteristic is God as the Lord of history, which is the main feature of the Old Testament understanding of God: God selects a special people and contracts a special covenant with them. Through the Law the divine agent binds this "people of God" in a special way. God sets before them a definite goal of salvation—the establishment of a divine dominion—and through the prophets admonishes the people by proclamations of salvation and calamity whenever they are unfaithful to the covenant and promise. (4) This God of history also is the God of judgment. The genuinely Israelite belief that the disclosure of God comes through the history of divinely-led people leads, with an inner logic, to the proclamation of God as the Lord of world history and as the Judge of the world.

The special position of humans in creation and history

**The specific concept of God as Father.** What is decisively new in the Christian, New Testament faith in God lies in the fact that this faith is so closely bound up with the person, teaching, and work of Jesus Christ that it is difficult to draw boundaries between theology (doctrines of God) and Christology (doctrines of Christ). Jesus himself embraced the God of the Hebrew patriarchs (Abraham, Isaac, and Jacob), but he also understood himself as the fulfiller of the promise of the Messiah–Son of man, who is the bringer of the Kingdom of God. The religious experience that forms the basis of the messianic self-understanding of Jesus is the recognition that the Messiah–Son of man is the Son of God.

The special relationship of Jesus to God is expressed through his designation of God as Father. In prayers Jesus used the Aramaic word *abba* (father) for God, which is otherwise unusual in religious discourse in Judaism; it was usually employed by children for their earthly father, similar to "daddy" in English. This father–son relationship became a prototype for the relationship of Christians to God. Appeal to the sonship of God played a crucial role in the development of Jesus' messianic self-understanding. According to the account of Jesus' baptism, Jesus understood his sonship when a voice from heaven said: "This is my beloved Son, with whom I am well pleased." In the Gospel According to John, this sonship constitutes the basis for the self-consciousness of Jesus: "I and the Father are one" (John 10:30).

The special relationship of Jesus to God

**The belief in the oneness of the Father and the Son.** Faith in the Son also brought about a oneness with the Father. The Son became the mediator of the glory of the Father to those who believe in him. In Jesus' high priestly prayer (in John, chapter 17) he says: "The glory which thou hast given me I have given to them, that they may be one even as we are one, I in them and thou in me, that they may become perfectly one." In the Lord's Prayer Jesus taught his disciples to address God as "our Father."

The Father–God of Jesus after Jesus' death and Resurrection becomes—for his disciples—the God and Father of our Lord Jesus Christ (*e.g.,* 2 Corinthians 1:3), who revealed his love through the sacrifice of his Son who was sent into the world. Faithful Christians can thus become the children of God, as noted in Revelation 21:7: "I will be his God and he shall be my son." For Christians, therefore, faith in God is not a doctrine to be detached from the person of Jesus Christ.

Medieval theologians often spoke of a "Beatific Vision," a blessed vision of God. They did so on the basis of their own mystical experience that constituted the fulfillment of salvation in the Kingdom of God, which the Son will deliver to the Father.

In the history of Christian mysticism, this visionary experience of the transpersonal "Godhead" behind the personal "God" (as in the works of the medieval mystic Meister Eckhart)—also called an experience of the "trans-deity," the "divine ground," "groundlessness," the "abyss," and the divine "nothingness"—constantly breaks through and is renewed. Occasionally, this experience of transpersonal divine transcendence has directed itself against the development of a piety that has banalized the personal idea of God so much so that the glory and holiness of God has been trivialized. The attempt of the 20th-century theologian Paul Tillich to reduce the Christian idea of God to the impersonal concept of "the Ground of Being," or "Being Itself," pointed toward an understanding of the prepersonal depths of the transcendence of Godhood.

Transpersonal mystical and mono-Christological tendencies

Nevertheless, in the Christian understanding of Christ as being one with the Father, there is a constant possibility that faith in God will be absorbed in a "monochristism"—*i.e.,* that the figure of the Son in the life of faith will overshadow the figure of the Father and thus cause it to disappear and that the figure of the Creator and Sustainer of the world will recede behind the figure of the Redeemer. The history of Christian piety and of Christian theology has constantly moved in this field of tension. Thus, the primacy of Christology and of the doctrine of justification in Reformation theology led to a depreciation of the creation doctrine and a Christian cosmology. This depreciation accelerated the estrangement between theology and the sciences during the period of the Enlightenment. This was subsequently distorted into a form of materialism. On the other hand, some 20th-century dialectical theologians, among them Karl Barth, in opposing materialism and humanism sometimes evoked a monochristic character that strongly accented the centrality of Christ at the expense of some cultural ties.

**The revelatory character of God.** The God of the Bible is the God who presses toward revelation. The creation of the world is viewed as an expression of God's will toward self-revelation, for even the pagans "knew God." In Paul's so-called Areopagus speech in Athens, he said of God: "Yet he is not far from each one of us, for 'in him we live and move and have our being,' " in allusion to the words of the pagan writer Aratus: "For we are indeed his offspring" (Acts 17:27–28). This was the beginning of a knowledge of God that has manifested itself under the catchphrase of the "natural revelation" of God or God's revelation in the "book of nature." It has survived as one strand of theory throughout much of Christian history.

God's bent toward self-revelation

The self-revelation of God presupposes, however, a basic biblical understanding of the existing relationship between God and human beings. It cannot be separated from the view that God created humans according to the divine image and that in Jesus Christ, who "reflects the glory of God and bears the very stamp of his nature" (Hebrews 1:3), the heavenly man has appeared among humans as the "last Adam." The inner connection between the "nat-

ural" and the biblical revelation takes place through the view of Christ as the divine Logos become human.

The tension between monotheism and the Logos doctrine

Hellenistic thinkers had already been attracted by the emphasis in later Judaism on monotheism and transcendence. This tendency was sketched out earlier in Plato and later Stoicism, but it came to its mature development in Neoplatonism in the 3rd century AD. In the 1st century Philo of Alexandria had interpreted the Old Testament concept of God in terms of the Logos idea of Hellenistic philosophy, but this Hellenization led to a characteristic tension that was to dominate the entire further history of Christian piety, as well as the Western history of ideas. The Greeks traced the idea of God to a "first cause" that stood behind all other causes and effects. Theologians under their influence used this understanding to contribute to a doctrine of God as "first cause" in Christian theology.

**God as Creator, Sustainer, and Judge.** The biblical understanding of God, however, was based upon the idea of the freedom of the Creator, Sustainer, and Judge vis-à-vis the world. This idea also included the concept that God could suspend the natural order or break the causal chain through miracles. This led to two specific problems that theology, inspired by Greek philosophy, set for itself: (1) the attempt to prove the existence of God, and (2) the attempt to justify God in view of both the apparent shortcomings of the creation and the existence of evil in history (*i.e.*, the problem of theodicy). Both attempts have occupied the intellectual efforts of Western theology and have inspired the highest of intellectual achievements. These attempts, however, often presumed that human reason could capture and define the transcendent. Even by the speculative theologians' definition such an effort was inherently impossible to conclude. Although such theologians creatively addressed the issue, it was often simple Christian piety that served to guard the notion of transcendence, while concentrating on the historical revelation of God in the more accessible instrument of God's self-disclosure in Jesus Christ.

Efforts to explain the ways of God to humans, particularly in respect to the problem of the existence of evil, are called theodicy. This form of justification of God has addressed profound human impulses and has relied upon strenuous exercises of human reason, but it has also led to no finally satisfying conclusions. The problem, which was already posed by Augustine and treated in detail by Thomas Aquinas, became of pressing importance in the European Thirty Years' War (1618–48) and its aftermath. At that time Gottfried Wilhelm Leibniz, who did more than anyone to develop the concept of theodicy, endeavoured to defend the Christian notion of God against the obvious atheistic consequences that were evoked by the critical thinkers of his time. This was because of the behaviour of the Christian churches, which were engaged in a war of mutual extermination. The result of such theological efforts, however, was either to declare God himself as the originator of evil, to excuse evil as a consequence of divine "permission," or instead—as with Hegel—to understand world history as the justification of God ("the true theodicy, the justification of God in history"). These answers satisfied neither the Christian experience of faith nor thoughtful reflection. Literature is full of examples in which writers influenced by the Christian tradition react against such justifications. One example can be found in the Russian novelist Fyodor Dostoyevsky's treatment of the suffering of children in *The Brothers Karamazov*.

The German philosopher Immanuel Kant set the terms for much modern reflection on God's existence when he challenged the grounds of most previous efforts to prove it. Kant contended that it was finally impossible for the human intellect to achieve insights into the realm of the transcendent. Even as he was arguing this, modern science was shifting from grounds that presumed the nature of God and God's universe to autonomous views of nature that were grounded only in experiment, skepticism, and research. During the 19th century, philosophers in Kantian and scientific traditions despaired of the attempt to prove the existence of God.

Modern attempts to abandon the idea of God

During the same period some Western intellectuals turned against the very idea of God. One strand of Hegelian thinkers, typified by the German philosopher Ludwig Feuerbach, attempted to unmask the idea of religion as illusion. To Feuerbach, faith was an ideology designed to help humans delude themselves. The idea of dialectical materialism, in which the concept of "spirit" was dropped by thinkers such as Karl Marx, developed in this tradition. It also characterized religion as "bad faith" or "the opiate of the people," designed to seduce them from efforts to build a good society through the hope of rewards in a life to come.

At the same time, at first chiefly in Britain, scientific thinkers in the tradition of Charles Darwin hypothesized that evolutionary processes denied all biblical concepts of divine creation. Some dialectical materialists incorporated Darwinian theories in a frontal attack on the Christian worldview. Some Christians contended that this was a perversion of evolution, since certain Christian teachings on divine creation, such as *creatio continua* ("continuing creation"), were both biblical and compatible with evolutionary theory. At the turn of the 20th century, some thinkers in both Britain and the United States optimistically reworked their doctrine of God in congruence with evolutionary thought.

**Modern views of God.** If 18th- and 19th-century rationalism and scientific attacks on the idea of God were often called "the first Enlightenment" or "the first illumination," in the 20th century a set of trends appeared that represented, to a broader public, a "second illumination." This included a rescue of the idea of God, even if it was not always compatible with previous Christian interpretations. Notable scientists of the 20th century, such as Albert Einstein, Max Planck, Max Born, and others, have on occasion, and against the testimony of the majority of their colleagues, allowed for an idea of God or religion in their concepts of life, the universe, and human beings.

Resurgence of a recognition of God

Corresponding to recognition of the idea of God by some leading scientific thinkers, there has been a new surge of experience of God noticeable in the different revival movements in the churches of Asia and Africa, as well as in America and Eurasia—in the midst of people either de-Christianized or attached to a purely conventional "cultural Christianity."

When the German philosopher Friedrich Nietzsche prophesied what he called "the death of God," many Christian thinkers agreed that a certain set of culturally conditioned and dogmatic concepts of God were inaccessible, implausible, and dying out. Some of these apologists argued that such a "death of God" was salutary, because it made room for a "God beyond the gods" of argument, or a "greater God." The French Jesuit thinker Pierre Teilhard de Chardin for a time attracted a large following as he set out to graft the theory of evolution onto "greater God" proclamations.

Certainly new understandings of physics, astronomy, and cosmology made it more difficult to defend pictures of God based upon notions of a small and young universe. The public grasped the change as it received images of a small Earth transmitted from space by Soviet and U.S. astronauts. Yet space exploration also inspired the imagination and led Christian thinkers who had positive views of science to use the occasion to enlarge the understanding of God and relate it to the Christian system.

**The view that God is not solitary.** The leaders of an 18th-century movement called Deism saw God as impersonal and unempathic—a principle of order and agent of responsibility not personal or addressable as the Christian God had been. Deism contributed to some intellectualizations of the idea of God, approaches that had sometimes appeared in the more sterile forms of medieval Scholasticism. God appeared to have been withdrawn from creation, which was pictured as a world machine; this God, at best, observed its running but never interfered.

According to the original Christian understanding of God of the early church, the Middle Ages, and the Reformers, God neither is solitary nor wishes to be alone. Instead, God is encircled with a boundless realm of angels, created in the divine image. They surround God in freely expressed love and devotion. They appear in a graduated, individuated hierarchy. These ranks of angels offer God

The role of angels

their praise, and they appear active in the universe as messengers and executors of the divine will. From the beginning God appears as the ruler and centre in this divinely fashioned realm, and the first created of this realm are the angels. The church of the angels is the upper church; the earthly church joins with them in the "cherubic hymn," the Trisagion ("Holy, Holy, Holy"), at the epiphany of the Lord and the angelic choirs surrounding him in the Eucharist. The earthly church is thus viewed as a participant—co-liturgist—in the angelic liturgy. Because the angels are created as free spiritual beings in accordance with the image of God, the first fall takes place in their midst—the first misuse of freedom was in the rebellion of the highest prince of the angels, Lucifer ("Light-bearer"), against God.

According to the view of the Fathers of the early church, teachers of the Middle Ages, and the Reformers, humans are only the second-created. The creation of human beings serves to refill the Kingdom of God with new spiritual creatures who are capable of offering to God the free love that the rebellious angels have refused to continue. In the realm of the first-created creatures, there already commences the problem of evil, which appears immediately in the freedom or misuse of freedom.

**Satan and the origin of evil.** In the Old and New Testaments, Satan (the devil) appears as the representative of evil. The philosophy and theology of the Enlightenment endeavoured to push the figure of the devil out of Christian consciousness as being a product of the mythological fantasy of the Middle Ages. It is precisely in this figure, however, that some aspects of the ways God deals with evil are especially evident. The devil first appears as an independent figure alongside God in the course of the Old Testament history of religion. In the Old Testament evil is still brought into a direct relationship with God; even evil, insofar as it has power and life, is effected by God: "I form light and create darkness, I make weal and create woe, I am the Lord, who do all these things" (Isaiah 45:7).

Satan gives expression to the demonic side of the divine wrath. In the Book of Job he appears as the partner of God, who on behalf of God puts the righteous one to the test. Only in postbiblical Judaism does the devil become the adversary of God, the prince of angels, who, created by God and placed at the head of the angelic hosts, entices some of the angels into revolt against God. In punishment for his rebellion he is cast from heaven together with his mutinous entourage, which were transformed into demons. As ruler over the fallen angels he henceforth continues the struggle against the Kingdom of God in three ways: he seeks to seduce man into sin; he tries to disrupt God's plan for salvation; and he appears before God as slanderer and accuser of the saints, so as to reduce the number of those chosen for the Kingdom of God.

By courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co. Ltd.




Satan leaves the presence of God to test God's faithful servant, Job. Engraving by William Blake, 1825, for an illustrated edition of *The Book of Job.*

The functions of Satan

Thus, Satan has a threefold function: he is a creature of God, who has his being and essence from God; he is the partner of God in the drama of the history of salvation; and he is the rival of God, who fights against God's plan of salvation. Through the influence of the dualistic thinking of Zoroastrian religion during the Babylonian Exile (586–538 BC) in Persia, Satan took on features of a countergod in late Judaism. In the writings of the Qumrān sects (who preserved the Dead Sea Scrolls), Belial, the "angel of darkness" and the "spirit of wickedness," appears as the adversary of the "prince of luminaries" and the "spirit of truth." The conclusion of the history of salvation is the eschatological battle of the prince of luminaries against Belial, which ends with judgment upon him, his angels, and people subject to him and ushers in the cessation of "worry, groaning, and wickedness" and the beginning of the rule of "truth."

In the New Testament the features of an anti-godly power are clearly prominent in the figures of the devil, Satan, Belial, and Beelzebub—the "enemy." He is the accuser, the evil one, the tempter, the old snake, the great dragon, the prince of this world, and the god of this world, who seeks to hinder the establishment of God's dominion through the life and suffering of Jesus Christ. Satan offers to give to Christ the riches of this world if Christ will acknowledge him as supreme lord. Thus, he is the real antagonist of the Messiah–Son of man, Christ, who is sent by God into the world to destroy the works of Satan.

He is lacking, however, the possibility of incarnation: he is left to rob others in order to procure for himself the appearance of personality and corporeality. As opposed to *philanthrōpia,* the love of man of Christ, who presents himself as an expiatory sacrifice for the sins of mankind out of love for it, Satan appears among early church teachers, such as Basil of Caesarea in the 4th century, as the *misanthrōpos,* the hater of humanity; vis-à-vis the bringer of heavenly beauty, he is the hater of beauty, the *misokalos.* With Gnosticism, dualistic features also penetrated the Christian sphere of intuitive vision. In the *Letter of Barnabas* (early 2nd century) Satan appeared as "the Black One"; according to the 2nd-century apologist Athenagoras he is "the one entrusted with the administration of matter and its forms of appearance," "the spirit hovering above matter." Under the influence of Gnosticism and Manichaeism (a syncretistic religion founded by Mani, a 3rd-century Persian prophet), there also followed—based on their dualistic aspects—the demonization of the entire realm of the sexual. This appears as the special temptational sphere of the devil; in sexual activity, the role of the instrument of diabolic enticement devolves upon woman. Manichaeistic and Gnostic tendencies remained as a permanent undercurrent in the church and determined, to a great extent, the understanding of sin and redemption. Satan remained the prototype of sin as the rebel who does not come to terms with fulfilling his godlikeness in love to his original image and Creator but instead desires equality with God and places love of self over love of God.

Gnostic and Manichaean elements in Christian concepts of angels and sin against God

Among the Fathers of the early church, the idea of Satan as the antagonist of Christ led to a mythical interpretation of the incarnation and disguise in the "form of a servant." Through this disguise the Son of God makes his heavenly origin unrecognizable to Satan. In some graphic, and almost comic, medieval depictions Christ appears as the "bait" cast before Satan, after which Satan grasps because he believes Christ to be an ordinary human being subject to his power. In the Middle Ages a further feature was added: the understanding of the devil as the "ape of God," who attempts to imitate God through spurious, malicious creations that he interpolates for, or opposes to, the divine creations.

In church history, the eras of the awakening of a new

Consciousness of sin against God and a renewed sense of evil

consciousness of sin are identical with those of a newly awakened sense for the presence of "evil"—as was the case with Augustine, Bernard of Clairvaux, Luther, Calvin, and Wesley. In the Christian historical consciousness the figure of Satan plays an important role, not least of all through the influence of the Revelation to John. The history of salvation is understood as the history of a continuous struggle between God and the demonic antagonist, who with constantly new means tries to thwart God's plan of salvation. The idea of the "stratagems of Satan," as developed by a 16th-century fortress engineer, Giacomo Aconcio, had its roots here: the history of the world is a constant attempt of Satan to disrupt the salvation events of God through ever new counter-events. This altercation constitutes the religious background of the drama of world history. Characteristic here is the impetus of acceleration already indicated in Revelation: blow and counterblow in the struggle taking place between God and Satan follow in ever shorter intervals; for the devil "knows that his time is short" (Revelation 12:12), and his power in heaven has already been laid low. On Earth the possibility of his efficacy is likewise limited by the return of the Lord. Hence, his attacks upon the elect of the Kingdom so increase in the last times that God is moved to curtail the days of the final affliction, for "if those days had not been shortened, no human being would be saved" (Matthew 24:22). Many of these features are retained in the philosophy of religion of German Idealism as well as in Russian philosophy of religion. According to the 20th-century Russian philosopher Nikolay Berdyayev, like the Germans Friedrich Schelling and Franz von Baader before him, the devil has no true personality and no genuine reality and, instead, is filled with an insatiable "hunger for reality," which he can attain by stealing reality from the people of whom he takes possession. Since the Enlightenment, Christian theologians who found the mythical pictures of Satan to be irrelevant, distorting, or confusing in Christian thought and experience have set out to demythologize this figure. Apologists such as the British literary figure C.S. Lewis and the Russian philosopher Vladimir Solovyov, however, have written cautionary words. They conceive that it would represent the devil's most cunning attempt at self-camouflage to be demythologized and that camouflage would be a certain new proof of his existence.

### GOD THE SON

Dogmatic teachings about the figure of Jesus Christ go back to the spontaneous faith experiences of the original church. The faithful of the early church experienced and recognized the incarnate and resurrected Son of God in the person of Jesus. The disciples' testimony served as confirmation for them that Jesus really is the exalted Lord and Son of God, who sits at the right hand of the Father and will return in glory to consummate the Kingdom.

**Different interpretations of the person of Jesus.** From the beginning of the church different interpretations of the person of Jesus have existed alongside one another. The Gospel According to Mark, for example, understands Jesus as the man upon whom the Holy Spirit descends at the baptism in the Jordan and who is declared the Son of God through the voice of God from the clouds. Two schools of thought developed—one associated with Antioch in Syria and the other with Alexandria in Egypt. Attempts at Christology that derive from the theological school of Antioch have followed one line of interpretation: they proceed from the humanity of Jesus and view his divinity in his consciousness of God, founded in the divine mission that was imposed upon him by God through the infusion of the Holy Spirit.

Another view was adopted by the catechetical school of Alexandrian theology. This view is expressed by the Gospel According to John, which regards the figure of Jesus Christ as the divine Logos become flesh. Here, the divinity of the person of Jesus is understood not as the endowment of the man Jesus with a divine power but rather as the result of the descent of the divine Logos—a preexistent heavenly being—into the world: the Logos taking on a human body of flesh so as to be realized in history. Thus it was that the struggle to understand the figures of Jesus Christ created a rivalry between the theologies of Antioch and Alexandria. Both schools had a wide sphere of influence, not only among the contemporary clergy but also in monasticism and among the laity. Characteristically, Nestorianism (a heresy founded in the 5th century), with its strong emphasis upon the human aspects of Jesus Christ, arose from the Antiochene school, whereas Monophysitism (a heresy founded in the 5th century), with its one-sided stress upon the divine nature of Christ, emerged from the Alexandrian school of theology.

Emphasis on the Logos

**The Christological controversies.** New intermediate solutions for resolving the Christological problem constantly were proposed between the two extreme positions of Antioch and Alexandria. As in the area of the doctrine of the Trinity, the general development of Christology has been characterized by a plurality of views and formulations. Also, the creeds of the major churches have by no means agreed with each other word for word. After Constantine, the great ecumenical synods occupied themselves essentially with the task of creating uniform formulations binding upon the entire imperial church.

Even the Christological formulas, however, do not claim to offer a rational, conceptual clarification; instead, they emphasize clearly three contentions in the mystery of the sonship of God. These are: first, that Jesus Christ, the Son of God, is completely God, that in reality "the whole fullness of deity dwells bodily" in him (Colossians 2:9); second, that he is completely human; and third, that these two "natures" do not exist beside one another in an unconnected way but, rather, are joined in him in a personal unity. Once again, the Neoplatonic metaphysics of substance offered the categories so as to settle conceptually these various theological concerns. Thus, the idea of the unity of essence (*homoousia*) of the divine Logos with God the Father assured the complete divinity of Jesus Christ, and the mystery of the person of Jesus Christ could be grasped in a complex but decisive formula: two natures in one person. The concept of person, taken from Roman law, served to join the fully divine and fully human natures of Christ into an individual unity. Christology is not the product of abstract, logical operations but instead originates in the liturgical and charismatic sphere wherein Christians engage in prayer, meditation, and asceticism. Not being derived primarily from abstract teaching, it rather changes within the liturgy in new forms and in countless hymns of worship—as in the words of the Easter liturgy:

Christology as originating in the charismatic sphere

> The king of the heavens appeared on earth out of kindness to man and it was with men that he associated. For he took his flesh from a pure virgin and he came forth from her, in that he accepted it. One is the Son, two-fold in essence, but not in person. Therefore in announcing him as in truth perfect God and perfect man, we confess Christ our God.

**Messianic views.** Faith in Jesus Christ is related in the closest way to faith in the Kingdom of God, the coming of which he proclaimed and introduced. Christian eschatological expectations, for their part, were joined with the messianic promises, which underwent a decisive transformation and differentiation in late Judaism, especially in the two centuries just before the appearance of Jesus. Two basic types can be distinguished as influencing the messianic self-understanding of Jesus as well as the faith of his disciples.

The old Jewish view of the fulfillment of the history of salvation was guided by the idea that at the end of the history of the Jewish people the Messiah will come from the house of David and establish the Kingdom of God—an earthly kingdom in which the Anointed of the Lord will gather the tribes of the chosen people and from Jerusalem will establish a world kingdom of peace. Accordingly, the expectation of the Kingdom had an explicitly inner-worldly character. The expectation of an earthly Messiah as the founder of a Jewish kingdom became the strongest impulse for political revolutions, primarily against Hellenistic and Roman dominion. The period preceding the appearance of Jesus was filled with continuous new messianic uprisings in which new messianic personalities appeared and claimed for themselves and their struggles for liberation the miraculous powers of the Kingdom of God. Especially

in Galilee, guerrilla groups were formed in which hope for a better future blazed all the more fiercely, because the present was so unpromising. Jesus disappointed the political expectations of these popular circles; he did not let himself be made a political Messiah. Conversely, it was his opponents who used the political misinterpretation of his person to destroy him. Jesus was condemned and executed by the responsible Roman authorities as a Jewish rioter who rebelled against Roman sovereignty. The inscription on the cross, "Jesus of Nazareth, king of the Jews," cited the motif of political insurrection of a Jewish messianic king against the Roman government as the official reason for his condemnation and execution.

Alongside this political type of messianic expectation there was a second form of eschatological expectation. Its supporters were the pious groups in the country, the Essenes and the Qumrān community on the Dead Sea. Their yearning was directed not toward an earthly Messiah but toward a heavenly anointed one, who would bring not an earthly but a heavenly kingdom. Fulfillment lay not in the old world but in the future, coming world, for which the main thing was to prepare oneself through repentance. These pious ones wanted to know nothing of sword and struggle, uprising and rebellion. They believed that the wondrous power of God alone would create the new time. The birth of a new eon would be preceded by intense messianic woes and a frightful judgment upon the godless, the pagan peoples, and Satan with his demonic powers. The Messiah would come not as an earthly king from the house of David but as a heavenly figure, as the Son of God, a heavenly being of the ages, who would descend into the world of the Evil One and there gather his own to lead them back into the realm of light. He would take up dominion of the world and, after overcoming all earthly and supernatural demonic powers, lay the entire cosmos at the feet of God.

Hope for a future world

A second new feature, anticipation of the Resurrection, was coupled with this transcending of the old expectation. According to the old Jewish eschatological expectation, the beneficiaries of the divine development of the world would be only the members of the last generation of humanity who were fortunate enough to experience the arrival of the Messiah upon Earth; all earlier generations would be consumed with the longing for fulfillment but would die without experiencing it. Ancient Judaism knew no hope of resurrection. In connection with the transcending of the expectation of the Kingdom of God, however, even anticipations of resurrection voiced earlier by Zoroastrianism were achieved: the Kingdom of God was to include within itself in the state of resurrection all the faithful of every generation of humanity. Even the faithful of the earlier generations would find in resurrection the realization of their faith. In the new eon the Messiah–Son of man would rule over the resurrected faithful of all times and all peoples. A characteristic breaking free of the eschatological expectation was thereby presented. It no longer referred exclusively to the Jews alone; with its transcendence a universalistic feature entered into it.

Jesus—in contrast to John the Baptist (a preacher of repentance who pointed to the coming bringer of the Kingdom)—knew himself to be the one who brought fulfillment of the Kingdom itself, because the wondrous powers of the Kingdom of God were already at work in him. He proclaimed the glad news that the long promised Kingdom was already dawning, that the consummation was here. This is what was new: the promised Kingdom, supra-worldly, of the future, the coming new eon, already reached redeemingly into the this-worldly from its beyondness, as a charismatic reality that brought people together in a new community.

Jesus as the Son of man

Jesus did not simply transfer to himself the promise of heavenly Son of man, as it was articulated in the apocryphal First Book of Enoch. Instead, he gave this expectation of the Son of man an entirely new interpretation. Pious Jewish circles, such as the Enoch community and other pietist groups, expected in the coming Son of man a figure of light from on high, a heavenly conquering hero, with all the marks of divine power and glory. Jesus, however, linked expectations of the Son of man with the figure of the suffering servant of God (as in Isaiah, chapter 53). He would return in glory as the consummator of the Kingdom. This self-understanding of Jesus was compatible with Christologies derived from the concept of the divine Logos.

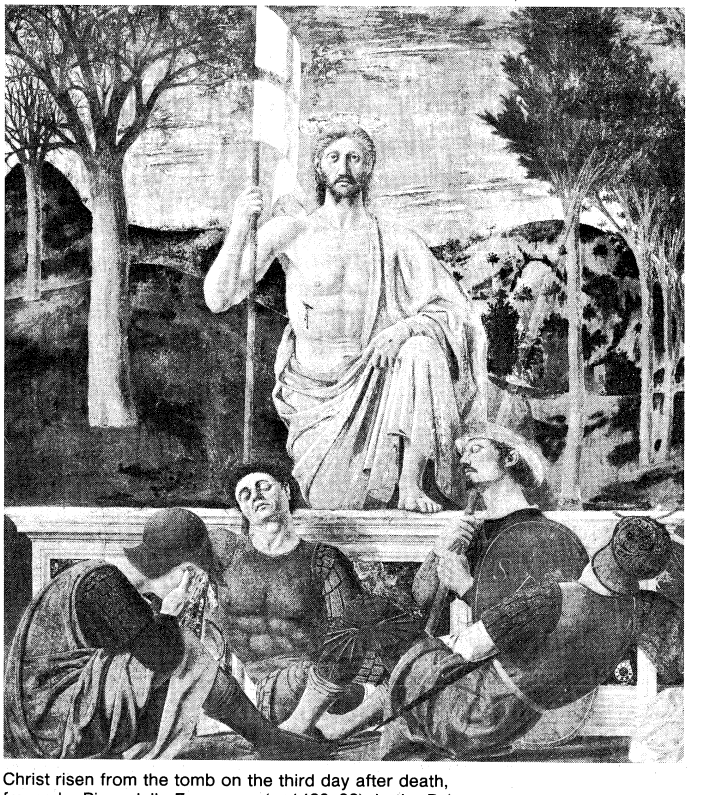
**The doctrine of the Virgin Mary and holy Wisdom.** The dogma of the Virgin Mary as the "mother of God" and "bearer of God" is connected in the closest way with the dogma of the incarnation of the divine Logos. The theoretical formation of doctrine did not bring the cult of the mother of God along in its train; instead, the doctrine only reflected the unusually great role that the veneration of the mother of God already had taken on at an early date in the liturgy and in the church piety of Orthodox faithful.

The expansion of the veneration of the Virgin Mary as the bearer of God (Theotokos) and the formation of the corresponding dogma is one of the most astonishing occurrences in the history of the early church. The New Testament offers only scanty points of departure for this development. Mary completely recedes behind the figure of Jesus Christ, who stands in the centre of all four Gospels. From the Gospels themselves it can be recognized that Jesus' development into the preacher of the Kingdom of God took place in sharp opposition to his family, who were so little convinced of his mission that they held him to be insane (Mark 3:21). Accordingly, all the Gospels stress the fact that Jesus separated himself from his family. Even the Gospel According to John still preserved traces of Jesus' tense relationship with his mother. Mary appears twice without being called by name the mother of Jesus; and Jesus himself regularly withholds from her the designation of mother. The saying, "Woman, what have you to do with me?" (John 2:4), is indeed the strongest expression of a conscious distancing.

The opposition of Jesus' family toward him

Nevertheless, with the conception of Jesus Christ as the Son of God, a tendency developed early in the church to grant to the mother of the Son of God a special place within the church. This development was sketched quite hesitantly in the New Testament. Only the prehistories in Matthew and Luke mention the virgin birth, which, however, cannot be simply coordinated or reconciled with the statements of the preceding genealogical tables. On these scanty presuppositions the later cult of the mother

Anderson—Alinari from Art Resource/EB Inc.



Christ risen from the tomb on the third day after death, fresco by Piero della Francesca (*c.* 1420–92). In the Palazzo Comunale, Sansepolcro, Italy.



The Virgin Mary as the Mother of Mercy, panel by Lippo Memmi (c. 1285–1361). In the dome of the Orvieto Cathedral, Italy.

Anderson—Alinari from Art Resource/EB Inc.

of God was developed. The view of the virgin birth entered into the creed of all Christianity and became one of the strongest religious impulses in the development of the dogma, liturgy, and ecclesiastical piety of the early church.

Veneration of the mother of God received its impetus when the Christian Church became the imperial church under Constantine and the pagan masses came under Christian influences and became members of the church. The peoples of the Mediterranean area and the Middle East could not make themselves conversant with the absolute power of God the Father and with the strict patriarchalism of the Jewish idea of God, which the original Christian message had taken over. Their piety and religious consciousness had been formed for millennia through the cult of the "great mother" goddess and the "divine virgin," a development that led all the way from the old popular religions of Babylonia and Assyria to the mystery cults of the late Hellenistic period. Despite the unfavourable presuppositions in the tradition of the Gospels, cultic veneration of the divine virgin and mother found within the Christian Church a new possibility of expression in the worship of Mary as the virgin mother of God, in whom was achieved the mysterious union of the divine Logos with human nature. The spontaneous impulse of popular piety, which pushed in this direction, moved far in advance of the practice and doctrine of the church. In Egypt, Mary was, at an early point, already worshiped under the title of Theotokos—an expression that Origen used in the 3rd century. The Council of Ephesus (431) raised this designation to a dogmatic standard. To the latter, the second Council of Constantinople (553) added the title "eternal Virgin." In the prayers and hymns of the Orthodox Church the name of the mother of God is invoked as often as is the name of Christ and the Holy Trinity.

Absorption of the cult of the "great mother"

The doctrine of the heavenly Wisdom (Sophia) represents an Eastern Church particularity. In late Judaism, speculations about the heavenly Wisdom—a heavenly figure beside God that presents itself to humanity as mediator in the work of creation as well as mediator of the knowledge of God—abounded. In Roman Catholic doctrine, Mary, the mother of God, was identified with the figure of the divine Wisdom. To borrow a term used in Christology to describe Jesus as being of the same substance (hypostasis) as the Father, Mary was seen as possessing a divine hypostasis.

Mary and heavenly Wisdom

This process of treating Mary and the heavenly Wisdom alike did not take place in the realm of the Eastern Orthodox Church. For all its veneration of the mother of God, the Eastern Orthodox Church never forgot that the root of this veneration lay in the incarnation of the divine Logos that took place through her. Accordingly, in the tradition of Orthodox theology, a specific doctrine of the heavenly Wisdom, Sophianism, is found alongside the doctrine of the mother of God. This distinction between the mother of God and the heavenly Sophia in 20th-century Russian philosophy of religion (in the works of Vladimir Solovyov, Pavel Florensky, W.N. Iljin, and Sergey Bulgakov) developed a special Sophianism. Sophianism did, however, evoke the opposition of Orthodox academic theology. The numerous great churches of Hagia Sophia, foremost among them the cathedral by that name in Constantinople (Istanbul), are consecrated to this figure of the heavenly Wisdom.

## GOD THE HOLY SPIRIT

**Contradictory aspects of the Holy Spirit.** The Holy Spirit of God becomes one of the most elusive and difficult themes in Christian theology, because it refers to one of the three Persons in the Godhead but does not evoke concrete images the way "Father" or "Creator" and "Son" or "Redeemer" do. Reference to the Holy Spirit includes the true creative element in the life of the church. It works in an apparently contradictory sense: by virtue of its authority, the Holy Spirit establishes law and breaks law, it institutes order and breaks order, it founds tradition and breaks tradition. It is the conservative as well as the revolutionary principle in church history. It guarantees the continuity of the church and yet it interrupts this continuity ever again through new creations. Both sides of this activity stand in a characteristic relationship of tension to one another.

The essence of free spontaneity

The essence of the expression of the Holy Spirit is free spontaneity. The Spirit blows like the wind, "where it wills," but where it blows it establishes a firm norm by virtue of its divine authority. The spirit of prophecy and the spirit of knowledge (*gnōsis*) are not subject to the will of the prophet and the enlightened one; revelation of the Spirit in the prophetic word or in the word of knowledge becomes Holy Scripture, which as "divinely breathed" "cannot be broken" and lays claim to a lasting validity for the church.

The Spirit, which is expressed in the various officeholders of the church, likewise founds the authority of ecclesiastical offices. The laying on of hands, as a sign of the transference of the Holy Spirit from one person to another, is a characteristic ritual that visibly represents and guarantees the continuity of the working of the Spirit in the officeholders chosen by the Apostles; it becomes the sacramental sign of the succession of the full power of spiritual authority of bishops and priests. The Holy Spirit also creates the sacraments and guarantees the constancy of their action in the church. All the expressions of church life—doctrine, office, polity, sacraments, power to loosen and to bind, and prayer—are understood as endowed by the Spirit.

The same Holy Spirit, however, also comes forth as the revolutionizing, freshly creating principle in church history. All the decisive reformational movements in church history, which broke with old institutions, have appealed to the authority of the Holy Spirit. This is probably the main reason that in the history of church dogma the article of the Holy Spirit has been developed only hesitantly and incompletely in comparison with the Christological article. A characteristic view of the Holy Spirit is sketched out in the Gospel According to John: the outpouring of the Holy Spirit takes place only after the Ascension of

Christ; it is the beginning of a new time of salvation, in which the Holy Spirit is sent as the Paraclete (Counsellor) to the church remaining behind in this world. The ecstatic phenomena, which are prominent in the church at Pentecost, are understood as fulfillment of this promise. With this event (Pentecost) the church entered into the period of the Holy Spirit. After a process of institutionalization in the church, opposition against it—through appeal to the Holy Spirit—became noticeable for the first time in Montanism, in the mid-2nd century. Montanus, a Phrygian prophet and charismatic leader, understood himself and the prophetic movement sustained by him as the fulfillment of the promise of the coming of the Paraclete. In the 13th century a spiritualistic countermovement against the feudalized institutional church gained attention anew in Joachim of Fiore, who understood the history of salvation in terms of a continuing self-realization of the divine Trinity in the three times of salvation: (1) the time of the Father, (2) the time of the Son, and (3) the time of the Holy Spirit. He promised the speedy beginning of the period of the Holy Spirit, in which the institutional papal church, with its sacraments and its revelation hardened in the letter of scripture, would be replaced by a community of charismatic figures, filled with the Spirit, and by the time of "spiritual knowledge." This promise became the spiritual stimulus of a series of revolutionary movements within the medieval church—*e.g.,* the reform movement of the radical Franciscan spirituals and the Hussite reform movement led by Jan Hus in 15th-century Bohemia. Their effects extend to a 16th-century radical reformer, Thomas Müntzer, who substantiated his revolution against the princes and clerical hierarchs with a new outpouring of the Spirit. Quakerism represents the most radical mode of rejection—carried out in the name of the freedom of the Holy Spirit—of all institutional forms, which are regarded as shackles and prisons of the Holy Spirit. In the 20th century a revival of charismatic forms of Christianity, called Pentecostalism and the charismatic movement, centred on the recovery of the experience of the Holy Spirit and necessitated some fresh theological inquiry about the subject.

Revolutionary movements appealing to the Holy Spirit

**Conflict between order and charismatic freedom.** As the fundamentally uncontrollable principle of life in the church, the Holy Spirit considerably upset Christian congregations from the very outset. Paul struggled to restrict the anarchist elements, which are connected with the appearance of free charismata (spiritual phenomena), and, over against these, to achieve a firm order in the church. Paul at times attempted to control and even repress charismatic activities, which he seemed to regard as irrational or prerational and thus potentially disruptive of fellowship. Among these were glossolalia, or speaking in tongues, a form of unrepressed speech. Paul preferred rational discourse in sermons. He also felt that spontaneity threatened the focus of worship. This tendency led to an emphasis on ecclesiastical offices with their limited authority vis-à-vis the uncontrolled appearance of free charismatic figures.

The conflict between church leadership resident in the locality and the appearance of free charismatic figures in the form of itinerant preachers forms the main motif of the oldest efforts for a church order. This difficulty became evident in the *Didachē,* the *Teaching of the Twelve Apostles* (early 2nd century). The authority of the Holy Spirit, in whose name the free charismatic figures speak, does not allow its instructions and prophecies to be criticized in terms of contents; its evaluation had to be made dependent upon purely ethical qualifications. This tension ended, in practical terms, with the exclusion of the free charismatic figures from the leadership of the church. The charismatic continuation of the revelation, in the form of new scriptures of revelation, was also checked. In the long historical process during which the Christian biblical canon took shape, Bishop Athanasius of Alexandria, in his 39th Easter letter (367), selected the number of writings—of apostolic origin—that he considered "canonical." Revelation in the form of Holy Scriptures binding for the Christian faith was thereby considered definitively concluded. The canon, henceforth fixed, can no longer be changed, abridged, or supplemented.

The fixing of revelation in a set number of sacred books

In a similar way, individual charismatic offices become institutionalized. A lower degree of consecration—a first stage for priestly ordination—still holds for the exorcist, the ritual figure who drives the devil from the possessed or those who are to be baptized. The teacher (*didaskalos*) also becomes institutionalized. In the Roman Church, only ordained priests are permitted to be church teachers—in contrast to the Eastern Orthodox Church, which until the 20th century did not require ecclesiastical ordination of a professor of theology.

The article about the Holy Spirit in the church creeds reflects little of these struggles. It suppresses the revolutionary principle of the Holy Spirit. Neither the so-called Apostles' Creed nor the Nicene Creed goes beyond establishment of faith in the Holy Spirit and its participation in the incarnation. In the Nicene Creed it is further emphasized that the Holy Spirit is the life-creating power—*i.e.,* the power both of creation and of rebirth—and that the Spirit has already spoken in the prophets.

The emergence of Trinitarian speculations in early church theology led to great difficulties in the article about the "person" of the Holy Spirit. In the New Testament the Holy Spirit tended to be present more as power than as person, though there was distinctive personal representation in the form of the dove at Jesus' baptism. But it was difficult to incorporate this graphic or symbolic representation into dogmatic theology. Nevertheless, with Athanasius the idea of the complete essence (*homoousia*) of the Holy Spirit with the Father and the Son was achieved. This was in opposition to all earlier attempts to subordinate the Holy Spirit to the Son and to the Father and to interpret the Spirit—similarly to anti-Trinitarian Christology—as a prince of the angels. According to Athanasius, the Holy Spirit alone guarantees the complete redemption of humanity: "through participation in the Holy Spirit we partake of the divine nature." In his work *De Trinitate,* Augustine undertook to render the essence of the Trinity understandable in terms of the Trinitarian structure of the human person: the Holy Spirit appears as the Spirit of love, which joins Father and Son and draws people into this communion of love. In Eastern Church theological thought, however, the Holy Spirit and the Son both proceed from the Father. In the West, the divine Trinity is determined more by the idea of the inner Trinitarian life in God; thus, the notion was carried through that the Holy Spirit goes forth from the Father and from the Son. Despite all the efforts of speculative theology, a graphic conception of the person of the Holy Spirit was not developed even later in the consciousness of the church.

The relationship of the Holy Spirit to the Father and the Son

**The operations of the Holy Spirit.** For the Christian faith, the Holy Spirit is clearly recognizable in charismatic figures (the saints), in whom the gifts of grace (charismata) of the Holy Spirit are expressed in different forms: reformers and other charismatic figures. The prophet, for instance, belongs to these charismatic types. The history of the church knows a continuous series of prophetic types, which reaches from New Testament prophets, such as Agabus (in Acts 11:28), through the 12th-century monk Bernard of Clairvaux to such Reformers as Luther and Calvin. Christoph Kotter and Nicolaus Drabicius—prophets of the Thirty Years' War period—were highly praised by the 17th-century Moravian bishop John Amos Comenius. Other prophets have existed in Pietism, Puritanism, and the Anglo-Saxon Free churches.

Prophetic women are especially numerous. In church history they begin with Anna (in Luke 2:36) and the prophetic daughters of the apostle Philip. Others are: Hildegard von Bingen, St. Bridget of Sweden, Joan of Arc, and the prophetic women of the Reformation period. In the modern world numbers of pioneers in the "holiness" and Pentecostal traditions were women, and women's gifts of prophecy have sometimes been cherished among Pentecostalists when they were overlooked or disdained by much of the rest of Christianity.

A further type of charismatic person is the healer, who functioned in the early church as an exorcist but who also emerged as a charismatic type in healing personalities of more recent church history (*e.g.,* Vincent de Paul in the 17th century). Equally significant is the curer-of-souls

Types of charismatic leaders

type, who exercises the gift of "distinguishing between spirits" in daily association with people. This gift is found especially emphasized among many of the great saints of all times. In the 19th century it particularly stands out in Johann Christoph Blumhardt, in Protestantism, and in Jean-Baptiste Vianney, the curé of Ars, in Roman Catholicism.

The charismatic wanderer type, who leads a roving life in imitation of Jesus Christ, who "has nowhere to lay his head," was molded through the ideal of "ascetic homelessness." The latter drove Scots-Irish monks, for example, not only through all of Europe but also to the remotest islands of the northern seas and as far as Iceland and Newfoundland. This ideal is still alive today in the Eastern Orthodox Church in the form of the *strannik* ("wanderer"). The "holy fool" type conceals a radical Christianity under the mask of foolishness and holds the truth of the gospel, in the disguise of folly, before the eyes of highly placed personalities: the worldly and the princes of the church who do not brook unmasked truth. This type, which frequently appeared in the Byzantine Church, has been represented especially in Western Christianity by Philip Neri, the founder of the religious order known as the Oratorians, in the 16th century.

The charismatic teacher (*didaskalos*), on the other hand, still appears. Filled with the spirit of intelligence or knowledge of the Holy Spirit, he carries out his teaching office, which does not necessarily need to be attached to an academic position. Many Free Church and ecclesiastical reform movements owe their genesis to such spirit-filled teachers, who are often decried as anomalous. The deacon likewise is originally the holder of a charismatic office of selfless service. Christian service, or *diakonia,* was not confined to Christian offices. Some of the energies that once went into it are now found in social service outside the church. Many of the agents of such service were originally or still may be inspired by Christian norms and examples in the care of the sick and the socially outcast or overlooked. Alongside such men as the Pietist August Hermann Francke, the Methodist John Wesley, Johann Wichern (the founder of the Inner Mission in Germany), and Friederich von Bodelschwingh (the founder of charitable institutions), important women have appeared as bearers of this charisma (*e.g.,* the English nurse Florence Nightingale and the Salvation Army leader Catherine Booth).

The Holy Spirit that "blows where it wills" has often been recognized as the impulse behind an enlargement of roles for women in the church. However limited these have been, they enlarged upon those that Christians inherited from Judaism. Partitions had screened women in a special left-hand section of the synagogue. While the pace of innovation was irregular, in the ecstatic worship services of the Christian congregations women tended to participate in speaking in tongues, hymns, prayer calls, or even prophecies. Evidently, this innovation in the face of the strict synagogal custom was held admissible on the basis of the authority of the Holy Spirit: "Do not quench the Spirit" (1 Thessalonians 5:19). Inasmuch as the appearance of charismatic women upset traditional concepts, however, Paul, acting on any number of personal and social motivations, reverted to the synagogal principle and inhibited the speaking role of women: "the women should keep silence in the churches." (1 Corinthians 14:34).

Because expressions of free charisma were increasingly suppressed in the institutional churches, the emergence of Pentecostal movements outside the institutional churches and partly in open opposition to them arose. This movement led to the founding of various Pentecostal Free churches at the end of the 19th century and the beginning of the 20th; today, it is represented through numerous independent Pentecostal groups, such as the Church of God and the Assemblies of God. At first scorned by the established churches and devalued as "demonic," the Pentecostal movement has grown to a world movement with strong missionary activity not only in Africa and South America but also in the European countries. In the United States, a strong influence of the Pentecostal movement—which has returned high esteem to the proto-Christian charismata of speaking in tongues, healing, and exorcism—is noticeable in the older churches as well, even in the Roman Catholic, Lutheran, and Anglican. This has occurred especially in liturgy and church music but also in preaching style and the return to faith healing.

## THE HOLY TRINITY

**The basis for the doctrine of the Trinity.** The central Christian affirmations about God are condensed and focused in the classic doctrine of the Trinity, which has its ultimate foundation in the special religious experience of the Christians in the first communities. This basis of experience is older than the doctrine of the Trinity. It consisted of the fact that God came to meet Christians in a threefold figure: (1) as Creator, Lord of the history of salvation, Father, and Judge, as revealed in the Old Testament; (2) as the Lord who, in the figure of Jesus Christ, lived among human beings and was present in their midst as the "Resurrected One"; and (3) as the Holy Spirit, whom they experienced as the power of the new life, the miraculous potency of the Kingdom of God. The question as to how to reconcile the encounter with God in this threefold figure with faith in the oneness of God, which was the Jews' and Christians' characteristic mark of distinction from paganism, agitated the piety of ancient Christendom in the deepest way. In the course of history, it also provided the strongest impetus for a speculative theology, which inspired Western metaphysics for many centuries. In the first two centuries of the Christian Era, however, a series of different answers to this question stood in juxtaposition. At first none of the Christian theologians had considered them speculatively.

The revelation of God in three figures

The diversity in interpretation of the Trinity was conditioned especially through the understanding of the figure of Jesus Christ. According to the theology of the Gospel According to John, the *divinity* of Jesus Christ constituted the departure point for understanding his person and efficacy. The Gospel According to Mark, however, did not proceed from a theology of incarnation but instead understood the baptism of Jesus Christ as the adoption of the

Novosti Press Agency



Testament Trinity, detail from the Four-Part-Icon, Novgorod School, *c.* 1400. In the State Russian Museum, Leningrad.

*man* Jesus Christ into the Sonship of God, accomplished through the descent of the Holy Spirit. The situation became further aggravated by the conceptions of the special personal character of the manifestation of God developed by way of the historical figure of Jesus Christ; the Holy Spirit was viewed not as a personal figure but rather as a power and appeared graphically only in the form of the dove and thus receded, to a large extent, in the Trinitarian speculation.

**Introduction of Neoplatonic themes.** In the Johannine literature in the Bible there appeared the first traces of the concept of Christ as the Logos, the "word" or "principle" that issues from eternity. Under the influence of subsequent Neoplatonic philosophy, this tradition became central in speculative theology. There was interest in the relationship of the "oneness" of God to the "triplicity" of divine manifestations. This question was answered through the Neoplatonic metaphysics of being. The transcendent God, who is beyond all being, all rationality, and all conceptuality, is divested of divine transcendence. In a first act of becoming self-conscious the Logos recognizes itself as the divine mind (Greek: *nous*), or divine world reason, which was characterized by the Neoplatonic philosopher Plotinus as the "Son" who goes forth from the Father. The next step by which the transcendent God becomes self-conscious consists in the appearance in the divine *nous* of the divine world, the idea of the world in its individual forms as the content of the divine consciousness. In Neoplatonic philosophy both the *nous* and the idea of the world are designated the hypostases of the transcendent God. Christian theology took the Neoplatonic metaphysics of substance as well as its doctrine of hypostases as the departure point for interpreting the relationship of the "Father" to the "Son" in terms of the Neoplatonic hypostases doctrine. This process stands in direct relationship with a speculative interpretation of Christology in connection with Neoplatonic Logos speculation.

The assumption of the Neoplatonic hypostases doctrine meant from the beginning a certain evaluation of the relationships of the three divine figures to one another, because for Neoplatonism the process of hypostatization is at the same time a process that includes a diminishing of being. Thus, in flowing forth from the transcendent source, the divine being is progressively weakened with the distance from the transcendent origin. Diminution of being, on these terms, is brought about through approach to matter, which for its part is understood in Neoplatonism as nonbeing. In transferring the Neoplatonic hypostases doctrine to the Christian interpretation of the Trinity there existed the danger that the different manifestations of God—as known by the Christian experience of faith: Father, Son, Holy Spirit—would be transformed into a hierarchy of gods graduated among themselves and thus into a polytheism. Though this danger was consciously avoided and, proceeding from a Logos Christology, the complete sameness of essence of the three manifestations of God was emphasized, there arose the danger of a relapse into a triplicity of equally ranked gods, which would displace the idea of the oneness of God.

**Attempts to define the Trinity.** By the 3rd century it was already apparent that all attempts to systematize the mystery of the divine Trinity with the theories of Neoplatonic hypostases metaphysics were unsatisfying and led to a constant series of new conflicts. The high point, upon which the basic difficulties underwent their most forceful theological and ecclesiastically political actualization, was the so-called Arian controversy. Arius belonged to the Antiochene school of theology, which placed strong emphasis upon the historicity of the man Jesus Christ. In his theological interpretation of the idea of God, Arius was interested in maintaining a formal understanding of the oneness of God. In defense of the oneness of God, he was obliged to dispute the sameness of essence of the Son and the Holy Spirit with God the Father, as stressed by the theologians of the Neoplatonically influenced Alexandrian school. From the outset, the controversy between both parties took place upon the common basis of the Neoplatonic concept of substance, which was foreign to the New Testament itself. It is no wonder that the continuation of the dispute on the basis of the metaphysics of substance likewise led to concepts that have no foundation in the New Testament—such as the question of the sameness of essence (*homoousia*) or similarity of essence (*homoiousia*) of the divine persons.

Emphasis on the historicity of Jesus

The basic concern of Arius was and remained disputing the oneness of essence of the Son and the Holy Spirit with God the Father, in order to preserve the oneness of God. The Son, thus, became a "second God, under God the Father"—*i.e.,* he is God only in a figurative sense, for he belongs on the side of the creatures, even if at their highest summit. Here Arius joined an older tradition of Christology, which had already played a role in Rome in the early 2nd century—namely, the so-called angel-Christology. The descent of the Son to Earth was understood as the descent to Earth of the highest prince of the angels, who became man in Jesus Christ; he is to some extent identified with the angel prince Michael. In the old angel-Christology the concern is already expressed to preserve the oneness of God, the inviolable distinguishing mark of the Jewish and Christian faiths over against all paganism. The Son is not himself God, but as the highest of the created spiritual beings he is moved as close as possible to God. Arius joined this tradition with the same aim—*i.e.,* defending the idea of the oneness of the Christian concept of God against all reproaches that Christianity introduces a new, more sublime form of polytheism.

The angel-Christology of the 2nd century

This attempt to save the oneness of God led, however, to an awkward consequence. For Jesus Christ, as the divine Logos become human, moves thereby to the side of the creatures—*i.e.,* to the side of the created world that needs redemption. How, then, should such a Christ, himself a part of the creation, be able to achieve the redemption of the world? On the whole, the Christian Church rejected, as an unhappy attack upon the reality of redemption, such a formal attempt at saving the oneness of God as was undertaken by Arius.

The main speaker for church orthodoxy was Athanasius of Alexandria, for whom the point of departure was not a philosophical-speculative principle but rather the reality of redemption, the certainty of salvation. The redemption of humanity from sin and death is only then guaranteed if Christ is total God and total human being, if the complete essence of God penetrates human nature right into the deepest layer of its carnal corporeality. Only if God in the full meaning of divine essence became human in Jesus Christ is deification of man in terms of overcoming sin and death guaranteed as the resurrection of the flesh.

Augustine, of decisive importance for the Western development of the Trinitarian doctrine in theology and metaphysics, coupled the doctrine of the Trinity with anthropology. Proceeding from the idea that humans are created by God according to the divine image, he attempted to explain the mystery of the Trinity by uncovering traces of the Trinity in the human personality. He went from analysis of the Trinitarian structure of the simple act of cognition to ascertainment of the Trinitarian structure both of human self-consciousness and of the act of religious contemplation in which people recognize themselves as the image of God.

A second model of Trinitarian doctrine—suspected of heresy from the outset—which had effects not only in theology but also in the social metaphysics of the West as well, emanated from Joachim of Fiore. He understood the course of the history of salvation as the successive realization of the Father, the Son, and the Holy Spirit in three consecutive periods of salvation. This interpretation of the Trinity became effective as a "theology of revolution," inasmuch as it was regarded as the theological justification of the endeavour to accelerate the arrival of the third state of the Holy Spirit through revolutionary initiative.

The final dogmatic formulation of the Trinitarian doctrine in the so-called Athanasian Creed (*c.* 500), *una substantia—tres personae* ("one substance—three persons"), reached back to the formulation of Tertullian. In practical terms it meant a compromise in that it held fast to both basic ideas of Christian revelation—the oneness of God and divine self-revelation in the figures of the Father, the Son, and the Holy Spirit—without rationalizing

The Athanasian formula

the mystery itself. In the final analysis the point of view thereby remained definitive that the fundamental assumptions of the reality of salvation and redemption are to be retained and not sacrificed to the concern of a rational monotheism.

Characteristically, in all periods of the later history of Christendom in which a rationalistic philosophy was achieved and the history of salvation aspect of the Trinitarian question receded, anti-Trinitarian currents returned. Many, to some extent, consciously rejoined ties with Arius: the humanist Enlightenment of the 16th century and the so-called anti-Trinitarians of the Italian Renaissance. A direct connection exists between anti-Trinitarianism and 18th-century research into the life of Jesus. The oldest life of Jesus researchers in the 18th century, such as Hermann Reimarus and Karl Bahrdt, who portrayed Jesus as the agent of a secret enlightenment order that had set itself the goal of spreading the religion of reason in the world, were at the same time anti-Trinitarians and pioneers of the radical rationalistic criticism of dogma. The Kantian critique of the proofs of God contributed further to a devaluation of Trinitarian doctrine. In the philosophy of German Idealism, Hegel, in the framework of his attempt to raise Christian dogma into the sphere of the conceptual, took the Christian Trinitarian doctrine as the basis for his system of philosophy and, above all, for his interpretation of history as the absolute spirit's becoming self-conscious. In more recent theology, at least in the accusations of some of its critics, the school of dialectical theology in Europe and the United States tended to reduce the doctrine of the Trinity and supplant it with a monochristism.

In a brief but well-publicized episode in the mid-1960s in the United States, a number of celebrated Protestant theologians engaged in cultural criticism observed or announced "the death of God." The theology of the death of God downplayed any notion of divine transcendence and invested its whole claim to be Christian in its accent on Jesus of Nazareth. Christian dogma was reinterpreted and reduced to norms of human sociality and freedom. Before long, however, the majority of theologians confronted this small school with the demands of classic Christian dogma, which insisted on confronting divine transcendence in any assertions about Jesus Christ.

Rediscovery of the transcendence of God

The transcendence of God has been rediscovered by science and sociology; theology in the closing decades of the 20th century endeavoured to overcome the purely anthropological interpretation of religion and once more to discover anew its transcendent ground. Theology has consequently been confronted with the problem of Trinity in a new form, which, in view of the Christian experience of God as an experience of the presence of the Father, Son, and Holy Spirit, cannot be eliminated.

## THE DOCTRINE OF MAN

**What it is to be human.** Most Christian theologians, following the culture and habits of their day, used the general term man to cover both sexes when referring to human beings. The literature on the subject consistently refers to anthropology in theology as "the doctrine of man," but it must be understood as that of "human" or "human being." The starting point for the Christian understanding of what it is to be human is the recognition that humans are created after the image of God. This idea views God and humans joined with one another through a mysterious connection. God is thought of as incomprehensible and beyond substance; yet God desired to reflect the divine image in one set of creatures and chose humans for this. Man as the image of God belongs, therefore, to the self-revelation of God in quite a decisive way. God, being reflected in the human creature, makes this being a partner in the realization of the divine self; there is constant interaction. God and humans belong so closely together that one can say that they are intended for each other. For this the statements of the great mystics are of significance. Man finds fulfillment in God, the divine prototype, but God also first comes to the fulfillment of the divine essence in relation, in this case, with the human.

**The human as a creature.** The idea of the human being as the creature created according to the image of God was already being interpreted in a twofold direction in the early church. For one thing, man, like all other creatures of the universe, is a creation of God. According to human nature, the creature is thus not divine but at the same time is not created out of nothingness; as creatures human beings stand in a relationship of utter dependency on God. They have nothing from themselves but owe everything, even their being, exclusively to the will of the divine Creator; they are joined with all other fellow creatures through a relationship of solidarity. Later, this idea of the solidarity of the creatures among one another almost completely receded behind the idea of the special position of humans and their special commission of dominion. The idea of solidarity with all creatures has been expressed and practiced by but few charismatic personalities in the history of Western piety, such as by Francis of Assisi in his "Canticle of the Sun": "Praised be Thou, my Lord, with all Thy creatures, especially with our sister sun."

The second aspect of the idea of the human being as a creature operated very much more emphatically: the superiority of humans over all other creatures. God placed humans in a special relationship to the divine. God created them in the divine image, thereby assigning to humans a special commission vis-à-vis all other creatures.

**The human as the image of God.** Under the influence of the dualistic philosophy of Plato, Christian theology attempted for some time to regard the image of God in human beings as restricted simply to their intellectual capability and faculty of perception. In his work *De Trinitate,* Augustine attempted to ascertain traces of divine Trinity in the human intellect. Christian mysticism confronted this dualistic view of humans. It understood humans in their mind–body entirety as being in the image of God. The image of God is stamped all the way into the sphere of human corporeality. The idea of human creation according to the image of God is already based upon the intention of the Incarnation, the self-representation of God in corporeality. Even according to their somatic (bodily) condition, humans are the universal form of being, in whom the powers and creative principles of the whole universe are combined in a personal unity of spirit, soul, and body.

The Christian understanding of evil is also linked with the idea of human creation according to the image of God. Evil cannot, in the Christian view, be derived from the dualistic assumption of the contrasts of spirit and body, reason and matter. According to the Christian understanding, the triumph of evil is not identical with the victory of matter, the "flesh," over the spirit. Such a dualistic interpretation has frequently been furthered by the fact that for many centuries the Christian understanding of sin, even among many of the church's teachers, was influenced by the philosophical assumptions of Neoplatonic dualism. Moreover, in Augustine there are still the aftereffects of Manichaeism, which—out of the dualistic conceptions of Zoroastrian religion—ultimately viewed the main motive force of sin in "concupiscence"—*i.e.,* the sex drive.

The idea of freedom

The only genuine departure point for the Christian view of evil is the idea of freedom, which is based in the concept of the human being as the image of God. The human is person because God is person. It is apparent in Christian claims that the concept of the human as "being-as-person" is the real seal of that human as "being-as-the-image-of-God," and therein lies the true nobility that distinguishes human beings from all other creatures. If the Christian faith is differentiated from other religions through the fact that for the Christian God is person, then this faith takes effect in the thereby resulting consequence that the human being, too, is person.

God at the same time entered into a great risk in creating the human as person. The real sign of God as personal being is freedom. When God created humans according to his image, he also gave over to them this mark of nobility—*i.e.,* freedom. This alone constitutes the presupposition of love. Only through this freedom can the human being as partner of God offer free love to God; only in this freedom can God's love be answered through free love in return. Love in its fulfilled form, according to the Christian understanding, is possible only between persons;

conversely, the person can be realized only in the complete love to another person. Humans can use this freedom to offer God, their Creator, their freely given love.

Yet, in the gift of freedom itself there also lay enclosed the possibility for humans to decide against God and to raise themselves to the goal of divine love. The event that is portrayed in the Mosaic creation story as the Fall of man (Genesis, chapter 3) is essentially the trying out of freedom, the free decision of humans against God. This rebellion consists of the fact that human beings improperly use their God-given freedom to set themselves against God and even to wish to be "like God."

**Human redemption.** This special interpretation of sin likewise renders understandable the specifically Christian understanding of human redemption, namely, the view of Jesus Christ as the historical figure of the Redeemer—*i.e.*, the specifically Christian view of the incarnation of God in Jesus Christ.

Members of Asian high religions have found it difficult to understand the fundamental Christian idea of the incarnation. The religious person of the East is inclined to understand the Christian idea of incarnation as an analogy to the Hindu concept of the *avatāra* (best rendered as incarnation). The starting point of the latter is that the divine descends to Earth ever again and is constantly clothed anew in a human figure, in order to reveal the heavenly truth to every era and all people in a manner comprehensible to them. Thus, it was natural to understand the figure of Jesus Christ also as such an *avatāra,* as a form of descent of the divine to mankind. In the realm of Hinduism ever-new attempts are found to comprehend Christology in this sense.

The difference between the Hindu *avatāra* and the Christ

The Christian understanding of the incarnation, however, is based upon a fundamentally different idea, which is enclosed in the simple saying of the Gospel According to John: "The Word became flesh" (chapter 1, verse 14). Whereas the *avatāra* concept assumes that the divine appears in the cyclic lapse of time periods—continually occurring, now in this, now in that earthly veil—the incarnation of the divine Logos in Jesus Christ is, according to the Christian view, a definitively unique happening. One might say that the Christian view of incarnation has an extremely material, even materialistic, feature. In Christianity, it is not a transcendent, divine being that takes on the appearance of an earthly corporeality, so as to be manifested through this semblance of a body; instead, God himself as human, as member of a definite people, a definite family, at a certain time—"suffered under Pontius Pilate"—enters into the corporeality, carnality, and materiality of the history of mankind. In the midst of history God creates the beginning of a thorough transformation of humans that in like manner embraces all spheres of human being—matter, soul, and mind. Incarnation so constituted did not have the character of veiling God in a human form, which would enable the divine being to reveal a new teaching with human words. The incarnation is not the special instance of a cyclic descent of God always occurring afresh in constantly new veils. Instead, it is the unique intervention of God in the history of the human world. Therein God took the figure of a single historical person into the divine being, suffered through the historical conditions of being, and overcame in this person, Jesus Christ, the root of human corruption—the misuse of freedom. God thereby established the dawn of a transformed, renewed, exalted form of human being and opened a realm in which love to God and to neighbour can be tranquilly fulfilled.

**The problem of suffering.** Here is raised the decisive question of the place of suffering within the Christian anthropology. Christianity's opponents have ever again reproached it with glorifying suffering instead of overcoming it. This reproof seems to many to be not entirely unjustified. There have in fact been eras in the history of Christian piety in which suffering as such underwent a frankly ecstatic glorification. This was especially so in several periods of the Middle Ages, in which the Christian Church was convulsed by the severest inner and outer crises and Christ appeared predominantly in the figure of the man of suffering.

The starting point for the Christian understanding of suffering is the messianic self-understanding of Jesus himself. A temptation to power and self-exaltation lay in the late Jewish promise of the coming of the Messiah–Son of man. The Gospel According to Matthew described the temptation of Jesus by Satan in the wilderness as a temptation to worldly power. Jesus himself deeply disappointed his disciples' notions aiming at power and exaltation, in that he taught them, in accordance with Isaiah, chapter 53: "The Son of man will suffer many things." Already in Jesus' announcements of suffering the Christian understanding of suffering is brought clearly to expression: suffering is not the final aim and end in itself in the realization of human destiny; it is the gateway to resurrection, to rebirth, to new creation. This idea receives its clarification from the Christian understanding of sin. Sin as the misuse of human freedom has led humans into total opposition against God, who in return delivers them over to death. Turning to God can therefore take place only when the results of this rebellion are overcome in all levels of human being, all the way to physical corporeality.

The Christian understanding of suffering

In the early church the sign of the cross was not considered a glorification of suffering but a "sign of victory" (*tropaion*) in the sense of the ancient triumphal sign that was set up at the place where the victorious turning point of the battle took place. The cross was likewise considered the "dread of the demons," since as a victory sign it struck terror into the hitherto ruling demonic powers of the world. An ancient church hymn of the cross spoke of the "cross of the beauty of the Kingdom of God." Christ generally appeared upon early church representations of the cross as the crowned victor, who in such figures is represented as the lord of the new eon, the new age promised in the coming of Christ. The emperor Constantine thus fastened to the standards of the imperial legions the cross, which was considered the victory sign for the community of Christians hitherto persecuted by the Roman Empire, and elevated it to a token of military triumph over the legions of his pagan foes that were assembled under the sign of the old gods.

The crowned victorious Christ

In the Christian understanding, suffering also does not appear—as in Buddhism—as suffering simply under the general conditions of human existence in this world; it is instead coupled with the specifically Christian idea of the imitation of Christ. Individual Christians are called to become imitators of Christ; incorporation into the body of Christ is granted to those who subsequently are ready to carry out within themselves Christ's destiny of suffering, death, and resurrection. The early church's characterization of the Christian was that of *Christophoros*—"bearer of Christ." Suffering was an unalterable principle in the great drama of freedom, which was identical with the drama of redemption.

**The resurrection of the body.** Just as clear, however, is the real, indeed materialistic, significance that lies in the Christian understanding of the resurrection. A dualistic understanding of what it is to be human, which assumes an essential difference between the spiritual and the material-bodily sides of human existence, necessarily leads to the idea of the immortality of the soul. According to this view, imperishableness belongs to spiritual nature alone. The Christian hope, however, does not aim at the immortality of the soul but at the resurrection of the body. Corporeality is not a quality that is foreign to the spiritual. Everything spiritual presses toward corporealization; its eternal figure is a corporeal figure. This hope was expressed by Vladimir Solovyov:

The resurrection of the body rather than immortality of the soul

> What help would the highest and greatest moral victory be for man, if the enemy, "death," which lurks in the ultimate depth of man's physical, somatic, material sphere, were not overcome?

The goal of redemption is not separation of the spirit from the body; it is rather the new human in the entirety of body, soul, and mind. It is appropriate to say that Christianity has contended for a "holistic" view of the human. The Christian image of the human being has an essentially corporeal aspect that is based in the idea of the incarnation and finds its most palpable expression in the idea of the resurrection.

**Progressive human perfection.** For a long time Christian anthropology in academic theology was dominated by static thinking. The human appeared as a complete being, placed in a finished world like a methodically provided-for tenant in a prefabricated, newly built residence ready for occupation. Redemption was understood just as statically: salvation appeared in the teachings of church dogma as restitution and restoration of the lost divine image and often in fact more a patching up of fragments through ecclesiastical remedies than as a real new creation.

Although it is not an uncontroversial point, there is in the New Testament, in the observation of many, a progression of salvation in history. Indeed, there is a progress of both the individual human being and of mankind as a whole, what might be thought of under some terms and conditions as a potential for the progressive perfection of the human being. This characteristic stands out already in the proclamation of Jesus. He promises his disciples: "Then the righteous will shine like the sun in the kingdom of their Father. He who has ears, let him hear." (Matthew 13:43). In the Gospel According to John, Jesus promises his disciples an increase of their divine powers that is to exceed even the spiritual powers at work in himself (John 14:12). Similar expectations are also expressed in the First Letter of John: "Beloved . . . it does not yet appear what we shall be, but we know that when he appears we shall be like him, for we shall see him as he is" (chapter 3, verse 2).

The New Testament concept of the progression of salvation

The idea of the Christian "superman," which was expressed by Montanus, is a result of this view. In connection with the breakthrough of the idea of evolution through Darwin in the areas of biology, zoology, and anthropology, the tendency asserted itself—above all in 19th-century American theology—of interpreting the Christian history of salvation in terms of the evolution and expectation of future human perfection in the form of reaching even higher charismatic levels and ever higher means of spiritual knowledge and communication.

In and after the mid-20th century a number of theologians, some of them of schools called "process theology" and some in evolutionary camps, have used these biblical clues to develop new understandings of Christian anthropology. These understandings challenge old orthodoxies, and it cannot be said that any of them have securely been worked into the development of classical thought. Yet these schools, influenced by thinkers such as the British philosopher Alfred North Whitehead or the Jesuit paleoanthropologist Teilhard de Chardin, see the human progressing toward later stages of fulfillment "in Christ."

The thought of Teilhard de Chardin

In such forms of Christian natural theology, Christ is not only a past reference point through the incarnation and a present experience in worship and devotion but also a focal point of the collective salvation of the world. In Teilhard's term, this is the "Omega Point" toward which creation is striving and in respect to which it is unfolding. Such Christian naturalists refer to the New Testament Letter to the Ephesians, where the goal of Christian motion is described: "until we all attain to the unity of the faith and of the knowledge of the Son of God, to mature manhood, to the measure of the stature of the fulness of Christ." (Ephesians 4:13). It must be said that evolutionary theology also met reaction and resistance from many more traditional Christians who operated with other (some would say more scholastic or more static) metaphysics, or who found fault with the thought of Whitehead or Teilhard. At the same time, it is safe to say that through the centuries of change in scientific thought, and with the enlarged cultural experience of Christianity apart from the Western world, ways of thinking about God are certain to be altered.

**The justified human.** Since the Reformation of the 16th century in the West, the Christian anthropology of Luther, Zwingli, and Calvin has been oriented primarily toward the schema of justification. The Christian is the one to whom the righteousness of God is ascribed in faith for the sake of the merit of Jesus Christ, which he earned for himself through his expiatory sacrifice on the cross. In the 20th century, however, the schema of justification seems less understandable as the starting point for a Christian anthropology, because Jewish law and the Roman Catholic concept of penance based on Roman law (against which the Reformers fulminated) are scarcely found any more in religious consciousness. Paul only speaks of justification when he becomes "as a Jew to the Jews," but if he speaks to Gentile Christians, then he becomes "as a Greek to the Greeks" and talks to them in ideas and images that are more suitable to the Greek ways of thinking in terms of the mystery religions: the new being, the freed and ransomed human, the new creation, the resurrection with Christ, the process of human transformation and supraformation, and the Sonship and friendship of God.

**The "new man": The human being in the light of Christ.** Probably no idea and no sentiment in the early church dominated the Christian feeling for life so thoroughly and comprehensively as the consciousness of the newness of the life into which persons viewed themselves transposed through participation in the life and body of Christ. The newness of the Christian message of salvation not only filled the hearts of the faithful but was also striking to the non-Christian milieu. The new humans experience and recognize the newness of life as the life of Christ that is beginning to mature in themselves, as the overwhelming experience of a new state already now commencing. In the New Testament statements about the new man, it was not a settled, complete new condition that was being spoken of, into which people are transposed through grace, but rather the beginning of a coming new state, the consummation of which will first take place in the future. The new human is one who is engaged in the process of renewal; new life is a principle of growth of the Christian maturing toward "perfect manhood in Christ." The new situation of human beings, for their part, works anew as fermenting "leaven" within old humankind, as "fresh dough," and contributes to transforming the old form of humanity through its fermentation into the state of the Kingdom of God.

**The "reborn human."** "Rebirth" has often been identified with a definite, temporally datable form of "conversion." Especially the pietistic and revival type of Christianity has contributed to a certain leveling of this concept. In the history of Christian piety a line of prominent personalities experienced their rebirth in the form of a temporally datable and also locally ascertainable conversion event. Fixation upon a single type of experience, however, is factually not justified. There are numerous other forms of completion of that mysterious event characterized with the expression rebirth. The mode of experience of rebirth itself is as manifold as the individuality of the person concerned, his special intellectual or emotional endowment, and his special history. The different forms of rebirth experience are distinguished not only according to whether the event sets in suddenly with overwhelming surprise, as when one is "born again" or "sees the light," or as the result of a slow process, a "growing," a "maturing," and an "evolution." They are also distinguished according to the psychic capability predominant at the time that thereby takes charge (will, intellect), the endowment at hand, and the personal type of religious experience. With the voluntaristic type, rebirth is expressed in a new alignment of the will, in the liberation of new capabilities and powers that were hitherto undeveloped in the person concerned. With the intellectual type, it leads to an activation of the capabilities for understanding, to the breakthrough of a "vision." With others it leads to the discovery of an unexpected beauty in the order of nature or to the discovery of the mysterious meaning of history. With still others it leads to a new vision of the moral life and its orders, to a selfless realization of love of neighbour. In the experience of Christian rebirth, the hitherto existing old condition of humanity is not simply eliminated so far as the given personality structure is concerned—a structure dependent upon heredity, education, and earlier life experiences. Instead, each person affected perceives his life in Christ at any given time as "newness of life."

Forms of "rebirth"

**Human liberation.** The condition of "fallen" humanity is frequently characterized in the New Testament as "slavery." It is the slavery of human willfulness that wants to have and enjoy all things for itself: the slavery of alienated love, which is no longer turned toward God but toward

The problem of slavery: internal and external

one's own self and the things of this world and which also degrades one's fellows into the means for egoism and exploitation. The servitude of people fallen away from God is much more oppressive than mere slavery of the senses and of greed for life. It is the enslavement not only of their "flesh" but also of all levels of their being, even the "most spiritual."

In a bold reversal of the language of Platonic dualism, Luther expressed it thusly in his commentary to the Letter of Paul to the Romans: "The entire man who is not reborn is flesh, even in his spirit; the entire man who is reborn is spirit, even when he eats and sleeps." Only from this perspective do Martin Luther's words about the "Freedom of a Christian Man" (1520) receive their true meaning. The freedom that Christians receive is the freedom that Christ, spoken of by Paul as the new Adam, gained for them by fighting. The freedom of Christians is the freedom reattained in Christ, in which the possibility of the misuse of freedom is addressed and overcome.

In the initial centuries of the church a special significance fell to the evangelical schema of liberation—and to the corresponding schema of ransom—in a society that, in its social structure, was constructed entirely upon the system of slavery. On the one hand, wide strata of the population lived in the permanent state of slavery; on the other hand, on the basis of the prevailing usage of war, even the free population was constantly exposed to the danger of passing into possession of the victor as a slave in case of a conquest. The schema of liberation could therefore count upon a spontaneous understanding.

Freedom alone also makes a perfect community possible. Such a community embraces God and the neighbour, in whom the image of God confronts human beings in the flesh. Community is fulfilled in the free service of love. Luther probably most pertinently articulated the paradox of Christian freedom, which includes both love and service: "A Christian man is a free lord of all things and subordinate to no one. A Christian man is a submissive servant of all things and subject to everyone." Christian freedom is thus to be understood neither purely individually nor purely collectively. The motives of the personal and the social are indivisibly joined by the idea that each person is indeed an image of God for himself alone, but that in Christ he also recognizes the image of God in the neighbour and with the neighbour is a member in the one body of Christ. Here, too, the evolutive principle of the idea of freedom is not to be mistaken; in it, for example, lay the spiritual impetus to the social and racial emancipation of slaves, as it was demanded by the great Christian champions of human rights in the 18th and 19th centuries and, through great efforts, pursued and achieved.

**Joy in human existence.** Friedrich Nietzsche summarized his critique of the Christians of his time in the words of Zarathushtra (Zoroaster): "They would have to sing better songs to me that I might believe in their Redeemer: his disciples would have to look more redeemed!" The critique is to the point. In the New Testament testimonials, joy appears as the characteristic mark of distinction of the Christian. It is the spontaneous result of being filled with the Holy Spirit and is among the main fruits of the Holy Spirit. Joy was the basic mood of congregational gatherings and was often expressed in an exuberant jubilation; it has its origin in the recognition that the dominion of evil is already broken through the power of Christ, that death, devil, and demons no longer possess any claim upon believers, and that the forces of forgiveness, reconciliation, resurrection, and transfiguration are already effective in humankind. This principle of the joy of the Christian is most strongly alive in the liturgy of the Eastern Orthodox Church.

Joy as a characteristic mark of a Christian

The roots of a specifically Christian sense of humour also lie within this joy. Its peculiarity consists of the fact that in the midst of the conflicts of life the Christian is capable of regarding all sufferings and afflictions from the perspective of overcoming them in the future or from the perspective of victory over them already achieved in Christ. In Christian humour, freedom and joy are combined. The Christian does not let himself be confused and tempted through cross and suffering but already perceives in the cross and in suffering a foretaste of eschatological triumph and joy. At one extreme the humour of the Danish philosopher Søren Kierkegaard is too dialectical and too bitter to exhaust the entire fullness of the Christian joy. More of it is found in the "hallelujah" of black spirituals.

**The charismatic believer.** In the New Testament the Christian is depicted as the person who is filled with the powers of the Holy Spirit. The view of the gifts of the Spirit stands in a direct relationship with the understanding of the human as the image of God. For the believing Christian of the original period of the church, the Holy Spirit was the Spirit of the Lord Jesus Christ, who is already now made manifest in his body, the community of the faithful, as the miraculous principle of life of the new eon. Throughout the centuries the Holy Spirit has remained the ferment of church history—all great reformations and numerous foundings of new churches and sects stand under the banner of new charismatic breakthroughs.

The Holy Spirit as the ferment of the church

**Christian perfection.** The demand for perfection is frequently repeated in the New Testament and has played a significant role in the history of Christian spirituality. In the Gospel According to Matthew, Jesus directs the demand to his disciples: "You, therefore, must be perfect, as your heavenly Father is perfect" (chapter 5, verse 48). This demand seems to exceed by far the measure of reasonableness for man. Yet, it is meant literally, for it is asserted again in the writings of the New Testament. The meaning of this claim is recognizable only from the understanding of the human as the image of God and from the apprehension of Christ as the "new Adam." The perfection of believers is the perfection with which they reflect the image of God. They have, to be sure, disfigured this image through willful alienation from the original, but in Christ they recover the perfection of the image of God.

The idea of the deification of man, which captures the Greek notion of "partaking" of the divine character, also points in the direction of perfection. Post-Reformation theology, out of anxiety before "mysticism," struck almost entirely from its vocabulary this concept that originated in the techniques of the mystical experience. In the first one and a half millennia of the Christian Church, however, the idea of deification—of partaking in God's being—constituted a central concept for Christian anthropology. Athanasius created the fundamental formula for the theology of deification: "God became man in order that we become God." In the theology of the early church these words became the basis of theological anthropology. Only the idea of perfection makes understandable a final enhancement of the Christian image of the human—the intensification from "child of God" to "friend of God." This appears as the highest form of communion reached between God and human beings; in it love is elevated to the highest form of personal communication between prototype and image.

**Fellow humans as the present Christ.** That revolutionizing idea, which constitutes the basis of Christian ethics, also becomes comprehensible through the foundation of Christian anthropology in the image of God: in the eye of Christian faith Christ is present in everyone, even the most debased. According to Matthew (chapter 25, verses 40 and 45) the Judge of the world says to the redeemed: "Truly, I say to you, as you did it to one of the least of these my brethren, you did it to me," and to the damned: "As you did it not to one of the least of these, you did it not to me." Another saying of the Lord is cited by Tertullian: "If you have seen your brother, you have seen your Lord." For the Christian the fellow human is the present Christ himself. In the fellow, Christians see, under the wrapping of misery, degeneration, and suffering, the image of the present Lord, who became human, who suffered, died, and was resurrected in order to lead all humanity back into the Kingdom of God.

Christ as present in everyone

In the self-understanding of the Christian community two tendencies battle with one another from the beginning of church history. They lead to completely different consequences in the basic orientation of Christians toward fellow Christians and fellow human beings.

The one attitude stands under the governing idea of election. God chooses some out of the human race, which

exists in opposition to all that is divine, and erects a Kingdom from these elect. This idea underlines the aristocratic character of the Kingdom of God; it consists of an elite of elect. In the Johannine apocalypse the 144,000 " . . . who have not defiled themselves with women" (Revelation 14:4) constituted the picked troops of the Kingdom of God. For Augustine and his theological successors up to Calvin, the community of the elect is numerically restricted; their number corresponds to the number of fallen angels, who must again be replaced through the matching number of redeemed men so that the Kingdom of God would again be restored numerically as well. The church is here understood as a selection of a few out of the masses of perdition who constitute the jetsam of the history of salvation. A grave endangering of the consciousness of community is concealed in this orientation, for self-righteousness, which is the root of self-love and thereby the death of love of neighbour, easily enters again via this consciousness of exclusive election.

The other attitude proceeds from the opposite idea that the goal of the salvation inaugurated through Jesus Christ can only be redemption of all humanity. According to this view God's love of humans (*philanthrōpia*), as the drama of divine self-surrender for human salvation shows, is greater than the righteousness that craves the eternal damnation of the guilty. Since the time of Origen, this second attitude is found not only among the great mystics of the Eastern Church but also among some mystics of Western Christendom. The teaching of universal reconciliation (*apokatastasis pantōn*) has struck against opposition in all Christian confessions. This is connected with the fact that such a universalistic view easily leads to a disposition that regards redemption as a kind of natural process that no one can evade. Such an orientation can lead to a weakening or loss of a consciousness of moral responsibility before God and neighbour; it contains the temptation to spiritual security and moral indolence.

## THE CHURCH

The relationship of the Christian and Jewish concepts of the people of God

The Christian view of the church was influenced by the Old Testament concept of the *qahal,* the elected people of God of the end time, and by the expectation of the coming of the Messiah in Judaism. The Greek secular word *ekklēsia,* the term used for the church, means an assembly of people coming together for a meeting.

In Christianity the concept received a new meaning through its relationship to the person of Jesus Christ as the messianic inaugurator of the Kingdom of God: (1) with Christ the elected community of the end time has appeared; (2) the church is the eschatological gift of the Holy Spirit, which already flows through the life of the church (Acts 2:33); (3) the community of the end time consists of those who believe in Jesus Christ—both Jews and pagans; the idea of the elected convenant people (*i.e.,* the Jews) is transferred to the "new Israel"; (4) the church forms the body of its Lord; and (5) the church consists of "living stones," from which its house is "built" (1 Peter 2:5).

Jesus himself created no firm organization for his community; the expectation of the immediate imminence of the Kingdom of God provided no occasion for this. Nevertheless, the selection of Apostles and the special position of individual Apostles within this circle pointed to the beginnings of a structuralization of his community. After the community was constituted anew because of the impressions made by the appearances of the Resurrected One, the succession of the appearances apparently effected a certain gradation within the community.

The unity of the church, which was dispersed geographically, was understood from the viewpoint of the Diaspora (James 1:1—the scattered churches of the new Israel represent "the twelve tribes in the Dispersion"). The *Didachē,* or the *Teaching of the Twelve Apostles* (late 1st century), viewed the church in terms of the bread of the Eucharist, whose wheat grains "are gathered from the mountains." The idea of the preexistence of the divine Logos brought into existence the concept of the preexistence of the church, which included the view that the world was created for the sake of the church. The earthly church is thus the representative of the heavenly church.

**Normative defenses in the early church.** Establishment of norms for the church was necessary because diverse kinds of interpretations of the Christian message were conceived under the influence of the religions of late antiquity, especially Gnosticism—a syncretistic religious dualistic belief system that incorporated many Christian motifs and became one of the strongest heresies of the early church. In Gnostic interpretations, mixed Christian and pagan ideas appealed to divine inspiration or claimed to be revelations of the Resurrected One. The church erected three defenses against the apparently uncontrollable prophetic and visionary efficacy of pneumatic (spiritual) figures as well as against pagan syncretism, which was represented by a mixing together of many divine images and expressions: (1) the New Testament canon, (2) the apostolic "rules of faith," or "creeds," and (3) the apostolic succession of bishops. The common basis of these three defenses is the idea of "apostolicity."

Reasons for the development of the canon, creed, and episcopacy

The early church never forgot that it was the church that created, selected the books, and fixed the canon of the New Testament, especially because of the threat of Gnostic writings. This is one of the primary distinctions between the Orthodox Church vis-à-vis the Reformation churches, which view the Scriptures as the final norm and rule for the church and church teaching. The Orthodox Church, like the Roman Catholic Church, emphasizes the fact that the Christian Church existed prior to the formation of the canon of Scripture—that it is indeed the source and origin of the Scripture itself. Thus, tradition plays a significant role alongside the Holy Scriptures in the Orthodox and Roman churches.

The apostolic rule of faith—*i.e.,* the creed—issued from the apostolic tradition of the church as a second, shorter form of its solidification, at first oral and then written. It also served as a defense against Gnosticism and syncretistic heretical interpretations of the Christian faith.

The third defense that the church used against both Gnostic and syncretistic movements and free charismatic movements within the church was the office of bishop, which became legitimized through the concept of apostolic succession. The mandate for missions, the defense against free prophecy, the polemics with Gnosticism and other heresies, the persecution of the church, and, not least of all, management of church discipline—all allowed the monarchical episcopacy to emerge as a strong jurisdictional office in the early centuries. The bishop, in his capacity as leader of the eucharistic worship service, as teacher, and as curer of souls, became the chief shepherd of the church and was considered its representative.

The basic idea of apostolic succession is as follows: Christ appointed the original Apostles and entrusted to them his full spiritual authority; the original Apostles then appointed overseers (bishops) for the churches founded by them and passed on to them, through the sacramental laying on of hands, their authority of office. These men transmitted the office of overseer to their successors also by the laying on of hands. In this manner, apostolic succession guaranteed the legitimacy of episcopal church government, episcopal doctrine, and the validity of the sacraments dispensed by the bishops.

**Evolution of the episcopal office.** The evolution of the episcopal office followed a different development in the East and in the West. The Orthodox Church accepts the monarchical episcopacy insofar as it involves the entire church, both the visible earthly and the invisible heavenly churches bound together inseparably. The monarchical principle, however, finds no application to the organization of the visible church. The latter is based upon democratic principles that are grounded in the polity of the early church. Just as all Apostles without exception were of equal authority and none of them held a paramount position over against the others, so too their successors, the bishops, are of equal authority without exception.

Democratic principles in the Orthodox Church

Thus, the politics of the Eastern Orthodox churches have a decidedly synodal character. Not only the priesthood but also the laity have been able to participate in Orthodox synods. Election to ecclesiastical offices (*i.e.,* pastor, bishop, or patriarch) involves participation by both clergy and laity. The individual polities of modern Orthodox

churches (*e.g.,* Greek or Russian) are distinguished according to the amount of state participation in the settlement of ecclesiastical questions.

The ecumenical council, which consists of the assembly of all Orthodox bishops, constitutes the highest authority of Orthodox synodal polity. The bishops gathered at an ecumenical council resolve all questions of Orthodox faith as well as of worship and canon law according to the principle that the majority rules. The councils recognized by the Orthodox Church as ecumenical councils are: Nicaea in 325, Constantinople in 381, Ephesus in 431, Chalcedon in 451, second Council of Constantinople in 553, third Council of Constantinople in 680, and second Council of Nicaea in 787. No council since then has been regarded as ecumenical by Eastern Orthodoxy.

Orthodoxy was divided into various old and new types of churches. Some of these were "patriarchal," which meant that they were directly responsible to a patriarch. Others were "autocephalous," which has come to mean in the modern world that as national churches they are in communion with Constantinople but are responsible for authority to their own national synods. This division, plus the fact that Orthodoxy has so often been the victim of revolutionary change and political onslaught, has served as a hindrance against any new ecumenical council, even though many Orthodox have asked for such a council.

The causes of the monarchical episcopate in the West

On the basis of the joint action of special circumstances, in the Roman Church the papacy evolved out of the monarchical episcopate. Rome, as the capital of the Roman Empire, in which a numerically significant Christian community was already formed in the 1st century, occupied a special position. A leading role devolved upon the leading bishop of the Roman community in questions of discipline, doctrine, and ecclesiastical and worship order. This occurred in the Latin provinces of the church in the West (Italy, Gaul, Spain, Africa), whose organization followed the provincial organization of the Roman Empire. A special leadership position devolved upon the Roman bishop after the collapse of the Western Roman Empire. The theological underpinning of this special position was emphasized by Petrine theology, which saw in the words of Jesus, "You are Peter, and on this rock I will build my church" (Matthew 16:18), a spiritual-legal instituting of the papacy by Jesus Christ himself. In the Greek Church of the East (*e.g.,* Origen) and also in Augustine in the West, however, these words were referred to Peter's confession of faith; since the time of the popes Gelasius I (reigned 492–496), Symmachus (reigned 498–514), and Gregory I (reigned 590–604), these words have served as the foundation for the claim of papal primacy over the entire Christian Church.

The tendency of Christianity toward intolerance

**Authority and dissent.** Christianity, from its beginning, tended toward an intolerance that was rooted in its religious self-consciousness. Christianity understands itself as revelation of the divine truth that became human in Jesus Christ himself. "I am the way, and the truth, and the life; no one comes to the Father, but by me" (John 14:6). To be a Christian is to "follow the truth" (3 John); the Christian proclamation is "the way of truth" (2 Peter 2:2). Those who do not acknowledge the truth are enemies "of the cross of Christ" (Philippians 3:18) who have "exchanged the truth about God for a lie" (Romans 1:25) and made themselves the advocates and confederates of the "adversary, the devil," who "prowls around like a roaring lion" (1 Peter 5:8). Thus, one cannot make a deal with the devil and his party—and in this lies the basis for intolerance in Christianity.

Christianity consistently practiced an intolerant attitude in its approach to Judaism and paganism as well as heresy in its own ranks. By practicing its intolerance vis-à-vis the Roman emperor cult, it thereby forced the Roman state, for its part, into intolerance. Rome, however, was not adapted to the treatment of a religion that negated its religious foundations, and this inadequacy later influenced the breakdown of paganism.

Early Christianity aimed at the elimination of paganism—the destruction of its institutions, temples, tradition, and the order of life based upon it. After Christianity's victory over Greco-Roman religions, it left only the ruins of paganism still remaining. Christian missions of later centuries constantly aimed at the destruction of indigenous religions, including their cultic places and traditions (as in missions to the Anglo-Saxons, Germans, and Slavs). This objective was not realized in mission areas in which Christian political powers did not succeed in conquests—*e.g.,* China and Japan; but in Indian Goa, for example, the temples and customs of all indigenous religions were eliminated by the Portuguese conquerors.

The furtherance of intolerance in confrontations with Islām

The attitude of intolerance was further reinforced when Islām confronted Christianity from the 7th century on. Islām understood itself as the conclusion and fulfillment of the Old and New Testament revelation; from the Christian view, however, Islām was understood eschatologically—*i.e.,* as the religion of the "false prophets," or as the religion of the Antichrist. The aggression of Christianity against Islām—on the Iberian Peninsula, in Palestine, and in the entire eastern Mediterranean area during the Crusades—was carried out under this fundamental attitude of intolerance. Intolerance of indigenous religions was also manifested in Roman Catholic missions in the New World; these missions transferred the methods of the struggle against Islām to the treatment of the Native Americans throughout the Western Hemisphere and destroyed their cults and cultic places. Against Protestants, the Counter-Reformation displayed the same kind of intolerance and was largely equated with the struggle against the Turks.

The idea of tolerance first arose during a series of historical catastrophes that forced Christianity into self-reflection: the devastating impressions of the military proceedings of the Inquisition troops against the heretical Cathari, Albigenses, and Waldenses during the Middle Ages; the psychological effect of the permanent inquisitional terror; the conquest of Constantinople by the Turks; the fratricidal struggle among the churches that arose during the Reformation; and the battles of the Protestant territorial churches against the sectarian and Free Church groups in their midst.

Thus, for Nicholas of Cusa (1401–64) the conquest of Constantinople became the occasion to demand, for the first time, the mutual toleration of Christianity and Islām as the presupposition for a religious peace. When the Reformation churches asserted the exclusive claim of possessing the Christian truth, they tried to carry it out with the help of the political and military power at their disposal. In the religious wars of the 16th and 17th centuries, Christian intolerance developed into an internal fratricidal struggle in which each side sought to annihilate the other party in the name of truth. Only the fact that such attempts did not succeed led to new reflections upon the justification of one's own exclusive claim to absoluteness.

The intolerance of the Reformation territorial churches found its counterpart in the intolerance of the revolutionary groups of the Reformation period, such as that of the German radical Reformer Thomas Müntzer, which wanted to force the coming of the Kingdom of God through the dominion of the "elect" over the "godless." In the intolerance of the ideology and techniques of many modern political revolutions and authoritarian regimes some see either a legacy or a mimicking of old Christian patterns and methods (*e.g.,* inquisition or brainwashing).

Among those who first spoke up consistently for tolerance were the Baptists and Spiritualists of the Reformation period. Their most important contribution consisted in that they stood up for their constantly reiterated demand for tolerance not only through their preaching but also through their courageous suffering.

The victory of tolerance over intolerance

The victory of tolerance contributed especially to the recognition of the evident contradiction between the theological self-conception of Christianity as a religion of love of God and neighbour and the inhumanity practiced by the churches in the persecution of dissenters. Recognition of this contradiction even provoked criticism of the Christian truths of faith themselves.

The Roman Catholic Church in the past has consistently opposed the development of religious toleration. Its claim to absolute power in a state is still practiced in the 20th century in some Catholic countries, such as Spain and Colombia, in relationships to Protestant minorities. Since

Pope John XXIII and the second Vatican Council (1962–65), however, a more tolerant attitude of the Roman Catholic Church has been demanded that is appropriate both to the ecumenical situation of Christendom in the latter part of the 20th century and to the personal character of the Christian faith.

**Creeds and confessions.** The faith of Christendom is present in the confessions of faith and the creedal writings of the different churches. Three creeds find general ecumenical acknowledgment: the Apostles' Creed, the Nicene–Constantinopolitan Creed (also called the Nicene Creed), and the Athanasian Creed. The Apostles' Creed is the baptismal confession of the Roman Catholic community; its original form as a Greek hymn can be traced back to the apostolic tradition (of the 2nd century). The Nicene–Constantinopolitan Creed is the confession of faith of the ecumenical Council of Nicaea in 325, which was later supplemented at the ecumenical Council of Constantinople in 381. Its principal use is in the liturgy of the Eucharist. The Athanasian Creed is a Latin creed whose theological content can be traced back to Athanasius of Alexandria (4th century) but that probably first originated in the 5th century in Spain or southern Gaul. It contains a detailed formulation of the doctrine of the Trinity and Christology (the two-natures doctrine), which was influenced by Augustine. All three creeds were accepted by the churches of the Reformation.

Acceptance of the three ecumenical creeds

Around central confessional statements about Jesus as the Christ in the New Testament—*e.g.,* "Jesus is Lord" (Romans 10:9); "You are the Christ" (Matthew 16:16)—are concentrated a series of further assertions that laud his significance for salvation and concern his suffering, death by crucifixion, Resurrection, and his exaltation to God. This tradition, through Mark, Luke, and Paul, was called "gospel," or kerygma (proclamation).

The original form of the creed possessed not a didactic but a hymnal character and had its locus in the worship service. Regular use of a creed as a baptismal confession, and, accordingly, in the preparation of candidates for baptism in catechetical instruction, influenced its fixed formulation. This was also true of its use in the eucharistic worship service as an expression of the congregation's unity in faith before receiving the elements of the Lord's Supper as well as its use as testimony before the world in times of persecution and as norm of faith (*regula fidei*) in the altercation with heresies.

Development of confessions

Development of confessions of faith into theological didactic creeds, which began during the Christological controversies of the 5th century, was continued in the Reformation. The relatively short creedal formulas grew into extensive creedal compositions, primarily because the Reformers conducted their battles with the Roman Church as a struggle for "pure doctrine" as well as for a foundation for the unity of the church. In the Diet of Augsburg in 1530 the feuding ecclesiastical parties were induced to deliver a presentation of their faith. Though the Roman Catholics did not accede to this challenge, the Protestants offered the Confessio Augustana (or the Augsburg Confession). First planned by Philipp Melanchthon, a follower of Luther, as a creed for union, it later became the basic confessional statement of the Lutheran Church.

The formation of various Protestant confessions was achieved in the individual territorial churches and led to the development of diverse *corpora doctrinae* ("bodies of doctrines"). The differences of the traditional creeds and adherence to them are still clearly noticeable in the ecumenical movement of the 20th century.

A similar development of doctrinal confessions occurred in Calvinism. The idea of the completion of confessional writings is missing in the Lutheran churches but not in Calvinistic churches: the revision of old and the formation of new creedal writings are permitted and in part are provided for in the rules of the church. Thus the Barmen Declaration in 1934, against the "German Christians" and the Nazi worldview, arose primarily from Reformed circles. The Anglican Church incorporated the Thirty-nine Articles (a confessional statement) and a short catechism into *The Book of Common Prayer* of 1559/1662 (revised in the United States by the Protestant Episcopal Church in 1928 and 1979) and thereby emphasized the unity of doctrine and worship.

Of the denominations that arose out of the Reformation churches, most created doctrinal documents that are comparable to the reformational confessional writings (*e.g.,* among Methodists, Baptists, and Congregationalists). Some denominations (*e.g.,* the Quakers, the Disciples of Christ, and some Baptists), on the other hand, have rejected any form of creed because they believe creeds to be obstacles to the Christian faith, thus conflicting with the freedom of the Holy Spirit.

The shifting of the chief emphasis in church life to "pure doctrine" in the 16th and 17th centuries also obliged the Orthodox and Roman Catholic churches to formulate their teaching in confessional texts. Thus, under the influence of the reformational creedal writings, the Eastern Orthodox Church developed confessional texts. An example is *The Orthodox Confession of Faith* (*Confessio orthodoxa*) of the metropolitan Peter Mogila of Kiev against Cyril Lucaris, a Calvinist-influenced patriarch of Constantinople; it was approved in 1643 by the Greek and Russian patriarchs. At the Council of Trent (1545–63) the Roman Catholic Church countered the Protestant doctrinal creeds with a *Professio fidei Tridentina* ("The Tridentine Profession of Faith"), which at the end of every article of faith respectively anathematizes the dissenting Protestant article of faith.

Creedal formulations in the 20th century

In modern Christendom, creedal formulation is continued in two areas. (1) Within the ecumenical movement, since the formation of the World Council of Churches in 1948 there have been attempts to create a brief uniform confession as the common basis of faith for the Christians in that council. These efforts have not yet been concluded. According to its constitution, the World Council of Churches is "a fellowship of Churches which accepts our Lord Jesus Christ as God and Saviour." In 1960 at St. Andrews, Scot., the World Council's central committee unanimously accepted an expanded draft of the "basis":

> The World Council of Churches is a community of churches which confess the Lord Jesus Christ, according to the Holy Scriptures, as God and Savior and therefore seek to fulfill that to which they are jointly called, to the glory of God the Father, the Son and the Holy Spirit.

This new version ensued mainly at the instigation of the Orthodox churches, for whom the hitherto existing form of the "basis" was not adequate.

The movement of Roman Catholicism into the interconfessional orbit after the second Vatican Council complicated attempts to draft a modern ecumenical confession. The Roman Catholic Church is not a member of the World Council, but conciliar Protestant and Orthodox members are reluctant to make major moves without considering Roman Catholic interests.

(2) There are great numbers of churches—the majority, many would contend—that are products of missionary endeavours by the West. For a time they were called "the younger churches" but are now more frequently referred to simply as Asian or African churches, or churches in developing nations. Among them the doctrinal disputes and confessional battles of Western Christendom have often been viewed as alien, imported, and frequently incomprehensible. The union of churches in South India into the Church of South India (1947) occurred only on the basis of the participating churches dismantling their traditional creedal differences. The Church of South India's scheme of union substitutes biblical revelation for doctrinal formulation. Similarly, the United Church of Christ in Japan (Kyodan) renounced drawing up a new creed and limited itself to a preface to the Apostles' Creed. In the churches of Africa, the inadequacy of the confessions of the 16th century also has been strongly recognized as a result of their own indigenous cultural presuppositions.

Areas in which early Christians sometimes violated ideals of holiness

**Organization.** In the early church, discipline—qualified by the ideal of holiness demanded from baptized Christians—concerned four areas in which there arose violations of the demand for holiness: (1) the relationship to the pagan social milieu and the forms of life and culture connected with it (*e.g.,* idolatry, the emperor's cult, the theatre, and the circus); (2) the relationship of the

sexes within the Christian community (*e.g.,* rejection of polygamy, prostitution, pederasty, sodomy, and obscene literature and art); (3) other offenses against the community, especially murder and property crimes of all kinds; and (4) the relationship to teachers of false doctrine, false prophets, and heretics.

Employment of church discipline at an early date led to the formation of a casuistry that at first consisted simply of the distinction between "mortal" and "not mortal" sins (1 John 5:15 ff.)—*i.e.,* between sins that through their gravity resulted in loss of eternal life and those with which this was not the case. In earliest Christianity, the relapse of a baptized Christian into paganism (*i.e.,* apostasy) was believed to be the most serious offense. In the Letter to the Hebrews one who is baptized irrevocably forfeits salvation through a relapse into grievous sin. The various difficulties in substantiating the theory and practice of a second repentance were solved by Pope Calixtus (reigned 217/218–222). This question was especially important in Rome because of the great number of offenses against the idea of holiness. Pope Calixtus granted to bishops decisions about definitive exclusion from the congregation or readmission as well as the evaluation of church punishments. Among all the factors that led to the power of the episcopacy, the concentration of penitential discipline in the hands of the bishop probably contributed more to the strengthening of episcopal power and to the achievement of the monarchical episcopate in the church than any other single factor. This development did not take place without fierce opposition (*e.g.,* Montanism).

Attainment of the church's demand of holiness was made more difficult in the large cities, especially in reference to sexual purity. The period of persecution by the pagan emperors and the legal constraint to performance of sacrifice before the altars of the emperor's images brought countless new instances of apostasy. The so-called Lapsi (Lapsedones), who had performed sacrifices before the emperor's image but, after persecution, faded away and then moved back into the churches again, became a serious problem for the church, sometimes causing schisms (*e.g.,* the Donatists).

The execution of church discipline by the clergy was subordinated to the regulations of canon law provided for priests. A genuine practice of church discipline was maintained in the monasteries in connection with the public confession of guilt, which was made by every monk before the entire assembly in the weekly gatherings of the chapter. A strong revival of church discipline among the laity also resulted from the church discipline pursued within monasticism.

Results of the externalization of church discipline

On the whole, the casuistic regulation of church discipline led to its externalization and devaluation. The medieval sects, therefore, always stressed in their critique of the worldly church the lack of spiritual discipline and endeavoured to realize a voluntary church discipline in terms of a renewed radical demand of holiness based on early Christianity. The radical sects that emerged in the Reformation reproached the territorial churches by claiming that they had restricted themselves to a renovation of doctrine and not to a renewal of the Christian life and a restoration of the "communion of saints." Different groups of Anabaptists (*e.g.,* Swiss Brethren, Mennonites, and Hutterites), especially, attempted to realize the ideal of the purity and holiness of the church through the reintroduction of a strict church discipline.

The Reformed churches in particular endeavoured to make church discipline a valid concern of the community. In Geneva, church discipline was expressed, at the instigation of Calvin, in the establishment of special overseers, who, in the individual districts assigned to them, had to watch over the moral behaviour of church members. There likewise came about the creation of such social arrangements as ecclesiastically controlled inns and taverns, in which not only the consumption of food and drink but even the topics of conversation were subject to stern regulation. The cooperation of ecclesiastical discipline and state legislation found its characteristic expression in the United States in the Prohibition amendment to the Constitution. Its introduction came most strongly from congregational churches, above all those characterized by Evangelical, Fundamentalist, or Pentecostal outlooks. They united forces with more moderate or liberal churches that were experienced in trying to affect the social order through legislation. Together they battled against the misuse of alcohol as part of their ideal to extend Christian norms and influence to the whole of society.

In the 20th century, church discipline, in the original spiritual sense of voluntary self-control, is practiced only in smaller communities of evangelical Christians, in which the ideal of holiness of the community is still maintained and in which the mutual, personal bond of the congregational members in the spirit of Christian fellowship still allows a meaningful realization of a church discipline. It is also practiced in churches in developing nations. In these churches the practice of church discipline still appears as a vitally necessary centre of the credible self-representation of the Christian community. Characteristically, therefore, these churches' main criticism of the old institutional churches has been directed against the cessation of church discipline among their members.

The development of the episcopacy in the Orthodox and Roman Catholic churches has been covered in the general introduction of this section under evolution of the episcopal office. Occupying a special position is the episcopal polity of the Anglican Communion. Despite the embittered opposition of Puritan and independent groups during the period of the Reformation and Revolution in England, this polity has maintained the theory and practice of the episcopal office of apostolic succession. The Low Church tradition of the Anglican Communion views the episcopal office as a form of ecclesiastical polity that has been tested through the centuries and is therefore commendable for pragmatic reasons; the Broad Church tradition, however, emphatically adheres to the traditional worth of the episcopal office without allowing the faithful to be excessively dependent upon its acknowledgement. The High Church tradition, on the other hand, values episcopal polity as an essential element of the Christian Church that belongs to the church's statements of faith. The episcopal branch of the Methodist Church has also retained in its polity the bishop's office in the sense of the Low Church and Broad Church view.

Episcopacy in Anglican and other Reformation churches

In the Reformation churches an episcopal tradition has been maintained in the Swedish state church (Lutheran), whose Reformation was introduced through a resolution of the imperial Diet of Västerås in 1527, with the cooperation of the Swedish bishops. In the German Evangelical (Lutheran and Reformed) territories, the bishops' line of apostolic succession was ruptured by the Reformation. As imperial princes, the Roman Catholic German bishops of the 16th century were rulers of their territories; they did not join the Reformation in order to avoid renouncing the exercise of their sovereign (temporal) rights as demanded by Luther's Reformation. On the basis of a legal construction originally intended as a right of emergency, the Evangelical rulers functioned as the bishops of their territorial churches but only in questions concerning external church order. This development was promoted through the older conception of the divine right of kings and princes, which was especially operative in Germanic lands.

In matters of church polity, controversial tendencies that began in the Reformation still work as divisive forces within the ecumenical movement in the 20th century. For Luther and Lutheranism, the polity of the church has no divine-legal characteristics; it is of subordinate significance for the essence of the church, falls under human ordinances, and is therefore quite alterable. In Calvinism, on the other hand (*e.g.,* in the *Ecclesiastical Ordinances* [*Ordonnances ecclésiastiques*] of 1541 and in Calvin's *Institutes of the Christian Religion* [1536]), the Holy Scriptures appear as a codex from which the polity of the congregation can be inferred or certainly derived as a divine law. Thus, on the basis of its spiritual–legal character, church polity would be a component of the essence of the church itself. Both tendencies stand in a constant inner tension with one another in the main branches of the Reformation and within the individual confessions as well.

Even in Lutheranism, however, there has been a demand

for a stronger emphasis upon the independent episcopal character of the superintendent's or president's office. Paradoxically, in the Lutheran Church, which came forth with the demand of the universal priesthood of believers, there arose the development of ecclesiastical authorities but not the development of self-contained congregational polities. When a merger of three Lutheran bodies produced a new Evangelical Lutheran Church in America in 1988 it established the bishop as leader of the synodal jurisdictions. In Lutheranism these bishops replaced presidents. Bishops were regarded there, as in Methodism, as part of the *bene esse,* the well-being, and not the *esse,* the essence, of the church. More or less self-contained congregational polities were developed in many Reformed churches because the Reformed Church congregation granted greater participation in the life of the congregation to the laity as presbyters and elders. Furthermore, the Reformed Church areas in Germany, France, England, and Scotland, as well as in The Netherlands and Hungary, had to build up their own ecclesiastical structure without dependence upon state authorities.

Among the conservative but often spontaneous evangelical Protestant churches diverse forms of polity have developed. They have all been founded with an appeal to the Holy Scriptures. Their prototypes can, in fact, be identified in the multiformity of congregational polities in the first three centuries before the victory of the monarchical episcopal office.

The model of the original church

Presbyterian polity appeals to the model of the original church. The polity of the Scottish Presbyterian Church and the Presbyterian churches of North America is primarily based upon this appeal, which was also found among many English Puritan groups. It proceeds from the basic view that the absolute power of Christ in his church postulates the equality of rights of all members and can find expression only in a single office, that of the presbyter. The calling to this office is through election by church members, formally analogous to the democratic, republican political mode, and, accordingly, in contrast with the monarchy of the papal and the aristocracy of the episcopal church polity. In Presbyterian churches the differences between clergy and laity have been abolished in theory and, to a great extent, in practice. A superstructure of consistories and presbyteries is superposed one upon the other, with increasing disciplinary power and graduated possibilities of appeal. Through their emphases upon the divine–legal character of Presbyterian polity, the Presbyterian churches have represented a Protestant polity that counters the Roman Catholic concept of the church in the area of ecclesiastical polity. In ecumenical discussions in the 20th century, the divine–legal character of this polity is occasionally noticeable in its thesis of an apostolic succession of presbyters as a counter-thesis to that of the apostolic succession of bishops.

Congregationalism stresses the autonomous right of the individual congregation to order its own life in the areas of teaching, worship, polity, and administration. This demand had been raised and practiced by the medieval sects and led to differentiated polities and congregational orders among the Waldenses, the Hussites, and the Bohemian Brethren. Congregationalism was advanced in the Reformation period by the most diverse parties in a renewed and reinforced way not only by "Enthusiasts" (or, in German, Schwärmer) and Anabaptists, who claimed for themselves the right to shape their congregational life according to the model of the original church, but also by individual representatives of Reformation sovereigns, such as Franz Lambert (François Lambert d'Avignon), whose resolutions at the Homberg Synod of 1526 were not carried out because of a veto by Luther. The beginnings of modern Congregationalism probably lie among the English refugee communities on the European mainland, in which the principle of the established church was first replaced by the concept of a covenant sealed between God or Jesus Christ and the individual or the individual congregation.

The basic concepts of Congregationalism

The basic concepts of Congregationalism are: the understanding of the congregation as the "holy people" under the regent Jesus Christ; the spiritual priesthood, kingship, and prophethood of every believer and the exchange of spiritual experiences between them, as well as the introduction of a strict church discipline exercised by the congregation itself; the equal rank of all clergy; the freedom of proclamation of the gospel from every episcopal or official permission; and performance of the sacraments according to the institution of Jesus. By virtue of the freedom of self-determination fundamentally granted every congregation, no dogmatic or constitutional union but rather only county union of the Congregationalist churches developed in England. North America, however, became the classic land of Congregationalism as a result of the great Puritan immigration to New England, beginning with the Pilgrims on the *Mayflower* (1620). In the 20th century, acknowledgement of the full authority of the individual congregation runs through almost all Protestant denominations in the United States and is even found among the Lutherans. Congregationalism participates in the ecumenical movement, within which it presses for awakening the independent activity of the Christian churches in the entire world in terms of a proto-Christian ideal of the congregation.

Numerous other forms of congregational polity have arisen in the history of Christendom, such as the association idea in the Society of Friends. Even Pentecostal communities have not been able to maintain themselves in a state of unrestrained and constant charismatic impulses but instead have had to develop a legally regulated polity. This was what happened in the early church, which likewise was compelled to restrain the freedom of charisma in a system of rulers and laws. Pentecostal communities either have been constituted in the area of a biblical fundamentalism theologically and on the basis of a congregationalist church polity constitutionally or they have ritualized the outpouring of the Spirit itself. Thus, the characteristic dialectic of the Holy Spirit is confirmed: the Spirit creates law and the Spirit breaks law even in the most recent manifestations of its working.

**Liturgy.** The central focus of the liturgy of the early church was the Eucharist, which the Christian community interpreted as a fellowship meal with the resurrected Christ. Judaism at the time of Christ was dominated by an intense expectation of the Kingdom of God, which would be inaugurated by the Messiah–Son of man. The early Christian Church appropriated this expectation, which revolved around the image of the messianic meal in which the faithful would "sit at table" (Luke 13:29) with the coming Messiah–Son of man. At the centre of Jesus' preaching on the Kingdom of God is the promise that the blessed would "eat bread" with the exalted Messiah–Son of man (Luke 13:29). The Lord himself would serve the chosen community of the Kingdom at the messianic meal (Luke 12:37 ff.), which bears the features of a wedding banquet. The basic mood in the community gathered about him is thus one of nuptial joy over the inauguration of the promised end time, which Jesus emphasized in Matthew, chapter 9, verse 15: "Can the wedding guests mourn as long as the bridegroom is with them?" The supper that Jesus celebrated with his disciples "on the night when he was betrayed" (1 Corinthians 11:23) inaugurated the heavenly meal that will be continued in the Kingdom of God. Decisive for understanding the original meaning of the Eucharist are the words of Jesus in Matthew, chapter 26, verse 29: "I shall not drink again of this fruit of the vine until that day when I drink it new with you in my Father's kingdom."

The image of the messianic meal

The death of Jesus at first bewildered his community in the face of his promise, but the appearances of the Resurrected One, beginning with Easter morning, confirmed their expectations about the messianic Kingdom. These appearances influenced the expectations about the messianic meal and the continuation of fellowship with the exalted Son of man in the meal. Faith in the Resurrection and an expectation of the continuation of the fellowship meal with the exalted Son of man are two basic elements of the Eucharist that are a part of the liturgy from the beginnings of the church. In meeting the Resurrected One in the eucharistic meal the community sees all the glowing expectations of salvation confirmed.

The basic mood of the Eucharist

The basic mood of the community at the eucharistic meal is thereby one of joy. "And breaking bread in their homes, they partook of food with glad and generous hearts, praising God" (Acts 2:46). The Orthodox liturgy has maintained this original Christian mood of joy as at a wedding feast until the present. In Reformation churches, however, a mood of repentance and sorrow over sin often diminished and suppressed the original Christian attitude of joy.

What the Christian community experiences in the eucharistic meal is basically a continuation of the appearances of the Resurrected One in its midst. Thus, many liturgical forms developed, all of which served to enhance the mystery of the eucharistic meal. In the magnificent liturgical creations from the 1st to the 6th century, diversity rather than uniformity was a commanding feature of the development of worship forms. The eucharistic mystery developed from a simple form, as depicted in the 1st-century *Didachē*, to the fully developed liturgies of the 5th and 6th centuries in both the East and the West.

This diversity that was demonstrated in the liturgies of the early church is still preserved in the Clementine liturgy (Antioch), the Syrian liturgy, the Liturgy of St. James of the church of Jerusalem, the Nestorian liturgy in Iran, the Liturgy of St. Mark in Egypt, the Roman mass, the Gallic liturgies, and the Ambrosian (Milanese), Mozarabic (Spanish), and Scottish-Irish (Celtic) liturgies.

In the 6th century two types of liturgies were fixed by canon law in the Eastern Orthodox Church: the Liturgy of St. John Chrysostom (originally the liturgy of Constantinople) and the Liturgy of St. Basil (originally the liturgy of the Cappadocian monasteries). The Liturgy of St. Basil, however, is celebrated only 10 times during the year, whereas the Liturgy of St. John Chrysostom is celebrated most other times. In addition to these liturgies is the so-called Liturgy of the Preconsecrated Offerings, attributed to Pope Gregory I the Great of the 6th century. In this liturgy no consecration of the eucharistic offering occurs—because the eucharistic offerings used have been consecrated on the previous Sunday—and it is celebrated on weekday mornings during Lent as well as from Monday to Wednesday during Holy Week.

The period of liturgical improvisation apparently was concluded earlier in the Latin West than in the East. The liturgy of the ancient Latin Church is textually available only since the 6th century. Though the Gallic liturgies are essentially closer to the Eastern liturgies, the liturgy of Rome followed a special development. From the middle of the 4th century, the Roman mass was celebrated in Latin rather than in Greek, which had been the earlier practice. The fixing of the Roman mass by canon law is congruent with the historical impulse of the Roman Catholic Church to follow the ancient Roman pattern of rendering sacred observance in legal forms and with stipulated regularities.

The tendency of liturgies to become fixed in form

Because of the authority inhering in the sacred, every liturgy has the tendency to become fixed in form, and any alteration of the liturgy can thus be regarded as a sacrilege. The spiritual–legal fixation of the liturgy, however, through the process of constant repetition and habit, led to an externalization that can transfer the liturgy into a lifeless formalism for both the liturgist and the participating community.

Characteristically, all reformation eras in the history of Christianity, in which new charismatic impulses arise in the areas of piety and theology, are also periods of new liturgical creations. Thus in the late 16th-century Reformation a great diversity of new liturgical forms emerged. Luther in Germany restricted himself to a reformatory alteration of the Roman Catholic liturgy of the mass, whereas Zwingli in Switzerland attempted to create a completely new evangelical liturgy of the Eucharist based upon a New Testament foundation. The Free churches also showed a strong liturgical productivity; in the Herrnhut Brethren (Moravian) community, Graf von Zinzendorf ushered in the singing worship services. Methodism, influenced by the Moravian spiritual songs and melodies, also produced new liturgical impulses, especially through its creation of new hymns and songs and its joyousness in singing.

The innovative religious bodies, especially those that arose in the 19th and 20th centuries, have been especially productive in this area. The Mormons, for example, developed not only a new type of church song but also a new style of church music in the context of their liturgical new creation (*e.g.*, "sealing"). The mood of charismatic, liturgical new creations has also been preserved in the Baptist churches of American blacks, whose spirituals are the most impressive sign of a free and spontaneous liturgy. The Pentecostal churches of the 20th century quite consciously attempt to protect themselves against liturgical formalism. The free, often spontaneously improvised liturgy of the Pentecostal tent missions was transformed into patterns that became familiar to a wider audience through televised evangelism, which was often of a Pentecostal nature. Often ecstatic, strongly rhythmized music endeavours to retain certain features of the charismatic spontaneity of the early Christian worship.

Traditional liturgy fixed by canon law, which could develop into a lifeless formalism, occasionally led to the adoption of a fundamentally anti-liturgical attitude. Zwingli's reformation, for example, exhibited an emphatically anti-liturgical tendency in that it reduced the intricate Roman Catholic order of service to beginning song, prayer, sermon, concluding prayer, and concluding song. In many Reformed churches, some anti-liturgical currents developed, which, in terms of visual art, have been directed against encouraging expressions that might distract from the preached and prayed Word. In more radical instances this has even meant protests against the use of the organ in the worship service. The Society of Friends radically eliminated the liturgy and replaced it with mutual silence, expecting the spontaneous activity of the Holy Spirit.

Though definite and obligatory liturgies have been established as normative, the forms of the liturgy continue to develop and change. The impulse toward variations in worship services has been especially noticeable in the latter part of the 20th century. In the Eastern Orthodox liturgy, in the Roman Catholic mass and breviary, and in Anglican and Lutheran liturgies, there are both fixed and changing sections. The fixed parts represent the basic structure of the worship service concerned, and the alternating parts emphasize the individual character of a particular service for a certain day or period of the church year. The changing parts consist of special Old and New Testament readings that are appropriate for a particular church festival, as well as of special prayers and particular hymns.

The basic parts of the eucharistic liturgy

The eucharistic liturgy consists of two parts: the Liturgy of the Catechumens and the Liturgy of the Faithful. This basic liturgical structure goes back to a time in which the church was a missionary church that grew for the most part through conversion of adults. The latter were first introduced to the Christian mysteries as catechumens through instruction in religious doctrine. They also received permission to take part in the first part of the worship service (which was instructional), but they had to leave the service before the eucharistic mystery was celebrated. The first part of the Orthodox worship service still ends with a threefold exclamation, reminiscent of pre-Christian, Hellenistic mystery formulas: "You catechumens, go forth! None of the catechumens (may remain here)!"

The eucharistic liturgy of the Orthodox Church is a kind of mystery drama in which the advent of the Lord is mystically consummated and the entire history of salvation—the incarnation, death, and Resurrection of Christ the Logos, up to the outpouring of the Holy Sprit—is recapitulated. The Orthodox Church also attaches the greatest value to the fact that within the eucharistic mystery an actual transformation of the eucharistic elements in bread and wine takes place. This is not the same as the Roman Catholic dogma of transubstantiation, which teaches that the substance of the bread and wine is changed into the body and blood of Christ, though the properties of the elements remain the same, when the priest consecrates the bread and wine. According to some Orthodox authorities, the Orthodox view is similar to the Lutheran doctrine of the Real Presence. The essential and central happening in the Orthodox liturgy, however, is the descent of the resurrected Lord himself, who enters the community as "the

King of the universe, borne along invisibly above spears by the angelic hosts." The transformation of the elements is, therefore, the immediate emanation of this personal presence. Thus, the Orthodox Church does not preserve and display the consecrated host after and outside the eucharistic liturgy, as in the Roman Catholic Church, because the consecrated offerings are mystically apprehended and actualized only during the eucharistic meal.

In the Roman Catholic mass, the sacrificial character of the Eucharist is strongly emphasized, but it is less so in the Orthodox liturgy. This is because in the Orthodox liturgy the Eucharist is not only a representation of the crucifixion sacrifice (as in the Roman mass) but also of the entire history of salvation, in which the entire congregation, priest and laity, participates. Thus, the Orthodox Church has also held fast to the original form of Holy Communion in both kinds.

The Orthodox Church still preserves the liturgical gestures of the early church. Though in many Protestant churches parishioners sit while praying, the Orthodox worshiper prays while standing (because he stands throughout the service), with arms hanging down, crossing himself at the beginning and ending of the prayer.

Forms of liturgical gestures

The prayerful gesture of folded hands among Protestant churches derives from an old Germanic tradition of holding the sword hand with the left hand, which symbolizes one's giving himself over to the protection of God because he is now defenseless. The prayerful gesture of hands pressed flat against one another with the fingertips pointed upward—the symbol of the flame—is practiced among Roman Catholics as well as Hindus and Buddhists. Other liturgical gestures found in many Christian churches are crossing oneself, genuflecting, beating oneself on the chest, and kneeling during prayer or when receiving the eucharistic elements. Among some Holiness or Pentecostal churches spontaneous handclapping and rhythmic movements of the body have been stylized gestures in the worship services. These gestures are often familiar features of worship in churches in Africa, Asia, and Latin America. Liturgical dancing, widely spread in pagan cults, was not practiced in the early church; vestigial remnants of this ancient practice, however, have been admitted in liturgical processionals. In the latter part of the 20th century, liturgical dances have been reintroduced in some churches but only in a limited fashion. Among the many other gestures of devotion and veneration practiced in the liturgically oriented churches such as the Roman Catholic Church, the High Church Anglican churches, and the Orthodox Church, are kissing the altar, the gospel, the cross, and the holy icons.

Liturgical vestments have developed in a variety of fashions, some of which have become very ornate. The liturgical vestments all have symbolic meaning (see below *Church year: Liturgical colours*). In the Orthodox Church the liturgical vestments symbolize the wedding garments that enable the liturgists to share in the heavenly wedding feast, the Eucharist. The *epitrachēlion,* which is worn around the neck and corresponds to the Roman stole, represents the flowing downward of the Holy Spirit (see also RITES AND CEREMONIES, SACRED).

**The sacraments.** The interpretation and number of the sacraments vary among the Roman Catholic, Orthodox, Eastern independent, and Protestant churches. The Roman Church has fixed the number of sacraments at seven: baptism, confirmation, the Eucharist, penance, holy orders, matrimony, and anointing of the sick. In the early church the number of sacraments varied, sometimes including as many as 10 or 12. The theology of the Orthodox Church, under the influence of the Roman Catholic Church, fixed the number of sacraments at seven. The classical Protestant churches (*i.e.,* Lutheran, Anglican, and Reformed) have accepted only two sacraments—*i.e.,* baptism and the Eucharist, though Luther allowed that penance was a valid part of sacramental theology.

The number of sacraments and forms of sacramentals

The New Testament mentions a series of "holy acts" that are not, strictly speaking, sacraments. Though the Roman Catholic Church recognizes a difference between such "holy acts," which are called sacramentals, and sacraments, the Orthodox Church does not, in principle, make such strict distinctions. Thus, though baptism and the Eucharist have been established as sacraments of the church, foot washing, which in the Gospel According to John, chapter 13, replaces the Lord's Supper, was not maintained as a sacrament. It is still practiced on special occasions, such as on Holy Thursday in the Roman Catholic Church and as a rite prior to the observance of the Lord's Supper, as in the Church of the Brethren. The "holy acts" of the Orthodox Church are symbolically connected to its most important mysteries. Hence, baptism consists of a triple immersion that is connected with a triple renunciation of Satan that the candidates say and act out symbolically prior to the immersions. Candidates first face west, which is the symbolic direction of the Antichrist, spit three times to symbolize their renunciation of Satan, and then face east, the symbolic direction of Christ, the sun of righteousness. Immediately following baptism, chrismation (anointing with consecrated oil) takes place, and the baptized believers receive the "seal of the gift of the Holy Spirit."

**Tradition.** The disposition of Christianity toward tradition has exhibited a characteristic tension from its very beginnings; it has broken tradition and it has created tradition. This tension, which is grounded in its essence, has been continued throughout its entire history. It began with breaking the tradition of piety described and prescribed in the Hebrew Scriptures and synagogue practices, which to the followers of Jesus looked legalistic. In the Sermon on the Mount, Jesus set forth his message as a renunciation of the Old Testament tradition of the Law. Yet, with his coming, new revelation, life, death, and Resurrection, he himself created a new tradition, a "new law," that has been carried on in the church. The dogmatic controversies of the Reformation period give the impression that the tradition of the church has to do primarily, if not exclusively, with ecclesiastical doctrinal tradition. Tradition, however, includes all areas of life of the Christian community and its piety, not just the teachings but also the forms of worship service, bodily gestures of prayer and the liturgy, oral and written tradition and the characteristic process of transition of the oral into written tradition, a new church tradition of rules for eating and fasting, and other aspects of the Christian life.

The inherent tension regarding tradition in Christianity

The break with the tradition of Jewish legal piety was not total. The Old Testament was adopted from Jewish tradition, but its interpretation was based upon the concepts of salvation that emerged around the figure of Jesus Christ. The Old Testament book of Psalms, including its musical form, was taken over in Christian worship as the foundation of the liturgy. The new revelation became tradition in the oral transmission of the words of the Lord (the logia) and the reports (kerygma) concerning the events of his life that were important for the early church's faith in him; his baptism, the story of his Passion, his Resurrection, and his Ascension. The celebration of the Lord's Supper as anticipation of the heavenly meal with the Messiah–Son of man in the coming Kingdom of God, even to the point of preserving in the liturgy the Aramaic exclamation *maranatha* ("O Lord, Come") and its Greek parallel *erche kyrie* ("Come, Lord!") as the supplicant calling for the Parousia (Second Coming)—all this became tradition.

In addition to the traditions of the Old Testament synagogal worship service, traditions of the Hellenistic mystery cults also were absorbed and reinterpreted in Christian forms. Among the traditions taken over from the mystery religions were: the arcane discipline—the distinction between the true *mystae* (those initiated into the secrets of the Christian faith), who were permitted to participate in the esoteric worship service (*i.e.,* the Eucharist), and the catechumens; the introduction of hymn singing dependent upon the melodic style of the mystery hymns (in addition to the Jewish Psalms); the retention of the ancient gesture of upraised hands during the epiclesis, the prayer that calls down the Holy Spirit upon bread and wine as they are consecrated in the sacrament; and many others.

Oral and written tradition

Of special significance is the oral tradition of doctrinal transmission and its written record. Judaism over the centuries had developed its own unique tradition of the oral transmission of teachings. According to rabbinic doctrine,

orally transmitted tradition coexisted on an equal basis with the recorded Law. Both text and tradition were believed to have been entrusted to Moses on Mount Sinai. Within the unbroken chain of scribes the tradition was passed on from generation to generation and substantiated through scripture and exegesis. The doctrinal contents of the tradition were initially passed on orally and memorized by the students through repetition. Because of the possibilities of error in a purely oral transmission, however, the extensive and growing body of tradition was, by necessity, fixed in written form. The rabbinic tradition of the Pharisees (a Jewish sect that sanctioned the reinterpretation of the Mosaic Law) was established in the Mishnah (commentaries) and later in the Palestinian and Babylonian Talmud (compendiums of Jewish Law, lore, and commentary). Because the essence of tradition is never concluded—*i.e.,* by its very nature is never completely fixed in writing—the learned discussion of tradition by necessity continued in constant exegetical debate with the Holy Scriptures. The written record of tradition, however, never claimed to be equal to the Holy Scriptures in Judaism. A similar process of written fixation also occurred among the sectarians of the community at Qumrān, which in its *Manual of Discipline* and in the *Damascus Document* recorded its interpretation of the Law, developed first orally in the tradition.

In the Christian Church a tradition also was formed proceeding from Jesus himself. The oral doctrinal transmission of the tradition was written down between the end of the 1st and the first half of the 2nd century in the form of various gospels, histories of the Apostles, letters, sermonic literature, and apocalypses. Among Christian Gnostics the tradition also included secret communications of the risen Christ to his disciples.

A new element, however, inhered in the Christian vis-à-vis the Jewish tradition. For Jewish piety the divine revelation encompassed two forms of divine expression: the Law and the Prophets. Nevertheless, this revelation is considered concluded with the last Prophets; its actualization further ensues through interpretation. In the Christian Church the tradition is joined not only to the teachings of Jesus and the story of his life as prophet and teacher that terminated with his death but also to the central event of the history of salvation, which his life, Passion, death, and Resurrection represent—namely, to the resurrected Christ who is henceforth present as the living Lord of the church and guides and increases it through his Holy Spirit. This led to the literary form of church tradition—the Holy Scripture. As the "New Testament," it takes its place next to the Holy Scripture of Judaism, henceforth reinterpreted as the "Old Testament." The tradition of the church itself thereby entered into the characteristic Christian tension between spirit and letter. The spirit creates tradition but also breaks tradition as soon as the latter is solidified into an external written form and thus impedes charismatic life.

The role of kerygma in tradition

Throughout church history, however, the core of this field of tension is formed by the transmission of the Christ event—the kerygma—itself. On the one hand, the kerygma is the bearer and starting point for tradition; on the other hand, it molds the impetus for ever-new impulses toward charismatic, fresh interpretations and, under certain circumstances, suggests or even enforces a conscious elimination of accumulated traditions. Decisive in this respect is the self-understanding of the church. According to the self-understanding of the Roman Catholic and the Eastern Orthodox churches, the church, as the institution of Jesus Christ, is the bearer of the oral and the written tradition. It is the church that created the New Testament canon. The selection of canonical writings undertaken by it already presupposes a dogmatic distinction between "ecclesiastical" teachings—which, in the opinion of its responsible leaders, are "apostolic"—and "heretical" teachings. It thereby already presupposes a far-reaching intellectualization of the tradition and its identification with "doctrine." The oral tradition thus became formalized in fixed creedal formulas.

Accordingly, in the history of the Christian Church a specific, characteristic dialectic has been evidenced between periods of excessive growth and formalistic hardening of tradition that hindered and smothered the charismatic life of the church and periods of a reduction of tradition that follow new reformational movements. The latter occurred, in part, within the church itself, such as in the reforms of Cluny, the Franciscans, and the Dominicans; they also took on the form of revolutionary movements. The Reformation of the 16th century exhibited various degrees of positions toward tradition. All of the Reformers broke with the institution of monasticism, the liturgical and sacramental tradition of the Roman Catholic Church, and certain elements of doctrinal tradition. Luther, however, was more conservative in his attitude toward the Roman Catholic Church than were Zwingli and Calvin. He was thus especially hated among the representatives of the radical Reformation—*e.g.,* the Anabaptists and Enthusiasts (Schwärmer), who demanded and practiced a revolutionary break with the entire Roman Catholic tradition. The new churches that arose from the Reformation, however, soon created their own new traditions. This was made necessary by the predominance of both the didactic, doctrinaire principle and the founding of one's own church upon one's own "confessional writings." Practical manifestations against the tradition of the Roman Catholic Church also had public effects—*e.g.,* the eating of sausage on fast days in Zürich at the start of Zwingli's reformation or the provocative marriages of monks and nuns.

The Reformation view of tradition

In the 19th century, the period of a progressive revolutionizing of political life in Europe and North and South America, the Roman Catholic Church sought to safeguard its tradition—threatened on all sides—through an emphatic program of "antimodernism." It endeavoured to protect tradition both by law and through theology (*e.g.,* in returning to a strict, obligatory neo-Thomism). The representatives of this development were the popes from Pius IX (reigned 1846–78) to Pius XII (reigned 1939–58). With Pope John XXIII (reigned 1958–63), a dismantling (*aggiornamento*) of antimodernism and a more critical attitude vis-à-vis the "tradition" set in; this extended to traditional dogmatic views as well as to the liturgy and church structure. The second Vatican Council (1962–65) guided this development into moderate channels. On the other hand, an opposite development has taken place in the Soviet Union and the eastern European countries. In these nations the remains of the Orthodox Church, which survived extermination campaigns of the Leninist and Stalinist eras from the 1920s to the 1950s, preserved themselves in a political environment hostile to the church precisely through a retreat to their church tradition and religious functioning in the realm of the liturgy. In the World Council of Churches, the Orthodox Church in the latter part of the 20th century has viewed its task as the bearer of Christian tradition over against the predominant social-ethical tendencies of certain Protestant member churches that have disregarded or de-emphasized the tradition of the church in a wave of antihistorical sentiment.

The role of exegesis in scriptural traditions

The most important creation of church tradition is that of the Holy Scriptures themselves and, secondarily, the exegesis (critical interpretations and explanations) of the Scriptures. Exegesis first appeared in Christian circles among Gnostic heretics and the church catechists (teachers)—*e.g.,* in the Christian school systems, such as in Alexandria and Antioch. The heretics, who could not claim the unbroken apostolic tradition maintained by the Orthodox Christian churches, had a necessary interest in claiming the tradition to justify their own movements. Thus, exegesis was directly related to the development of a normative scriptural canon in the Orthodox churches. A similar need for the interpretation of an ecclesiastically fixed scriptural canon resulted in the Christian school system.

The first representatives of early church exegesis were not the bishops but rather the "teachers" (*didaskaloi*) of the catechetical schools, modeled after the Hellenistic philosophers' schools in which interpretive and philological principles had been developed according to the traditions of the founders of the respective schools. The allegorical interpretation of Greek classical philosophical and poetical texts, which was prevalent at the Library and Museum (the school) of Alexandria, for example, directly influenced

the exegetical method of the Christian Catechetical school there. Basing his principles on the methods of Philo of Alexandria and Clement of Alexandria, his teacher, and others, Origen—the Christian Catechetical school's most significant representative—created the foundation for the type of Christian exegesis (*i.e.,* the typological-allegorical method) that lasted from the patristic period and the Middle Ages up to the time of Luther in the 16th century. Origen based his exegesis upon comprehensive textual-critical work that was common to current Hellenistic practices, such as collecting Hebrew texts and Greek parallel translations of the Old Testament. His main concern, however, was that of ascertaining the spiritual meaning of the Scriptures, the trans-historical divine truth that is hidden in the records of the history of salvation in the Scriptures. He thus developed a system containing four types of interpretation: literal, moral, typological, and allegorical.

The view of "teachers" as charismatic figures (*i.e.,* those gifted by the Holy Spirit with the ability to uncover the hidden spiritual meaning of the letter) long hindered Western theologians in developing their own exegetical works. Exegetical literature was restricted to "chains" (*catenae*), in which excerpts from commentaries or homilies of the charismatic Fathers were joined together in a "chain" for the individual words and sentences of the Holy Scriptures. This was similar to the way in which early medieval theological works were composed of "sentences"—*i.e.,* individual doctrinal definitions from the writings of authoritative church teachers along with a limited commentary. Typological exegesis attained special significance for medieval Christian mysticism, which was inspired to a great extent by the allegorical interpretation of the Song of Solomon as the wedding between Christ and the soul.

The exegetical principles of the Reformation

Only with the Reformation, under the leadership of Luther, did there emerge an emphatic turning away from the allegorical exegesis and a turning toward the literal meaning of the Scriptures. This had its beginnings in the early church in the theological school of Antioch. In contrast to the Platonic tradition of the school of Alexandria, the school of Antioch was guided by Aristotelian philosophy. In place of allegorizing, which was consciously rejected, Antiochene exegesis was very much occupied with textual criticism. Both traditions often were included together in the so-called glosses of the Latin Middle Ages, such as in the *Glossa ordinaria* ("Ordinary Glosses"), edited by Anselm of Laon (died 1117), and the *Postillae*—the first biblical commentary to be printed (1471–72)—of Nicholas of Lyra (*c.* 1270–1349).

According to his own statement, Luther's reformational breakthrough came about through a fresh exegetical reflection—*legendo et docendo* ("by perusing and teaching")—in connection with his lectures on the Bible at the university of Wittenberg in Germany. He used the preliminary work of humanist philologists for the restoration of the Old and New Testament text (*e.g.,* Erasmus' 1516 edition of the Greek New Testament in the lectures on the Letter of Paul to the Romans). Luther replaced the traditional schema of the fourfold meaning of the Scripture with a spiritual interpretation of the letter—*i.e.,* one based on Christ. Inasmuch as the letter, which speaks historically of the work of Christ, at the same time always means this work as the salvation event that has happened "for us," it always contains the spiritual meaning in itself. In debates with the Spiritualists and Enthusiasts, who made use of the allegorical-tropological (figurative) method, Luther appealed ever more strongly to the unequivocal "clarity" of the letter of the Scriptures, which contains the "clarity" of the "subject" expressed by it. His exegesis is thus also a dogmatic one. The struggle between historical and tropological exegesis was emphasized in the debate between Luther and Zwingli over the understanding of the Lord's Supper.

During the early 18th century, biblical interpretation free of dogmatic interest was achieved among theologians accused of heresy by orthodox colleagues of their confession, such as among the Dutch Arminians (*e.g.,* Hugo Grotius and Johann Jakob Wettstein). Interest in the history of the Old and New Testament period was growing; ancient Middle Eastern history, biblical geography and archaeology, and the history of the religions of the ancient East and Hellenism were being included in the interpretation of the Scriptures. Under the influence of the Enlightenment, the historical criticism of the Bible, which was independent of the moral and edifying evaluation of the Holy Scriptures, was established. Soon including criticism of early church dogma, it led directly to the rise of historical criticism of the Bible in the 19th and 20th centuries.

Development of historical criticism of the Bible

In addition to the tradition of the Holy Scriptures and its interpretation, traditions centring on holy places also developed. The veneration of holy places is the oldest expression of Christian popular piety. From Judaism the Christian Church adopted the idea and practice of venerating holy places. In post-exilic Judaism (*i.e.,* after the 5th century BC), Jerusalem became the sanctuary and the centre of the Jews in Palestine as well as the goal of the pilgrimages of Jews of the Diaspora. After the destruction in AD 70 of Jerusalem, which was the holy city for the early church, it remained for Christians—as the site of the suffering and Resurrection of Jesus Christ and as the place of his return in glory—a holy city and a goal of pilgrimages. Such early bishops as Melito of Sardis and Alexander of Jerusalem and such theologians as Origen embarked on pilgrimages to Jerusalem. When the Christian Church became the state church in the 4th century, pilgrimages to the holy places in Palestine became popular.

The journey of the empress mother Helena to the Holy Land before AD 330 inaugurated the cult of relics through the alleged discovery of the holy cross. Constantine built the Church of the Holy Sepulchre in Jerusalem (335) and the Church of the Nativity over the Grotto of the Nativity in Bethlehem. The numerous other biblical commemorative places of the Old and New Testament history soon followed.

The cult of martyrs and saints led to establishment of shrines outside Palestine that were developed into pilgrimage places. The idea that the martyrs are present at the places of their martyrdom (*e.g.,* Peter's tomb at the Vatican) secured a prominent position for holy places connected with the cult of saints and martyrs. The cult of the martyrs was developed especially in the Roman catacombs, and it contributed to the formation of the Petrine doctrine and the teaching of the primacy of the Roman bishop. After the 4th century the cult of martyrs spread further and created an abundance of new holy places in the West: thus, Santiago de Compostela in Spain was connected with the tomb of James, to which equal rank with Rome and Jerusalem was later accorded; then Trèves in Germany, with the tomb of Matthew, which exerted a special power of attraction through the relic of the holy robe; and Marburg in Germany, with the shrine of St. Elizabeth. In the Middle Ages, during the development of the Roman Catholic sacrament of penance, holy places became places of grace, the visitation of which was considered a work of penance.

The veneration of relics

The original historical consciousness of the Christian Church is also alive in the cult of relics. In the relics of the body in which the saint suffered martyrdom, the saint himself is believed to be present, or at least something of the power of the Holy Spirit that filled him. The cult of relics began as a result of veneration of a martyr at his or her tomb, over which later was erected an altar of the church built to honour the saint. From the 4th century on in the East, and later also in the West, the remains of the martyrs were distributed in order that as many as possible could share in their miraculous power. Fragments of relics were sewn into a silken cloth (antimension), and the Eucharist could be celebrated only upon an altar that was covered with such an antimension. In times of persecution the Eucharist could be celebrated upon any table, as long as it was covered with the antimension and consecrated through the presence of the martyr. In the Latin Church the relics are enclosed in a cavity (*sepulcrum*) in the altar top. During the deconsecration of a church, the relic is again removed from the *sepulcrum.*

In the late Middles Ages the character of the pilgrimage, just like the veneration of relics, underwent a degeneration in connection with the degeneration of the sacrament of penance because of the abuse of the indulgence.

Luther's critique of the indulgence began with a criticism of the display of the elector of Saxony Frederick III the Wise's imposing collection of relics in the Schlosskirche (Castle Church) of Wittenberg on All Saints' Day (1516). Over against the attacks of Luther, the Council of Trent declared that

> the holy bodies of the holy martyrs and others living with Christ, whose bodies were living members of Christ and temples of the Holy Spirit, and will be by him raised to eternal life and glorified, are to be venerated by the faithful, since by them God bestows many benefits upon men.

In order to avoid the development of a holy place at his grave and a reliquary and saintly cult around his person, Calvin arranged by will that his body be buried at an unknown spot. The erection of the giant monument to the Reformer at the supposed place of his burial shows the futility of his effort and the strength of the Christian consciousness of tradition.

The roots, deals, and purposes of monasticism

**Monasticism.** Monasticism, an institution based on the Christian ideal of perfection, has its roots in New Testament Christianity, in which the baptized were designated as the "perfect ones." In the early church, monasticism equated perfection with world-denying asceticism, along with the view that perfect Christianity centred its way of life on the maximum love of God and neighbour.

Monastic discipline, in the course of time, became an external means for the attainment of this ideal of perfect love of God and neighbour. Only a few especially disciplined persons, however, have been able to live according to the path that leads to the ideal of perfection. The masses, on the other hand, are inwardly and outwardly incapable of exercising ascetic discipline. Therefore, the monastic rules of life were not generally binding "commands" but rather only "counsels" directed to those called to lead an ascetic life. The essential distinction between command and counsel is found in the words of Jesus: he did not command men to "make themselves eunuchs for the sake of the kingdom of heaven," but rather he recommended this condition only to those who were "able to receive this" (Matthew 19:12). Unmarried ascetics were recognized as a special class in the early church, forming the core of many churches. Later, with its distinction between counsel (*suasum*) and command (*iussum*), as in the writings of Tertullian in the late 2nd century, the church found itself in full accord with the oldest Christian view. During the latter part of the 2nd and the beginning of the 3rd century, the combination of asceticism and mysticism, which was to become the spiritual basis of later monasticism in the East and in part also in the West, was emphasized by Clement of Alexandria and Origen.

By the 4th century, monasticism had become an established institution in the Christian Church. This was not because of the decadence of the people of late antiquity, as has often been asserted, but rather because monasticism was sustained by the resilient and culturally unexhausted rural populations of Egypt and Syria, who had developed an enthusiasm for asceticism itself. Out of the desire for still further advanced isolation, ascetics moved from areas in proximity to inhabited places and established themselves in tombs, abandoned and half-deteriorated human settlements, caves, and, finally, into the wilderness areas of the deserts. The main task of the ascetics—*i.e.,* struggle with the demons—thereby underwent a heightened intensification: the desert was considered the abode of the demons, the place of refuge of the pagan gods falling back before a victorious Christianity. Hence, the expansion of Christianity in the cities of Egypt and the rise of Egyptian desert monasticism in the 4th century occurred both because the masses streamed into the churches as a result of the official imperial toleration and support policies and because ascetics striving for perfection left the cities and moved into the desert in significant numbers.

The influence of the monastic leaders

Certain writings that captured the spirit of monasticism further enhanced the development of this way of life in the church. Athanasius of Alexandria, the 4th century's most significant bishop spiritually and in terms of ecclesiastical politics, wrote the *Life of St. Antony,* which described the eremitic (hermit) life in the desert and the awesome struggle of ascetics with the demons as the model of the life of Christian perfection. This work indicates that the church sanctioned and propagated monasticism.

A former Roman soldier of the 4th century, Pachomius, created the first monastery in the modern sense. He united the monks under one roof in a community living under the leadership of an abbot (father, or leader). In 323 he founded the first true monastic cloister in Tabennisi, north of Thebes, in Egypt, and joined together houses of 30 to 40 monks, each with its own superior. Pachomius also created a monastic rule that, however, served more as a regulation of external monastic life than spiritual guidance. During the remainder of the 4th century, monasticism soon developed in areas outside Egypt. Athanasius brought the monastic rule of Pachomius to the West during his banishment (340–346) to Trèves in Germany—as a result of his opposition to the imperially sanctioned heretical doctrines of Arianism. Mar Awgin, a Syrian monk, introduced the monastic rule in Mesopotamia, and Jerome established a monastic cloister in Bethlehem.

Basil the Great, one of the three Cappadocian Fathers of the 4th century, definitively shaped monastic community life in the Byzantine Church. His ascetic writings furnished the theological and instructional foundation for the "common life" (cenobitism) of monks. He became the creator of a monastic rule that, through constant variations and modifications, became authoritative for later Orthodox monasticism. The Rule of Basil has preserved the Orthodox combination of asceticism and mysticism into the 20th century.

The development of Western monasticism

Western monasticism, founded by Benedict of Nursia (Italy) in the 6th century, has gone through a double form of special development vis-à-vis early church monasticism. The first consists of its clericalization. In modern Roman Catholic cloisters, monks are, except for the serving brothers (*fratres*), ordained priests and are thereby drawn in a direct way into the ecclesiastical tasks of the Roman Church. Originally, however, monks were laymen. Pachomius had explicitly forbidden monks to become priests on the ground that "it is good not to covet power and glory." Basil the Great, however, by means of a special vow and a special ceremony, enabled monks to cease being just laymen and to attain a position between that of the clergy and the laity. Even in the 20th century, monks of the Orthodox Church are, for the most part, lay monks; only a few fathers (abbots) of each cloister are ordained priests (*hieromonachoi*), who are thus allowed to administer the sacraments.

The second special development in Roman Catholicism consists of the functional characteristics of its many orders. The individual orders aid the church in its various areas of activity—*e.g.,* missions, education, care for the sick and needy, and combating heresy. Developing a wide-ranging diversification in its structure and sociological interests, Roman Catholic monasticism has extended all the way from the knightly orders to orders of mendicant friars, and it has included orders of decided feudal and aristocratic characteristics alongside orders of purely bourgeois characteristics. To the degree that special missionary, pedagogical, scholarly-theological, and ecclesiastically political tasks of the orders increased in the West, the character of ancient monasticism—originally focused completely on prayer, meditation, and contemplation—receded more and more in importance. Few monastic orders—the Benedictines and the Carmelites are notable exceptions—still attempt to preserve the ancient character and purposes of monasticism in Roman Catholicism in the 20th century.

The meaning of "saint"

**The saintly life.** In Christian popular piety the saint plays a very significant role. Originally a self-designation of all Christians collectively, "the saints," understood in this broad sense, are "sanctified through the name of the Lord Jesus Christ and through the Spirit of our God," according to the First Letter of Paul to the Corinthians, chapter 1, verse 31. On the one hand, the saint may be understood as a Christian who endeavours to fulfill the binding demand of moral holiness in obedience to God and in love of his neighbour (2 Corinthians 7:1; 1 Thessalonians 4:3), or a charismatic figure in whom the gifts of the Holy Spirit operate according to the personal and temporal circumstances of such an individual. Because

of certain views on being "called to holiness," members of many radical sects have designated themselves as "the saints"—from Oliver Cromwell's "saints" in 17th-century England to the Mormon "latter-day saints" in the 19th and 20th centuries.

The general meaning of "saint" was transformed during the period of the persecutions of Christians in the Roman Empire. The martyr, the witness in blood to Christ and follower in his suffering, became the prototype of the future ideal of the saint. Veneration of the saints began because of a belief that martyrs were received directly into heaven after their martyrdoms and that their intercession with God was especially effective—in the Revelation to John the martyrs occupy a special position in heaven, immediately under the altar of God (Revelation 6:9). Veneration of confessors (*i.e.,* those who had not denied their belief in Christ but had not been martyred), bishops, popes, early Church Fathers, and ascetics who had led a godlike life was established soon after cessation of the persecutions.

The veneration of saints

In the Greek church the saints were regarded as charismatic figures in whom the prototype of Christ is reflected in multifarious images. Veneration of the saints in the Orthodox churches was thus based more upon the idea that the saints provided instructional examples of the Christian life of sanctification. In the West, however, cultic veneration of the saints, the concept of patron saints, and the view that saints are helpers in need became predominant. The cult of the saints gradually came under the control of the papacy, which regulated cultic veneration of a saintly personality extolled in popular piety by means of a process of canonization strictly defined by canon law. The saints thus dominated the church calendar, which notes the names of the ecclesiastically recognized saints of each day of the year. They are venerated on a particular day in the prayer of intercession, and references are made to their deeds, sufferings, and miracles in the liturgy.

Under Pope Paul VI, the Roman Catholic Church attempted to reduce the significance of the veneration of saints—and thereby emphasize the idea of their historical exemplariness—by deleting some unhistorical, ostensibly mythological figures from the calendar of saints. The difference between historical and mythological saints, however, is difficult to maintain in details because mythological features from pre-Christian hero myths had often been intermixed, even in the lives of demonstrably historical saints. Thus, deletion of saints from the calendar has had little success in popular piety. Pope John Paul II, fully respectful of the directions of the second Vatican Council, did, however, pay renewed respect to some of the pre-council forms of devotion which the reformers had tended to displace.

In the early church the veneration of saints at first was restricted to celebrations at their tombs, but the cult of saintly relics soon spread the veneration of particular saints to many areas. The *Martyrdom of Polycarp,* for example, called the remains of the bishop Polycarp of Smyrna, martyred in 156/167, "more precious than costly stones and more excellent than gold." A belief in the need of special protection by saints is the basis of the system of patron saints: most Roman Catholic churches have a saint as their patron, whose presence in the church is represented by a relic of that particular saint. Saints, however, became patrons not only of churches but also of cities, regions, vocational groups, or classes. Saints also won a special significance as patrons of names: in the Roman Catholic and Eastern Orthodox churches a Christian generally received the name of the saint on whose holiday (day of death) he is baptized. The believer is thus joined for life with the patron of his name through the name and the name day, which, as the day of rebirth (*i.e.,* baptism), is of much greater significance than the natural birthday.

In the Eastern Orthodox Church, relics of saints appear less frequently, but icons of saints appear in greater numbers. Though cultic veneration of saints as patrons, tutelary saints, and helpers in need has increased through the centuries, the view that the saints are examples of the Christian life of sanctification has been preserved. The Roman Catholic Church, through its use of the canonization of saints, has constantly established new models for practical religiosity and morality to meet contemporary needs—raising to the position of sainthood personages all the way from the holy king to the holy servant girl.

Reformation and modern views of saints

In view of the excess of the veneration of saints, the Reformation not only eliminated the cultic veneration of saints but also images and relics of the saints from the churches and homes. Although the Reformation did not theoretically deny the saints their significance as historical witnesses to the power and grace of God, through such radical measures it virtually eliminated the meaning of saints as guiding images and examples of Christian life. Under the influence of Luther's view that all believers are saints, the veneration of the saints and their relics also was either de-emphasized or eliminated. The experience of martyrdom in the times of persecution in the Reformation and Counter-Reformation encouraged the development of a new saintly ideal in the radical Protestant sects in connection with the renewal of a strict demand for sanctification. Such was the case in the Baptists' "Chronicle of the Martyrs" as well as in Spiritualism. The Swedish archbishop Nathan Söderblom's attempt at awakening in Protestantism in the 20th century a new understanding of the saint received notice in Protestant ecumenical circles and led to a rediscovery of saints in the Protestant realm (*e.g.,* through Walter Nigg's book *Great Saints*). In modern Roman Catholicism, emphasis is increasingly being placed upon the charismatic aspects of the saints and their significance as models of a spiritual, holy Christian life.

**Art and iconography.** Christian art constitutes an essential element of the Christian religion. Until the 17th century the history of Western art was largely identical with the history of Western ecclesiastical and religious art. During the first three centuries of the Christian Church, however, there was no Christian art, and the church generally resisted it with all its might. Clement of Alexandria, for example, criticized religious (pagan) art in that it encouraged people to worship that which is created rather than the Creator. About the mid-3rd century an incipient pictorial art began to be used and accepted in the Christian Church but not without fervent opposition in some congregations. Only when the Christian Church became the Roman imperial church under Emperor Constantine in the early 4th century were pictures used in the churches, and they then began to strike roots in Christian popular religiosity.

Early attitudes concerning the use of religious art

Later, however, when pictorial art was publicly placed in the service of the church, warnings against this development were voiced by leading theologians. The church historian Eusebius, the most diligent glorifier of Constantine, characterized the use of images of the Apostles Paul and Peter as well as of Christ himself as a pagan custom. Asterius, bishop of Amaseia in Pontus during the late 4th and early 5th centuries, similarly stated in a sermon:

> Do not picture Christ on your garments. It is enough that he once suffered the humiliation of dwelling in a human body which of his own accord he assumed for our sakes. So, not upon your robes, but upon your soul, carry about his image.

Epiphanius (*c.* 315–402), bishop of Salamis in Greece, also energetically opposed in word and deed the disposition toward images in the imperial church:

> Have God always in your hearts, but not in the community house, for it does not become a Christian to expect the elation of his soul from recourse to his eyes and the roaming about of his senses.

Christian art developed at such a late stage because of its origins in Judaism and its opposition to paganism and the emperor's cult. In addition to a faith in God the Father, Creator of heaven and Earth, and faith in the uniqueness and holiness of God, Christianity also received from its Jewish origins a prohibition against the use of images to depict the sacred or holy, including humans, who were created in "the image of God." The early Christian Church was also deeply involved in a struggle against paganism, which, to the Christian observer, was viewed as idolatry in that its many gods were represented in various pictorial and statuary forms. In early Christian missionary preaching, the Old Testament attacks upon pagan veneration of images were transferred directly to pagan image veneration of the first three centuries AD. The struggle against

images was conducted as a battle against "idols" with all the intensity of faith in the oneness and exclusiveness of the imageless biblical God.

Abhorrence of images also was furthered because the emperor's cult was so despised by Christians. Christians were compelled, through anti-Christian legislation, to venerate the imperial images by offering sacrifices to them. Refusal to make the sacrifice was the chief cause of martyrdom. Characteristically, thus, the Christian Church's reaction after its public recognition was expressed in the riotous destruction of the pagan divine images.

In spite of these very strong religious and emotional restraints, the church developed a form of art peculiar to its needs. Protestants often have held that the development of ecclesiastical art was a part of the entire process of the inner decay of the Christian Church when it was elevated to the position of the officially favoured religious institution of the Roman Empire. In other words, some groups within Protestantism have claimed that the development of church art was part of the process of the church's inner paganization.

The incarnation as the basis for the use of images

The starting point for the development of Christian pictorial art, however, lies in the basic teaching of the Christian revelation itself—namely, the incarnation, the point at which the Christian proclamation is differentiated from Judaism. The incarnation of the Son of man, the Messiah, in the form of a human being—who was created in the "image of God"—granted theological approval of a sort to the use of images that symbolized Christian truths. Clement of Alexandria, at one point, called God "the Great Artist," who formed humans according to the image of the Logos, the archetypal light of light. The great theological struggles over the use of images within the church during the period of the so-called Iconoclastic Controversy in the 8th and 9th centuries indicate how a new understanding of images emerged on the basis of Christian doctrine. This new understanding was developed into a theology of icons that still prevails in the Eastern Orthodox Church in the 20th century.

The great significance of images of the saints for the Orthodox faithful is primarily expressed in the cultic veneration of the images within the worship service. Second, it is expressed in the dogmatic fixation of the figures, gestures, and colours in Eastern Church iconic art. In the West, the creative achievement of the individual artist is admired, but Orthodox painting dispenses with the predominance of the individual painter's freely creative imagination. Throughout the centuries the Eastern Church has been content with reproducing certain types of holy images, and only seldom does an individual artist play a predominant role within the history of Orthodox Church painting. Most Orthodox ecclesiastical artists have remained anonymous. Icon painting is viewed as a holy skill that is practiced in cloisters in which definite schools of painting have developed. In the schools, traditional principles prevail so much that different artist-monks generally perform only certain functions in the production of a single icon. Style motifs—*e.g.,* composition, impartation of colour, hair and beard fashions, and gestures of the figures—are fixed in painting books that contain the canons of the different monastic schools of icon painters.

The significance of icons in Eastern Orthodoxy

The significance of the image of the saint in the theology, piety, and liturgy of the Eastern Orthodox Church can be judged historically from the fact that the struggle over holy images within Orthodox Church history brought about a movement whose scope and meaning can be compared only with the Reformation of Luther and Calvin. In the 7th century a tendency hostile to images and fostered by both theological and political figures gained ground within the Byzantine Church and upset Orthodox Christendom to its very depths; known as the Iconoclastic Controversy, it was supported by some reform-minded emperors. Although opponents of icons had all the political means of power at their disposal, they were not able to succeed in overthrowing the use of icons. The conclusion of this struggle with the victory of the supporters of the use of icons is celebrated in the entire Orthodox Church on the first Sunday of Lent as the Feast of Orthodoxy.

Orthodox icon painting is not to be separated from its ecclesiastical and liturgical function. The painting of the image is, in fact, a liturgical act in which the artist-monks prepare themselves by fasting, penance, and consecrating the materials necessary for the painting. Before the finished icon is used, it likewise is consecrated. Not viewed as a human work, an icon (according to 8th- and 9th-century literature) was understood instead as a manifestation of a heavenly archetype. A golden background is used on icons to indicate a heavenly perspective. The icon is always painted two-dimensionally because it is viewed as a window through which worshipers can view the heavenly archetype from their earthly position. A figure in the three-dimensionality of the plastic arts, such as sculpture, would thus be an abandonment of the character of epiphany (appearance).

Ideas of the iconic liturgy dominate the manuals of the Orthodox icon painters. The model of the Christ figure for icon painters was found in an apocryphal writing of the early church—the *Letter of Lentulus,* which was a legendary letter supposedly written by a certain Lentulus, who was named consul in the 12th year of the emperor Tiberius. As the superior of Pontius Pilate, the procurator of Judaea, he by chance was staying in Palestine at the time of the trial of Jesus. In an official report to the Emperor about the trial of Jesus, Lentulus included an official warrant for Jesus with a description of the Christ. This apocryphal description furnished the basic model for the Byzantine Christ type.

The Trinity also may not be represented, except in those forms in which, according to the view of Orthodox church doctrine, the Trinity showed itself in the divine Word of the Old and New Testaments. Early church theology interpreted an Old Testament passage (Genesis 18:1 ff.) as an appearance of the divine Trinity—namely, the visit of the three men with the patriarch Abraham at Mamre in Palestine. Also included in icons of the Trinity are the appearance of the three divine Persons—symbolized as a hand, a man, and a dove—at the baptism of Jesus (Matthew 3:16 ff.) and the Pentecostal scene, in which the Lord, ascended to heaven, sits at the right hand of God and the Comforter (the Holy Spirit) is sent down to the Apostles in the form of fiery tongues (Acts 2). Another Trinitarian iconic scene is the Transfiguration of Jesus at Mount Tabor (Matthew 17:2).

Icons of Mary were probably first created because of the development of Marian doctrines in the 3rd and 4th centuries. The lack of New Testament descriptions of Mary was compensated by numerous legends of Mary that concerned themselves especially with wondrous appearances of miraculous icons of the mother of God. In Russian and many other Orthodox churches, including the monasteries at Mount Athos, such miraculous mother of God icons, "not made by hands," have been placed where the appearances of the mother of God took place.

The consecration liturgy of the icons of saints expresses the fact that the saints themselves, for their part, are viewed as likenesses of Christ. In them, the image of God has been renewed again through the working of salvation of the incarnate Son of God.

Varying views on the theology of icons

The foes of images explicitly deny that the New Testament, in relation to the Old Testament, contains any new attitude toward images. Their basic theological outlook is that the divine is beyond all earthly form in its transcendence and spirituality; representation in earthly substances and forms of the divine already indicate its profanation. The relationship to God, who is Spirit, can only be a purely spiritual one; the worship of the individual as well as the community can happen only "in spirit and in truth" (John 4:24). Similarly, the divine archetype can also be realized only spiritually and morally in life. The religious path of the action of God upon humans is not the path of external influence upon the senses but rather that of spiritual action upon the mind and the will. Such an effect does not come about through the art of painting. Opponents of icons thus claim that the only way to reach an understanding of the truth is by studying the writings of the Old and New Testaments, which are filled with the Spirit of God.

The decisive contrast between the iconodules (image

lovers) and the iconoclasts (image destroyers) is found in their understanding of Christology. The iconodules based their theology upon the view of Athanasius—who reflected Alexandrian Christology—that Christ, the God become human, is the visible, earthly, and corporeal icon of the heavenly Father, created by God himself. The iconoclasts, on the other hand, explain, in terms of ancient Antiochene Christology, that the image conflicts with the ecclesiastical dogma of the Person of the Redeemer. It is unseemly, according to their views, to desire to portray a personality such as Christ, who is himself divine, because that would mean pulling the divine down into the materialistic realm.

The theology of the iconoclasts of the Reformation period in the West made use, for the most part, of the same arguments. For the radical Protestants, the realization of God is only in the Word and sacrament.

After iconic theology had overcome opposition in the Byzantine imperial church, there were numerous Christian groups—especially in Asia Minor—in which the old hostility toward church icons was still maintained and which, in part, already had been forced into positions of heresy, such as the Paulicians (members of a 7th–9th-century dualistic sect).

The history of iconoclasm began in the early church with an emphatic (and, from the viewpoint of lovers of Greek and Roman culture, catastrophic) iconoclastic movement that led to the annihilation of nearly all of the sacred art of the pagan religions of the Roman Empire. In Western Christendom, an iconoclastic attitude was again expressed in various medieval lay movements and sects, such as the Cathari and the Waldenses. Iconoclasm underwent a revolutionary outbreak in the 16th-century Reformation in Germany, France, and England. Despite the different historical types of iconoclasm, a surprising uniformity in regard to their affective structure and theological argumentation exists. The Iconoclastic Controversy of the 7th and 8th centuries also became a point of contention in the Western Church. To be sure, the latter had recognized the seventh ecumenical council at Nicaea (787), in which iconoclasm was condemned. Nevertheless, an entirely different situation existed in the West. The Frankish–Germanic Church was a young church in which images were much more infrequent than in the old Byzantine Church, in which holy icons had accumulated over the centuries. In the West there was still no Christian pictorial art as highly developed as in the East. Also, Christianity there did not have to struggle against a highly developed pagan pictorial art. Donar, a Germanic god, reputedly whispered in a holy oak, and Boniface merely had to fell the Donar oak in order to demonstrate the superiority of Christ over the pagan god. Among the Germanic tribes in the West, there was no guild of sculptors or goldsmiths, as in Ephesus (Acts 19: 24 ff.), who would have been able to protest in the name of their gods against the Christian iconoclasts.

The difference between Eastern and Western views on icons

The Western viewpoint is revealed most clearly in the formulations of the synodal decisions on the question of images, as they were promulgated in the Frankish kingdom in the *Libri Carolini,* Charlemagne's code of laws. In this work it is emphasized that images have only a representative character. Thus, they are understood not as an appearance of the saint but only as a visualization of the holy Persons for the support of recollecting spiritual meanings that have been expounded intellectually through sermons. Hence, this led to an essentially instructional and aesthetic concept of images. The Western Church also viewed images as the Holy Scriptures' substitute for the illiterate—*i.e.,* for the overwhelming majority of church people in this period. Images thus became the Bible for the laity. Pope Adrian I, who encouraged Western recognition of the iconodulic Council of Nicaea, also referred to the perspicuity of the icons. This idea of perspicuity—*i.e.,* the appeal to one's imagination to picture the biblical persons and events to oneself—enabled him to recognize the Greek high esteem for the image without completely accepting the complicated theological foundation for icon veneration. The ideas articulated in the *Libri Carolini* remained decisive for the Western tradition. According to Thomas Aquinas, one of the greatest medieval theologians of the West, images in the church serve a threefold purpose: (1) for the instruction of the uneducated in place of books; (2) for illustrating and remembering the mystery of the incarnation; and (3) for awakening the passion of devotion, which is kindled more effectively on the basis of viewing than through hearing.

Western theology of icons

In the Western theology of icons, the omnipotence of the two-dimensionality of church art also was abandoned. Alongside church pictorial painting, ecclesiastical plastic arts developed; even painting in the three-dimensional form was introduced through the means of perspective. Art, furthermore, became embedded in the entire life of personal religiosity. The holy image became the devotional image; the worshiper placed himself before an image and became engrossed in his meditation of the mysteries of the Christian revelation. As devotional images, the images became the focal points for contemplation and mystical representation. Conversely, the mystical vision itself worked its way back again into pictorial art, in that what was beheld in the vision was reproduced in church art. The burden of ecclesiastical tradition, which weighs heavily upon Byzantine art, has been gradually abolished in the Western Church. In the Eastern Church the art form is just as fixed as ecclesiastical dogma; nothing may be changed in the heavenly prototypes. This idea plays little or no role in the West. There, religious art adjusts itself at any given time to the total religious disposition of the church, to the general religious mental posture, and also to religious needs. Religious art in the West also has been shaped by the imaginative fantasy of the individual artist. Thus, from the outset, a much more individual church art developed in the West. Thus, it became possible to dissociate sacred history from its dogmatic milieu and to transpose it from the past into the actual present, thereby allowing for an adaptable development of ecclesiastical art.

**Missions.** The missions and expansion of Christianity are among the most unusual of historical occurrences. Other world religions, such as Buddhism and Islām, also have raised a claim to universal validity, but no world religion other than Christianity has succeeded in realizing this claim through missionary expansion over the entire world (see also below *The Christian community and the world: Christian missions*).

The influence of the concept of the end time

The unique global expansion of Christianity is directly related to its expectation of the end time, in the imminent expectation of the return of Christ. The Christian expectation of the end time never consisted simply of a passive yearning for the coming Kingdom of God. Being grasped by faith in its immediately impending arrival was expressed instead in an intense activation and acceleration of efforts to prepare the world for the return of Christ and the coming of the Kingdom. This state of being grasped transformed itself into the pressing duty to "prepare the way of the Lord" (Matthew 3:3) and to remove all resistance to the establishment of the Kingdom on Earth.

This eschatological pressure stands behind both the earlier and the later achievements of an ever wider expansion of Christianity. Columbus, in undertaking to cross the ocean in a westerly direction in the 15th century, for example, believed that Satan had settled in India, thus successfully disrupting the extension of the gospel and delaying the return of Christ. According to his eschatological calculations, the time for the return of Christ was nearly at hand; thus, India had to be reached by the shortest way possible so that the last bulwark of Satan might be removed through Christian missions. The same eschatological expectation drove the Spanish Jesuit Francis Xavier to India and Japan in the 16th century. Protestant world missions, commencing a century later, also were influenced by the eschatological expectation of the end time (*e.g.,* the missions of the German Lutherans Bartholomäus Ziegenbalg and Heinrich Plütschau in India in the early 18th century and the missions of the Puritans among the Indians in Massachusetts in the late 17th century). The first seal of Massachusetts displayed an Indian with a beckoning hand and the inscription "Come over and help us"—the words of the Macedonian who appeared to the Apostle Paul in a night vision (Acts 16:9).

The leading missionaries of all times have accomplished great feats of extensive travels. On his numerous mission-

Extensive travels of Christian missionaries

ary journeys, the Apostle Paul showed a greater accomplishment in distances traveled than any known general of the Roman army, official of the Roman Empire, or trader of his time. Francis Xavier also traveled more than any other known person in his times and endured intense physical exertions on land and sea. John R. Mott, founder of the World's Student Christian Federation, was the most widely traveled man of the first half of the 20th century. The catchphrase coined by him, "Jesus Christ to the nations in this generation," has been the basic principle of all the great and small missionary impulses that have contributed to the worldwide expansion of Christianity.

This eschatological aspect of Christian missions has continued through the 20th century, especially among Pentecostals and Adventists. The missionary institutions of these churches come from the tradition of the conservative evangelical churches, which maintain a strong inclination toward an imminent expectation.

The ideal of ascetic homelessness

Related to the eschatological motif in missions is the ideal of ascetic homelessness. In imitation of the homeless Christ, who "has nowhere to lay his head" (Matthew 8:20), the early medieval Scots-Irish monks—as radical Christian ascetics—demanded the renunciation of that which is dearest to humans: one's own home. "For the sake of Christ" they assumed ascetic homelessness by leaving their cloisters—often in groups of 12 under the leadership of a 13th—and ventured abroad. They traveled to continental Europe—especially in Celtic areas—as far as Switzerland and over the Alps and also went to Iceland. Similarly, Russian Orthodox hermits and monks, who often had to flee because of repressive measures by the state and the state church, conducted missions in areas northeast of the Soviet Union, Siberia, the Aleutian Islands, and Alaska. An example of a modern ascetic missionary is the French nobleman Charles-Eugène de Foucauld (1858–1916), who became a martyred anchorite missionary among the Bedouin of the Sahara.

## LAST THINGS

The "last things" were the first things, in terms of urgency, for the faithful of the early church. The central content of their faith and their hope was the coming Kingdom of God. They believed that the promises of the Old Testament about the coming bringer of salvation had been fulfilled in Jesus Christ, but that the fulfillment was not yet complete. Thus, they awaited Christ's Second Coming, which they believed was imminent.

**Expectations of the Kingdom of God in early Christianity.** In early Christianity's expectation of the Kingdom of God, two types were inherited from Judaism. The first was the expectation of a messianic Kingdom in this world, with its centre in Jerusalem, which was to be established by an earthly Messiah from the house of David. The second expectation was that of a heavenly Kingdom, which was to be inaugurated by the heavenly Messiah, Son of man, and in which the elected comrades of the Kingdom from all times would share in the state of the resurrection.

Types of expectations inherited from Judaism

The two types of expectation of salvation did not remain neatly separated in the early church but rather intersected one another in manifold ways. Under the influence of the persecutions of the church, a characteristic combination of the end-time expectations was established. In Paul's letters and in the Revelation to John, the faithful Christians will first reign together with their returning Lord for some time in this world. Those Christians who are still alive at his return will take part in the reign without dying (1 Thessalonians 4:17). Christians who have already died will rise again and, as resurrected ones, share in the Kingdom upon Earth. Only after completion of this first act of the events of the end time will there then follow the general resurrection of all the dead and the Last Judgment, in which the elect will participate as co-judges (1 Corinthians 6:2).

In the Revelation to John this expectation is condensed into the concept of the 1,000-year (millennial) Kingdom. For 1,000 years the dragon (Satan) is to be chained up and thrown into the abyss. In John's vision, Christians, the first resurrected, "came to life and reigned with Christ a thousand years" (Revelation 20:4). Only later does the resurrection of all the dead take place, as well as the general judgment, creation of the new heaven and the new Earth, and the descent of the new Jerusalem. According to the view of the Revelation to John, this 1,000-year Kingdom is composed of the chosen comrades of the Kingdom, especially the martyrs and all who stood the test in times of persecution; it is a Kingdom of the privileged elect.

Expectations of the millennium

This promise has exerted revolutionary effects in the course of church history. In the early church the expectation of the millennium was viewed as a social and political utopia, a state in which the chosen Christians would rule and judge with their Lord in this world. Such chiliastic (or millennial) expectations provided the impetus for ecclesiastical, political, and social reformations and revolutions in the course of church history. The establishment of a 1,000-year Kingdom in which the elect, with Christ, will reign and receive the administrative and judgeship posts has fascinated religious expectations as well as political and social imagination far more than the second part of the eschatological expectation, the "Last Judgment."

The delay of the Parousia resulted in a weakening of the imminent expectation in the early church. In this process of "de-eschatologizing," the institutional church increasingly replaced the expected Kingdom of God. The formation of the Roman Catholic Church as a hierarchical institution is directly connected with the declining of the imminent expectation. The theology of Augustine constitutes the conclusion of this development in the West. He de-emphasized the original imminent expectation by declaring that the Kingdom of God has already begun in this world with the institution of the church; the church is the historical representative of the Kingdom of God on Earth. The first resurrection, according to Augustine, occurs constantly within the church in the form of the sacrament of baptism, through which the faithful are introduced into the Kingdom of God. The expectation of the coming Kingdom of God, the resurrection of the faithful, and the Last Judgment have in actuality finally become a doctrine of the "last things" because the gifts of salvation of the coming Kingdom of God are interpreted as being already present in the sacraments of the church.

**Expectations of the Kingdom of God in the medieval and Reformation periods.** Nevertheless, the original imminent expectation has spontaneously and constantly reemerged in the history of Christianity. In the period before the 16th-century Reformation, heretical groups—such as the 2nd-century Montanists and the medieval Cathari, Waldenses, the followers of Joachim of Fiore, and the Franciscan Spirituals—accused the Roman Catholic Church of betraying the original eschatological imminent expectation. These groups revived eschatological expectations. Even within the Roman Catholic Church itself, however, such movements have constantly reemerged to inspire reform efforts. In the medieval church new outbreaks of an imminent expectation occurred in connection with great historical catastrophes, such as epidemics of the plague, Islāmic invasions, schisms, and wars.

Eschatological interpretations of the Reformers

Luther's Reformation also was sustained by an imminent expectation. For the Reformers, the starting point for their eschatological interpretation of contemporary history was that the "internal Antichrist," the pope, had established himself in the temple at the Holy Place and that through persecution by the "external Antichrist," the Turk, the church had entered into the travails of the end time. The Reformation churches, however, soon became institutional territorial churches, which in turn repressed the end-time expectation, and thus doctrine of the "last things" became an appendix to dogmatics.

**Expectations of the Kingdom of God in the post-Reformation period.** In the post-Reformation period, the imminent expectation appeared in individual groups on the margin of the institutional Reformation churches; such groups generally made the imminent expectation itself the object of their sect formation. This has been the result of the fact that, since the Reformation, the Roman Catholic Church has been virtually immune to eschatological movements. The Lutheran Church has been less immune; a series of eschatological groups whose activity in the church was determined by their expectation of the

imminent return of Christ appeared in Pietism. Among the congregational and evangelical churches of England and America, the formation of new eschatological groups has been a frequent occurrence, especially during the period of the English Revolution in the 17th century and during the revival movements in the United States in the 18th–20th centuries. Such groups shared significantly in the renewal and expansion of Christianity in domestic and foreign missions. Indeed, by late in the 20th century much of the Christian missionary outreach had passed into the hands of millennial-minded groups.

**The role of imminent expectation in missions and emigrations.** The great missionary activities of the Christian Church are in most cases based upon a reawakened imminent expectation, which creates a characteristic tension. The tension between the universal mission of the church and the hitherto omitted missionary duties as well as the idea that the colossal task must be accomplished in the shortest time possible renders comprehensible the astonishing physical and spiritual achievements of the great Christian missionaries. After the inundation of the old Christian areas of Africa and Asia by Islām, Franciscan missionaries in the 13th and 14th centuries, enduring incredible hardships, went by land and by sea to India, China, and Mongolia to preach the gospel. In a similar way, the missionary movement of the 18th and 19th centuries also proceeded from such eschatological groups within Protestantism.

The expectation of the Kingdom of God, in the form of the imminent expectation, plays a strong role in emigration movements. In a sense, the Crusades could be included among such movements. Great masses of European Christians again and again set out for Palestine with a sense of finding there the land of their salvation and personally being present when Christ returns there to establish his Kingdom. The eschatological strain of the Crusades can be noted in the Crusade sermons of Bernard of Clairvaux in 1147, who kindled enthusiasm to liberate Jerusalem with reference to the pressing terminal dates of the end time.

Eschatological ideas in emigrations

The emigration movement toward America also was influenced by beliefs in eschatologically fixed dates (*e.g.,* Columbus). Puritans who traveled to America in the 17th century and Quakers, Baptists, and Methodists in the 18th century believed that America was the "wilderness" promised in the Revelation to John. William Penn gave the name Philadelphia to the capital of the woodland areas ceded to him (1681) because he took up the idea of establishing the true church of the end time, represented by the Philadelphia community of the Revelation to John. A great number of the attempts undertaken to found radical Christian communities in North America may be viewed as anticipations of the coming Jerusalem. The same holds true for the emigration of German revivalists of the 18th and early 19th centuries to Russia and Palestine. The "Friends of the Temple"—Swabians who went with Christoph Hoffmann to Palestine in 1866—and the Swabians, Franks, Hessians, and Bavarians, who after the Napoleonic Wars followed the call of Tsar Alexander I to Bessarabia, were all dominated by the idea of living in the end time and preparing themselves for the coming Kingdom of God. In Tsar Alexander I they saw the "eagle . . . as it flew in midheaven" (Revelation 8:13), which prepared the "recovery spot" for them in the East upon which Christ will descend.

**Eschatological expectations and secularization.** In the eyes of some theologians, the very process of secularization, which progressively rules out transcendent explanations of natural and historical conditions, has been a working out of a form of eschatological expectation. Of course, the substance is quite different in the cases where people work in expectation of the Kingdom of God and in the other cases where they become "futurologists." But the impulse to prepare oneself for such futures has analogues and origins, it is contended, in old Christian ideas of penance and preparation for the coming Kingdom.

In the Gospels the attitude toward the coming Kingdom of God led, over and beyond the expectation of nullifying sin and death, to certain worldly conclusions of an organizational kind. The disciples of Jesus knew that there will be "first ones" in the Kingdom of Heaven; they pressed for the administrative posts in the coming Kingdom of God (*e.g.,* the Apostles James and John). The promise, too, that they are to take part as judges at the Last Judgment (Luke 22:30) sparked definite conceptions of rank. Jesus castigated them in their disputes over rank with the words, "If any one would be first, he must be last of all and servant of all" (Mark 9:35).

Despite this warning, the imminent expectation of the coming Kingdom of God awakened concrete, substantial ideas that led ever closer to social utopias. With the 18th-century German Lutheran mystic and Pietist F.C. Oetinger, the end-time expectation generated definite social and political demands—*e.g.,* dissolution of the state, abolition of property, and elimination of class differences. Some of the aspects of the end-time expectation of Pietism were revived in the French Revolution's political and social programs. The transition from the end-time expectation to the social utopia, however, had already been achieved in writings from the 16th and early 17th centuries—*e.g.,* the English humanist and saint Thomas More's . . . *de optimo reipublicae statu deque nova Insula Utopia* (1516; "On the Highest State of a Republic and on the New Island Utopia"), the German theologian Johann Valentin Andrea's *Reipublicae Christianopolitanae Descriptio* (1619; "A Description of the Christian Republic"), the English philosopher Francis Bacon's *New Atlantis* (1627), and the English bishop Francis Godwin's *Man in the Moone* (1638). It is also found in early socialism of the 19th century—*e.g.,* the French social reformer Henri de Saint-Simon's *Nouveau Christianisme* (1825; "The New Christianity") and the French Socialist Étienne Cabet's *Voyage en Icarie* (1840; "Voyage to Icaria").

Concepts of social utopias and futurology

What distinguishes the Christian social utopia from the earlier kind of eschatology is the stronger emphasis upon social responsibility for the preparation of the Kingdom of God and a considerable preponderance of various techniques in the establishment of the utopian society. (In general, the end-time expectation has also inspired technical fantasy and science fiction.) Also characteristic is the basic attitude that people themselves must prepare the future perfect society in a formative and organizing manner and that "hoping" and "awaiting" are replaced by human initiative. A graduated transition from a social utopia still consciously Christian to a purely Socialist one can be observed in the writings and activities of the French Socialists Charles Fourier, Saint-Simon, and Pierre-Joseph Proudhon, the English Socialist Robert Owen, and the German Socialist Wilhelm Weitling. Secularized remnants of a glowing Christian end-time expectation are still found even in the Marxist view of the social utopia.

Modern planning and projection of alternative futures is a secularization of the end-time expectations previously envisioned in Christian terms. The future is thus manipulated through planning (*i.e.,* "horizontal eschatology") in place of eschatological "hoping" and "waiting for" fulfillment. "Horizontal eschatology" is thus taken out of the sphere of the unexpected and numinous (spiritual); it is made the subject not only of a detailed prognosis based upon statistics but also of a detailed programming undertaken on the basis of this prognosis. An eschatological remainder is found only in an ideological image of man, upon which programming and planning are based.

**Concepts of life after death.** The Christian end-time expectation is directed not only at the future of the church but also at the future of the individual believer. It includes definite conceptions of the personal continuance of life after death. Many baptized early Christians were convinced they would not die at all but would still experience the advent of Christ in their lifetimes and would go directly into the Kingdom of God without death. Others were convinced they would go through the air to meet Christ returning upon the clouds of the sky: "Then we who are alive, who are left, shall be caught up together with them in the clouds to meet the Lord in the air; and so we shall always be with the Lord" (1 Thessalonians 4:17). In the early imminent expectation, the period between death and the coming of the Kingdom still constituted no ob-

ject of concern. An expectation that one enters into bliss or perdition immediately after death is also found in the words of Jesus on the cross: "Today you will be with me in Paradise" (Luke 23:43).

Eternal life and eternal death

In the Nicene Creed the life of the Christian is characterized as "eternal life." In the Gospels and in the apostolic letters, "eternal" is first of all a temporal designation: in contrast to life of this world, eternal life has a deathless duration. In its essence, however, it is life according to God's kind of eternity—*i.e.,* perfect, sharing in his glory and bliss (Romans 2:7, 10). "Eternal life" in the Christian sense is thus not identical with "immortality of the soul"; rather, it is only to be understood in connection with the expectation of the resurrection. "Continuance" is neutral vis-à-vis the opposition of salvation and disaster, but the raising from the dead leads to judgment, and its decision can also mean eternal punishment (Matthew 25:46). The antithesis to eternal life is not earthly life but eternal death.

Eternal life is personal life, and precisely therein is fulfilled the essence of man who is created according to the image of God. Within eternal life there are differences. In the present life there are variations in talent, duty, responsibility, and breadth and height of life, just as there are also distinctions in "wages" according to the measure of the occupation, the sacrifice of suffering, and the trial (1 Corinthians 3:8). Correspondingly, the resurrected are also distinguished in eternal life according to their "glory":

> There is one glory of the sun, and another glory of the moon, and another glory of the stars; for star differs from star in glory. So it is with the resurrection of the dead (1 Corinthians 15:41–42).

This expectation has had a great influence upon the Christian conception of marriage and friendship. The idea of a continuation of marriage and friendship after death has contributed very much to the deepening of the view of marriage, as is shown by the strong influence of the 17th–18th-century Swedish mystic, philosopher, and scientist Emanuel Swedenborg's ideas upon the romantic philosophy of religion and its interpretation of marriage and friendship in the thought of the German scholars Friedrich Schelling and Friedrich Schleiermacher. The Western concept of personality was thus deepened through the Christian view of its eternal value.

The delay of the imminent expectation brought about the question of the fate of the dead person in the period between the death of the individual Christian and the resurrection. Two basic views were developed. One view is that of an individual judgment, which takes place immediately after death and brings the individual to an interim state, from which he enters into the realm of bliss or that of perdition. The idea of an individual judgment, however, cannot be readily harmonized with the concept of the general Last Judgment on the day of the general resurrection of the dead. It anticipates the decision of the general judgment and thus deprives of its significance the notion of the Last Judgment. A second view, therefore, also prevailed: the sleep of the soul—*i.e.,* the soul of the dead person enters into a sleeping state that continues until the Last Judgment, which will occur after the general resurrection. At the Last Judgment the resurrected will be assigned either to eternal life or eternal damnation. This conception, accepted in many churches, contains many discrepancies, especially the abandonment of the fundamental idea of the continuity of personal life.

Both views contain an inhuman consequence. The first leaves to people no further opportunity to improve the mistakes of their lives and to expiate their guilt. The second preserves the personality in an intermediate state for an indefinite period so as to later punish it for sins or reward it for good deeds from a time prior to entrance into the sleep of the soul.

The concept of purgatory

The belief in purgatory (an interim state in which a correction of a dead person's evil condition is still possible) of the Roman Catholic Church gives the deceased opportunities for repentance and penance to ameliorate their situation.

The presupposition of the doctrine of purgatory is that there is a special judgment for each individual at once after death. Hence, the logical conclusion is that purgatory ceases with the Last Judgment. The stay in purgatory can be shortened through intercession, alms, indulgences, and benefits of the sacrifice of the mass. The Eastern Orthodox Church has no doctrine of purgatory but does practice an intercession for the dead. It assumes that, on the basis of the connection between the church of the living and that of the dead, an exertion of influence upon the fate of the dead through intercession is possible before the time of the Last Judgment.

The idea of the Last Judgment has often become incomprehensible to the modern world. At the most, people apparently are still open to the concept of judgment of the guilt and innocence of the individual. The idea decisive for the early church's expectation of the Judgment, however, was that the Last Judgment will be a public one. This corresponds to the fundamental Christian idea that human beings—both the living and the dead—are bound together in an indissoluble communion; it presupposes the conception of the church as the body of Christ. All of humanity is as one person. Humans sin with one another, and their evil is connected together in the "realm of sin" in a manifold way, unrecognizable in the individual. Each person is responsible for the other and is guilty with the other. The judgment upon each person, therefore, concerns all. Judgment upon the individual is thus at the same time judgment upon the whole, and vice versa. The Judgment is also public in regard to the positive side—the praise and reward of God for that which is done rightly and practiced in the common life, often without knowing it.

For the most part, the churches of the latter part of the 20th century no longer have the courage to uphold the Christian teaching of life after death. The church has long neglected teachings about the entire area of the last things. The New Testament responses presuppose the imminent expectation and thus leave many questions unanswered that arose because of the delay of the Parousia. The doctrine of the sleep of the soul, on the other hand, contains many consequences that question the fundamental idea of the Christian view of the personality of the *imago Dei* ("image of God"). The beginnings of a further development of the Christian view of life after death, as are found in Swedenborg, have never been recognized positively by the church. For this reason, since the period of Romanticism and Idealism, ideas of the transmigration of souls and reincarnation, taken over from Hinduism and Buddhism, have gained a footing in Christian views of the end-time expectation. Some important impulses toward a new understanding of the view of life after death are found in Christian theosophy, such as the idea of a further development of the human personality upon other celestial bodies after death. (E.W.B./M.E.M.)

Effects of neglect of the doctrine of life after death

## Church year

The Christian church year is an annual cycle of seasons and days observed in the Christian churches in commemoration of the life, death, and Resurrection of Jesus Christ and of his virtues as exhibited in the lives of his saints. This section surveys the origin and meaning, development, and current revisions of this cycle.

### ORIGINS OF THE CHURCH YEAR

**Religious times and seasons.** The church year has deep roots in the primitive human impulse to mark certain times with sacral significance and ritual observance. These are times when conscious attention is given to the mysterious forces that surround and involve all living creatures in the natural and inexorable cycles of light and darkness; labour and rest; birth, growth, decay, and death.

Two interrelated cycles have had primary importance in the shaping of religious calendars. One is cosmic: the phases of the Moon and the solar equinoxes and solstices. The other is the periodic succession of the seasons of nature that determines times of sowing and reaping. Both cycles speak to the mystery of birth, death, and rebirth and to human dependence upon the fecundity of life given in the natural creation.

**Jewish background.** The Jewish religious year, grounded in the divinely revealed Law of the Old Testament, was the foundation for the church year of Christians. It is a lunar-

month calendar stemming from the primitive nomadic life of the Hebrews, with its chief festival at the first full moon of spring, known later as the Passover. Grafted onto this calendar after the settlement of the Hebrew tribes in Palestine were the agricultural festivals—dependent upon "the early and later rains"—the firstfruits at Passover, the first harvest at the Feast of Weeks or Pentecost, and the autumn harvest at the Feast of Tabernacles or Booths.

The sabbath

Of uncertain origin, but prior to the monarchical period (11th to 6th century BC), the Hebrews observed a seven-day week, of which the last day, or sabbath, was a holiday and day of rest. Whatever its original purpose, it became transformed into a sacral day, consecrated to Yahweh, the one God of the Hebrews, and increasingly surrounded with restrictions upon all activity other than worship. In the time of Jesus (1st century AD), "keeping holy the sabbath day" was a principal hallmark of adherence to Judaism.

The remarkable aspect of the Jewish religious year was its transformation, in successive codifications of the Old Testament Law, into a series of historical commemorations associated with God's deed in creation and in the redemption of God's people. At first, the sabbath was related to the Exodus, the deliverance of the Hebrews from Egypt in the 13th century BC (Deuteronomy 5:15), and, later, to the repose of God at the completion of creation (Exodus 20:8–11; Genesis 2:2–3). The three agricultural feasts became a sequence of remembrances of the Exodus from Egypt and the pilgrimage through the wilderness to the promised land (Exodus 12:1–20; Leviticus 23; Deuteronomy 16:1–17). Through these annual celebrations the devout Jew relived the saving events of the past and anticipated the final deliverance of the people of God in the age to come. Rabban Gamaliel, a contemporary of Jesus, said, "In every generation a man must so regard himself as if he came forth himself out of Egypt. . . . " (from Mishna, *Pesaḥim* 10:5).

**Formation of the church year.** In his earthly life Jesus was subject to the law of sabbath, feast, and fast prescribed in the Old Testament; but his ministry and teaching pointed to a new age, the coming Kingdom of God, when the Law would be fulfilled. He was, therefore, not so much concerned with outward conformity to legal regulations as he was with the spirit in which they were observed. "The sabbath was made for man, and not man for the sabbath" (Mark 2:27). It was in the context of a celebration of the Passover feast with his disciples that he was arrested, tried, and put to death.

Early Christians believed that the new age promised by Jesus had dawned with his Resurrection, on "the first day of the week" (Matthew 28:1; Mark 16:2; Luke 24:1; John 20:1). By this event the Law was fulfilled. Now every day and time were viewed as holy for the celebration and remembrance of Jesus' triumph over sin and death. Though many of his disciples continued to observe the special times and seasons of the Jewish Law, new converts broke with the custom because they regarded it as no longer needful or necessary. Paul, himself a dutiful observant of the Law, considered the keeping of holy days a matter of indifference, provided the devotion be "in honor of the Lord" (Romans 14:5–9). He warned his converts not to judge one another with regard "to a festival or a new moon or a sabbath" (Colossians 2:16).

From the beginning the church took over from Judaism the seven-day week. Before the end of the apostolic age (1st century AD), as the church became predominantly Gentile in membership, the first day of the week, or Sunday, had become the normative time when Christians assembled for their distinctive acts of worship, in commemoration of the Lord's Resurrection (Acts 20:7; 1 Corinthians 16:2). During the first two centuries AD, the Greco-Roman world in general adopted the planetary seven-day week of the astrologers.

Sunday

Christian writers of the 2nd century came to view Sunday, "the Lord's day," as a symbol of Christianity in distinction from Judaism. Most of the churches decided to observe the Lord's Passover (Easter) always on a Sunday, after the Jewish feast was over. In addition, local churches began to celebrate the anniversaries of the deaths of their martyrs, called "birthdays in eternity," for these also were regarded as witnesses to the resurrection triumph of Christ in his followers. The weekly Sunday and the annual Paschal (Passover) observance of 50 days from Easter to Pentecost (the Jewish harvest festival that also commemorated the revelation of the Law to Moses) were thus the principal framework of the church year until the 4th century—reminders of the new age to be brought by Christ at his coming again in glory at the end of time, when the true believers would enter their inheritance of perpetual joy and feasting with their Redeemer and Lord.

The establishment of Christianity as a state religion, following the conversion of the emperor Constantine (AD 312), brought new developments. The Paschal season was matched by a longer season of preparation (Lent) for the many new candidates for baptism at the Easter ceremonies, and the discipline and penance of those who for grievous sins had been cut off from the communion of the church.

A new focus of celebration, to commemorate the birthday of Christ, the world Redeemer, was instituted at ancient winter solstices (December 25 and January 6) to rival the pagan feasts in honour of the birth of a new age brought by the Unconquered Sun. Later, the Western churches created a preparatory season for this festival, known as Advent. Many new days were gradually added to the roster of martyr anniversaries to commemorate distinguished leaders, the dedication of buildings and shrines in honour of the saints, and the transferral of their relics.

### THE MAJOR CHURCH CALENDARS

Unlike the cycle of feasts and fasts of the Jewish Law, the Christian year has never been based upon a divine revelation. It is rather a tradition that is always subject to change by ecclesiastical law. Each self-governing church maintains the right to order the church year according to pastoral needs of edification. The pattern of the year therefore varies in the several churches of the East and of the West. The subtle adjustments of a lunar-month calendar, with its movable date of Easter, and a solar calendar of fixed dates require many rules to avoid conflict of observances.

Reforms in the church year

In the Western churches periodic reforms of the church year have occurred, notably in the Reformation era and again in the 20th century. The Protestant Reformers of the 16th century took differing attitudes toward such reforms. With their strong sense of the prime authority of Scripture and of the freedom of the gospel from all legalisms in liturgical matters, they revised the church year with varying degrees of radicalism. Lutherans and Anglicans took a conservative position, retaining the traditional seasons but eliminating commemorations that had no connection with the biblical record.

The Reformed churches, on the other hand, allowed only those feasts with a clear basis in the New Testament: Sundays, Holy Week and Easter, Pentecost, and in some cases Christmas. The Church of Scotland and Anabaptist and Puritan groups abolished the church year entirely, except for Sundays. In recent years this attitude has been very much modified. Their protest has been a reminder to the church that all days are regarded as belonging to Christ in the freedom of his Spirit, who cannot be controlled by rigid systems of fixed special observances.

In the late 20th century in the Western churches the church year was being subjected to an overall revision comparable in scope only to that of the 16th century. This was due to a number of currents of interest that were converging; *i.e.*, advances in historical and liturgical studies, changes in theological perspectives, and ecumenical encounters.

The basic structure of the church year was the creation of the ancient churches in the varied cultures surrounding the Mediterranean Sea that were embraced in the Roman Empire. Christian missionaries have carried the church year throughout the world—first in the Northern Hemisphere and, since the 16th century, in the Southern Hemisphere, where the natural seasons are reversed. It is unlikely that the dates of the two major feasts, Easter and Christmas, which control the seasons of the church year, will be changed. But new symbols and popular customs

associated with them will emerge in areas where, for example, Easter is celebrated in the autumn rather than as a spring festival.

The church year consists of two concurrent cycles: (1) the Proper of Time (Temporale), or seasons and Sundays that revolve around the movable date of Easter and the fixed date of Christmas, and (2) the Proper of Saints (Sanctorale), other commemorations on fixed dates of the year. Every season and holy day is a celebration, albeit with different emphases, of the total revelation and redemption of Christ, which are "made present at all times" or proclaim "the paschal mystery as achieved in the saints who have suffered and been glorified with Christ" (second Vatican Council, "Constitution on the Sacred Liturgy"). The church year is an epitome in time of the history of salvation in Christ.

Orthodox calendar

**Eastern churches.** The Orthodox churches of the Byzantine tradition recall the Resurrection of Christ every Sunday. Many Sundays take their title from the Gospel lesson for the day. In addition to Easter, "the feast of feasts," there are 12 other major feasts: Christmas, Epiphany, Hypapante (Meeting of Christ with Simeon, February 2), Palm Sunday, Ascension, Pentecost, Transfiguration (August 6), Exaltation of the Holy Cross (September 14), and four feasts of the Blessed Virgin Mary—her Nativity (September 8), Presentation in the Temple (November 21), Annunciation (March 25), and Falling Asleep (August 15).

The principal cycle consists of (1) 10 weeks before Easter, contained in the *Triōdion* (pre-Easter liturgical service book); the first four of these Sundays prepare for the Great Fast, or Lent (*i.e.,* the Sunday of the Pharisee and Publican; the Sunday of the Prodigal Son; Meat-Fast Sunday, after which abstinence from meat is enjoined; and Cheese-Fast Sunday, after which the fast includes cheese, eggs, butter, and milk), and (2) eight weeks after Easter, contained in the *Pentēkostarion* (post-Easter liturgical service book), including the Feast of Ascension, 40 days after Easter, and concluding with the Festival of All Saints on the Sunday after Pentecost. Other special commemorations of the period are the Feast of Orthodoxy, on the first Sunday in Lent, recalling the end of the Iconoclastic Controversy in 843, and the feast of the Fathers of the first Ecumenical Council of Nicaea in 325 on the sixth Sunday after Easter.

The schedule of fixed holy days in the Menaion (liturgical service book for each month) begins on September 1, the New Year's or Indiction Day of the Byzantine Empire. It includes the invariable feasts of Christ, St. Mary and other Christian saints, and many Old Testament saints.

Calendar of the Eastern independent churches

The separated churches of the East (those not accepting the jurisdiction of Orthodox patriarchs or bishops) have calendars basically similar to the Byzantine. West Syrians (Jacobites) and East Syrians (Nestorians) begin the year with a series of Sundays devoted to themes of the Dedication of the Church (consecration by a bishop) and the Annunciation (of the angel Gabriel to Mary that she would bear the Son of God)—the West Syrian sequence starting on November 1, the East Syrian on December 1. There are few saints' days in the Nestorian calendar. The Copts (Egyptians) and Ethiopians date their year from August 29, considered the beginning of the Christian Era in the persecution of the emperor Diocletian (AD 303–311). They have some 32 feasts of the Virgin Mary and many feasts of angels. The Armenian Church follows the Byzantine in beginning the year with the preparatory Sundays before Lent, but it commonly observes fixed holy days on the nearest Sunday. It is the only ancient church that never adopted the feast of Christmas on December 25 but celebrates the incarnation only on Epiphany, January 6.

**Roman Catholic Church.** The church year begins on the first Sunday in Advent, which is the fourth Sunday before Christmas Day. Until 1969, after Advent and Christmas, there followed the seasons of Epiphany, Pre-Lent, Lent, Easter, Ascension, and Pentecost. The first day of Lent is Ash Wednesday, being the 40th day (exclusive of Sundays) before Easter. A special festival of the Holy Trinity occurs on the first Sunday after Pentecost. Corpus Christi, a feast celebrating the Real Presence of Christ in the bread and wine of the Eucharist (Communion meal, or the Lord's Supper), was instituted in 1264 by Pope Urban IV and is observed on the Thursday after Trinity Sunday. In 1925 Pope Pius XI created the Feast of Christ the King, assigned to the last Sunday in October.

Fixed holy days

Until 1969, the fixed holy days began with St. Andrew (November 30), the nearest to the beginning of Advent. The three days before Ascension Day, called Minor Rogation Days ("Days of Asking"), are devoted to special prayers for fruitful harvests. Found only in the Roman Catholic Church are the fasts of the four seasons (*quatuor tempora*), known as Ember Days, and especially associated with ordinations to the ministry. They occur on the Wednesdays, Fridays, and Saturdays after the third Sunday of Advent and the first Sunday in Lent, in the week of Pentecost and the week after Holy Cross Day (September 14).

A revised calendar was issued by Pope John XXIII in 1960. The "Constitution on the Sacred Liturgy" of the second Vatican Council called for further reforms. These have been completed in the new calendar and lectionary promulgated by Pope Paul VI in 1969.

The most important feature of the new calendar was the restoration of all Sundays as feasts of Christ. No saints' days, even of the Virgin Mary, may take precedence of a Sunday. In the Proper of Time, the season of Pre-Lent was eliminated, and two cycles were provided: (1) the principal seasons, Sundays, and holy days from Advent to Pentecost and (2) a schedule of 33 Sundays per annum to be observed in numbered sequence in place of the Sundays previously designated "after Epiphany" and "after Pentecost." The ancient Roman Feast of St. Mary was restored to January 1; a new Feast of the Baptism of Christ was assigned to the first Sunday after Epiphany; and the Feast of Christ the King was shifted to the last Sunday of Ordinary Time. All octaves were eliminated. Fixed holy days are now arranged from January 1.

The revised Roman Catholic calendar

A considerable simplification, reclassification, and in many cases shifting of dates were made in the Proper of Saints. Except for 13 "solemnities" (including major feasts of Christ and Mary) and 25 "feasts," all other saints' days and holy days were reduced to "memorials," either obligatory or optional—with the right of national and regional episcopal conferences to alter their rank. Ember and Rogation Days were assigned as votive masses to be observed according to regional directives.

Regulations regarding holy days and processes leading to the canonization of saints are controlled by the Sacred Congregation for Divine Worship (formerly the Congregation of Rites). Certain feasts, in addition to all Sundays, are designated "holy days of obligation," when all the faithful must attend Mass. In the United States these are: Christmas Day (December 25), the Feast of St. Mary (New Year's Day), Ascension Day, the Assumption of the Blessed Virgin Mary (August 15), All Saints' Day (November 1), and the Immaculate Conception of the Blessed Virgin Mary (December 8). In addition to these, "days of obligation" observed elsewhere include: St. Joseph's Day (March 19), the Annunciation (March 25), SS. Peter and Paul Day (June 29), and the Feast of Corpus Christi.

**Protestant churches.** Lutheran and Anglican churches preserve in their liturgies the seasons of the Roman Catholic calendar; but in general they reduced the fixed holy days to primary feasts of Christ and the Apostles and evangelists, Michaelmas Day (September 29), and All Saints' Day (November 1). In the second half of the year, Sundays were named "after Trinity." In the late 20th century the revisions of Lutheran and Anglican service books were influenced by the new designs of the Roman Catholic calendar, notably proposals to eliminate Pre-Lent and to name Sundays "after Pentecost" instead of "after Trinity." Anglican and Lutheran calendars were also enriching their entries with many non-biblical saints and holy days, but for optional observance. Lutherans celebrate a festival of the Reformation on October 31 or the Sunday preceding that date.

In other Protestant churches, only Sunday observance remains obligatory, including Easter and Pentecost. Holy Week is frequently observed, and Christmas is commonly celebrated liturgically on the Sunday preceding Decem-

ber 25. Among these Protestant churches, new service books and hymnals have exhibited interest in recovering the major seasons of the Proper of Time, from Advent to Pentecost, and in some cases the Feast of All Saints. Especially significant was the restoration of the seasons in the Reformed (Presbyterian) and Methodist churches.

Many Protestant churches devote Sundays to special themes of a religious, charitable, or civic nature, such as Race Relations, Rural Life, Christian Home, and Labour Sundays. Harvest festivals, common in the Western churches since the Middle Ages, have a distinctive American tradition in Thanksgiving Day, on the fourth Thursday in November. Traditionally held to have originated in the Plymouth (Mass.) colony in 1621, it was first proclaimed a national holiday by President Abraham Lincoln in 1863. Ecumenical services, now worldwide, are observed during the Octave or Week of Prayer for Christian Unity, January 18–25—a custom started by Paul James Wattson of the Franciscan Friars of the Atonement and developed by Abbé Paul Couturier. The week is jointly sponsored by the World Council of Churches and the Vatican Secretariat for Promoting Christian Unity.

### HISTORY OF THE CHURCH YEAR

**Sunday.** Regular Christian corporate worship on Sundays goes back to the apostolic age, but New Testament writings do not explain how the practice began. Jewish Christians probably kept the sabbath at the synagogue, then joined their Gentile fellow believers for Christian worship after the close of the sabbath at sundown, either in the evening or early Sunday morning. When the church became predominantly Gentile, Sunday remained as the customary day of worship. Assemblies for the Eucharist were common on Saturday, however, as well as on Sunday in the Eastern churches into the 5th century, and Eastern canons forbade the practice, customary in the Roman Church, of fasting on the sabbath.

The term Lord's Day, signifying the triumph of Christ in his Resurrection and the beginning of a new creation, was in use by the end of the 1st century (Revelation 1:10; *Didachē* 14; Ignatius of Antioch, *Magnesians* 9:1). Some writers referred to the sabbath as the rest promised to the people of God at the end of time and to Sunday as "the eighth day," or beginning of a new world (Hebrews 4:4–11; *Letter of Barnabas* 15).

Sunday as a legal holiday

In 321 the Roman emperor Constantine decreed Sunday to be a legal holiday and forbade all trade and work other than necessary agricultural labour. Later emperors extended the prohibition to include public amusements in the theatre and circus. Church councils of the period were more concerned to enforce the obligation of Sunday worship, the earliest being the Spanish Council of Elvira (*c.* 300); but a synod of Laodicea (*c.* 381) enjoined Christians not to "Judaize" but to work on the sabbath and rest, if possible, on the Lord's Day. The Old Testament commandment of sabbath rest received a spiritual interpretation from the Church Fathers when they applied it to Sunday; *e.g.,* Augustine of Hippo held that the sabbath rest from servile work meant abstention from sin (compare *Tract. in Joannis,* Book III, chapter 19; Book XX, chapter 2).

A literal application of the sabbath law to Sunday became evident in conciliar canons and civil laws of the Frankish kingdoms in the 6th century, climaxed by Charlemagne's capitulary adopted by the Council of Aachen, 789 (canon 80). Medieval legislation thereafter repeatedly sought to enforce the "holiday" of Sunday, as also of many other holy days, for the benefit of serfs and labourers.

Sabbatarian laws applied to Sunday were also continued by the Protestant Reformers. The Acts of Uniformity of Edward VI in 1552 and of Elizabeth I in 1559 required all persons to attend worship on Sunday, the latter imposing a fine for neglect to do so. The Church of England's Canons of 1604 (number 13) make similar provision. Many Puritans were strongly sabbatarian in sentiment. Some of them referred to Sunday as "the sabbath." In the Puritan colonies of New England, the so-called Blue Laws of Sunday observance were especially severe. Today some states and cities in the United States have statutes restricting certain trades and amusements on Sunday. Church laws continue to insist upon the moral obligation to attend worship every Lord's Day.

Blue Laws

**Advent.** The Advent (from Latin *adventus,* "coming") season is peculiar to the Western churches, though its original impulse probably came from the East, where it was common, after the ecumenical Council of Ephesus in 431, to devote sermons on Sundays before Christmas to the theme of the Annunciation. In Ravenna—a channel of Eastern influences upon the Western Church—Peter Chrysologus (reigned *c.* 433–450) delivered such homilies (sermons). The earliest reference to a season of Advent is the institution by Bishop Perpetuus of Tours (reigned 461–490) of a fast before Christmas, beginning from St. Martin's Day on November 11. Known as St. Martin's Lent, the custom was extended to other Frankish churches by the Council of Mâcon in 581.

The six-week season was adopted by the church of Milan and the churches of Spain. At Rome, there is no indication of Advent before the latter half of the 6th century, when it was reduced—probably by Pope Gregory I the Great—to four weeks before Christmas. The longer Gallican season left traces in medieval service books, notably the Use of Sarum (Salisbury), extensively followed in England, with its Sunday before Advent. The coming of Christ in his Nativity was overlaid with a second theme, also stemming from Gallican churches, namely, his Second Coming at the end of time. This interweaving of the themes of two advents of Christ gives the season a peculiar tension both of penitence and of joy in expectation of the Lord who is "at hand."

Popular piety in Advent is chiefly devoted to musical and dramatic performances based upon biblical prophecies and stories of the Nativity of Christ. In many homes and churches simple devotions are associated with an Advent evergreen wreath, in which four candles are inserted and lighted, one by one, each week, as a symbol of the coming of the "Light" of the world.

**Christmas.** The word Christmas is derived from the Old English *Cristes maesse,* "Christ's Mass." There is no certain tradition of the date of Christ's birth. Christian chronographers of the 3rd century believed that the creation of the world took place at the spring equinox, then reckoned as March 25; hence the new creation in the incarnation (*i.e.,* the conception) and death of Christ must therefore have occurred on the same day, with his birth following nine months later at the winter solstice, December 25. The oldest extant notice of a feast of Christ's Nativity occurs in a Roman almanac (the Chronographer of 354, or Philocalian Calendar), which indicates that the festival was observed by the church in Rome by the year 336.

Many have posited the theory that the feast of Christ's Nativity, the birthday of "the sun of righteousness" (Malachi 4:2), was instituted in Rome, or possibly North Africa, as a Christian rival to the pagan festival of the Unconquered Sun at the winter solstice. This syncretistic cult that leaned toward monotheism had been given official recognition by the emperor Aurelian in 274. It was popular in the armies of the Illyrian (Balkan) emperors of the late 3rd century, including Constantine's father. Constantine himself was an adherent before his conversion to Christianity in 312. There is, however, no evidence of any intervention by him to promote the Christian festival. The exact circumstances of the beginning of Christmas Day remain obscure.

From Rome the feast spread to other churches of the West and East, the last to adopt it being the Church of Jerusalem in the time of Bishop Juvenal (reigned 424–458). Coordinated with Epiphany, a feast of Eastern origin commemorating the manifestation of Christ to the world, the celebration of the incarnation of Christ as Redeemer and Light of the world was favoured by the intense concern of the church of the 4th and 5th centuries in formulating creeds and dogmatic definitions relating to Christ's divine and human natures.

Christmas is the most popular of all festivals among Christians and many non-Christians alike, and its observance combines many strands of tradition. From the ancient Roman pagan festivals of Saturnalia (December 17) and New Year's come the merrymaking and exchange

of presents. Old Germanic midwinter customs have contributed the lighting of the Yule log and decorations with evergreens. The Christmas tree comes from medieval German mystery plays centred in representations of the Tree of Paradise (Genesis 2:9). Francis of Assisi popularized the Christmas crib, or crèche, in his celebration at Greccio, Italy, in 1223.

Another popular medieval feast was that of St. Nicholas of Myra (*c.* 340) on December 6, when the saint was believed to visit children with admonitions and gifts, in preparation for the gift of the Christ child at Christmas. Through the Dutch the tradition of St. Nicholas (Sinterklaas, hence "Santa Claus") was brought to America in their colony of New Amsterdam, now New York. The sending of greeting cards at Christmas began in Britain in the 1840s and was introduced to the United States in the 1870s.

**Epiphany.** In Hellenistic times an epiphany (from the Greek *epiphania,* "manifestation"), or appearance of divine power in a person or event, was a common religious concept. The New Testament uses the word to denote the final appearing of Christ at the end of time; but in 2 Timothy 1:10 it refers to his coming as Saviour on earth. In this latter sense, a festival of Christ's epiphany is first attested among heretical Gnostic Christians (those who believed that mankind was saved by secret knowledge, not faith, and that matter was evil and the spiritual world good) in Egypt in the late 2nd century (Clement of Alexandria, *Strōmateis,* Book I, chapter 21), on January 6, when he was manifested as Son of God at his baptism. The date is that of an Egyptian solstice, celebrated by pagans as a time of overflow of the waters of the Nile, and in certain mystery cults as the occasion of the birth of a new eon, or age, from the virgin goddess Kore, daughter of the earth-mother goddess Demeter. In other places of the Middle East, the time was associated with miraculous fountains from which wine flowed in place of water.

Nothing more is known of an Epiphany feast until the 4th century, when it appears in the Eastern churches as a festival second in rank only to Easter. It commemorated three "manifestations": the birth, the baptism, and the first miracle of the Lord at Cana (John 2:1 ff.). In the latter half of the century Eastern and Western churches adopted each other's incarnation festival, thus establishing the 12-day celebration from Christmas to Epiphany. The particular emphasis in the Eastern feast upon the baptism of Christ led to special liturgical ceremonies of the blessing of waters and the ministration of baptism at this time. In the West, where Christmas was the primary festival, the Epiphany was associated particularly with the Adoration of the Magi to the infant Jesus (Matthew 2:1–12), as anticipation of the universal redemption of Christ in his "Manifestation to the Gentiles."

The 12 days of Christmas

**Pre-Lent.** A season of Pre-Lent, peculiar to the Roman Catholic rite, was eliminated from that calendar in 1969. It had developed in the 6th century as a time of special supplication for God's protection and defense in a period of great suffering in Italy from war, pestilence, and famine. It was marked by three Sundays before the beginning of Lent, called, respectively, Septuagesima, Sexagesima, and Quinquagesima—roughly 70, 60, and 50 days before Easter. Though not included in the discipline of Lenten penitence and fast, the season was related by some authorities to influences from the East, especially upon Roman monastic customs, for a longer Lent of eight weeks.

Shrove Tuesday, the day before Ash Wednesday (the initial day of Lent), is in many places a day of carnival, though its name derives from the custom of going to confession for absolution and penance before Lent (from the Middle English word *shriven,* "to shrive"). A famous carnival is that of Mardi Gras (French: "Fat Tuesday") in New Orleans.

**Lent.** The Lenten (from Middle English *lenten,* "spring") season is rooted in the preparation of candidates for baptism at the Paschal vigil. For several weeks they received intensive instruction, each session followed by prayer and exorcism. The earliest detailed account of these ceremonies is in the *Apostolic Tradition* (*c.* 200) of Hippolytus. At the conclusion all the faithful joined the catechumens (inquirers for instruction) in a strict fast on the Friday and Saturday before Easter. These were the days "when the Bridegroom was taken away" (compare Mark 2:20).

As a 40-day period (six weeks) Lent is mentioned in canon 5 of the first ecumenical Council of Nicaea in 325. In the 4th century instruction of the baptismal candidates was normally given by the bishop. Several such "catechetical lectures" on the creed and sacraments have survived, notably those of Cyril of Jerusalem and Theodore of Mopsuestia. Augustine's treatise *De catechizandis rudibus* (*c.* 400) gave a less dogmatic and more biblical and historical approach. The Roman Church organized its instruction around three (later seven) "scrutinies," at which the catechumens were introduced to the Gospels, the Apostles' Creed, and the Lord's Prayer.

Since Sunday was never a fast day, piety sought to conform the Lenten fast exactly to 40 days, after the examples of the 40 days in the wilderness of Moses, Elijah, and Christ. In the Eastern churches, where Saturdays were also excluded from fasting, this developed into an eight-week Lent. At Rome, from the late 5th century, the fast began on Wednesday before the first Sunday in Lent.

Ash Wednesday and Holy Week

During Lent also, grievous sinners were excluded from Communion and prepared for their restoration. As a sign of their penitence, they wore sackcloth and were sprinkled with ashes (Tertullian, *De paenitentia* 11; compare the biblical precedents: Jeremiah 6:26; Jonah 3:6; Matthew 11:21). This form of public penance began to die out in the 9th century. At the same time, it became customary for all the faithful to be reminded of the need for penitence by receiving an imposition of ashes on their foreheads on the first day of Lent—hence the name Ash Wednesday.

The last week of Lent was one of special devotion in remembrance of the Lord's Passion. Athanasius in his *Festal Letter* of 330 called it "holy Paschal week." The Church of Jerusalem in particular organized dramatic ceremonies during the week at appropriate holy sites of its neighbourhood. A detailed description is contained in the account of a Spanish nun (*c.* 395), *Peregrinatio ad loca sancta* (or *Peregrinatio Etheriae*). From Jerusalem many of these ceremonies, such as the Palm Sunday procession and the Good Friday veneration of the cross, spread to other churches.

The Roman Catholic liturgy of Holy Week begins with the blessing of palms and a procession on Sunday, with a solemn rendition of St. Matthew's Passion narrative at the mass. On Thursday the bishop blesses the sacred oils for the catechumens and the sick and the chrism (oil) for confirmation, and, in ancient times, penitents were reconciled for their Easter Communion. After a festal mass commemorating the institution of the Eucharist, the altars are stripped and washed. An additional ceremony, of medieval origin, has given its name to this day—the washing of feet, in imitation of the Lord's action at the Last Supper (John 13:2–15). It is popularly called the Maundy, from the anthem sung during the ceremony (Mandatum, "a new commandment," John 13:34).

Maundy Thursday

Another medieval custom, which has had a popular revival in the late 20th century, is the service of Tenebrae, held on Wednesday, Thursday, and Friday, in the evening. It is the old choir office of Matins and Lauds, originally sung before dawn and marked by the gradual extinguishing of candles before the breaking of the light of day.

On Good Friday (the day commemorating the Crucifixion of Christ), the Mass of the Presanctified is observed. Its name is derived from the fact that there is no consecration of the sacred elements of bread and wine; instead, Communion is ministered from the Reserved Sacrament (consecrated elements retained from previous celebrations). Other features are the singing of the Passion according to John, the impressive series of intercessions, and the adoration of the cross with singing of the Reproaches and the hymn "Pange lingua" ("Sing, my tongue, the glorious battle"). Following the Communion and dismissal of the people, there are no further liturgical rites other than the daily choir offices until the vigil of Easter.

**Easter.** The term Easter, commemorating the Resurrection of Christ, comes from the Old English *ēaster* or *ēastre,*

a festival of spring; the Greek and Latin Pascha, from the Hebrew Pesaḥ, "Passover." The earliest Christians celebrated the Lord's Passover at the same time as the Jews, during the night of the first (paschal) full moon of the first month of spring (Nisan 14–15). By the middle of the 2nd century most churches had transferred this celebration to the Sunday after the Jewish feast. But certain churches of Asia Minor clung to the older custom, for which they were denounced as "Judaizing" (Eusebius, *Ecclesiastical History,* Book V, chapters 23–25). The first ecumenical Council of Nicaea in 325 decreed that all churches should observe the feast together on a Sunday. Yet many disparities remained in the way the several churches calculated the date of Easter. Today the Eastern churches follow the Julian calendar, the Western churches its correction by Pope Gregory XIII in 1582, so that in some years there may be a month's difference in the time of celebration.

Differences in the celebration of Easter

Since 1900, various religious, business, and professional groups have promoted the concept of a fixed world calendar, which would include a fixed date for Easter. Proposals have been placed before the League of Nations and its successor, the United Nations. The second Vatican Council in its "Constitution on the Sacred Liturgy" (1963) accepted the principle of a fixed date for Easter, subject to approval by other churches, provided that no world calendar impaired the regular succession of a seven-day week. The World Council of Churches in the early 1970s canvased its member bodies to this end, and a large majority replied in favour of such a change. An Easter message of Athenagoras I, the Orthodox Patriarch of Constantinople, in 1969, called for a resolution of the differences between the Eastern and Western churches and a search for a common date. Among those preferring a fixed date for the observance of Easter—regardless of the issue respecting a common world calendar—the second Sunday in April has been widely proposed.

The Easter celebration continues for 50 days, to and including the Feast of Pentecost. In the early church, as on all Sundays, there was no fasting or kneeling in prayer during the period.

The liturgy began with a solemn vigil on Saturday evening. A new fire was lit for the blessing of the Paschal candle (the Exultet)—symbol of the driving away of the powers of darkness and death by the Passover of the Lord. There followed a series of lessons from the Old Testament, with a homily based upon the narrative of Exodus 12. Then, toward midnight, while the faithful were engaged in prayers, candidates for baptism were taken to the baptistery for their initiation. Returning to the assembly, they were confirmed by the bishop with chrism and the laying on of hands, and toward dawn the Easter Eucharist was completed. A similar celebration was repeated on the eve of Pentecost for those who were hindered from receiving baptism at Easter.

As at Christmas, so also at Easter, popular customs reflect many ancient pagan survivals—in this instance, connected with spring fertility rites, such as the symbols of the Easter egg and the Easter hare or rabbit. The Easter lamb, however, comes from the Jewish Passover ritual, as applied to Christ, "the Lamb of God" (compare John 1:29, 36; 1 Corinthians 5:7).

**Ascension.** At first, the church commemorated the Ascension (from the Latin *ascensio,* "ascent") of Christ into heaven, after his Resurrection (Luke 24:50–51; Acts 1:1–11), as part of the total victory of Christ celebrated from Easter to Pentecost. A special feast of the Ascension is not mentioned before the 4th century. The Spanish Council of Elvira (*c.* 300) appears to have rejected it as an unwarranted innovation. But by the end of the 4th century the feast had become universal in the church, on the 40th day after Easter.

The old English popular name for the feast is Holy Thursday, but there is no liturgical tradition to support the idea of an "Ascensiontide" as a season distinct from Easter. From the 10th century there developed an "octave" of Ascension, adopted at Rome in the 12th century but suppressed in 1955. The three days before Ascension Day, known as Minor Rogation Days, were instituted by Bishop Mamertus of Vienne (Gaul) in 470 and extended to all the Frankish churches at the Council of Orléans in 511. Pope Leo III (reigned 795–816) adopted them at Rome. They are observed by processional litanies and fasting as a supplication for clement weather for the crops and deliverance from pestilence and famine. In 1969 the Minor Rogation Days were changed to votive masses.

**Pentecost.** The Jews had an early harvest festival seven weeks after the firstfruit offerings of Passover, called the Feast of Weeks. The Priestly Code (Leviticus 23:15–16) assigned it to "the morrow after the seventh sabbath"—which would be a Sunday. Early rabbinic tradition (Babylonian Talmud, *Pesaḥim* 68b) associated the festival with the giving of the Law at Sinai, on the basis of Exodus 19:1.

The Christian festival of Pentecost (from the Greek *pentecoste,* "50th day"), unlike Easter, is not rooted in Judaism but is based upon the narrative of Acts 2, recording the gift of the Holy Spirit to the disciples and the launching of the church's mission to all peoples on the Pentecost that followed the Lord's Resurrection. The outpouring of the Spirit was the final seal upon Christ's redemptive work, a sign of the inauguration of the new age when the Law was fulfilled and the way to salvation opened to the Gentile peoples. For this reason the early Christians considered Pentecost to be included in, but climactic of, the great "50 days" of Easter. Pentecost was in fact the name commonly given by the early Fathers to the whole season.

New Testament origins of Pentecost

As early as the 5th century, baptisms were administered at Pentecost to those unable to be initiated at Easter, and a vigil rite was developed comparable to that of the Pascha (Leo the Great, *Letters* 16; Leonine and Gelasian sacramentaries). The Anglo-Saxons called the feast White Sunday (Whitsunday), from the white garments bestowed upon the newly baptized (compare Bede, *Ecclesiastical History,* Book II, chapter 9; *Penitential* of Archbishop Theodore of Tarsus). The term Whitsunday has been customary in the Anglican churches since the First Prayer Book of Edward VI (1549).

**Pentecost or Trinity Season.** The Sundays after Pentecost mark the season of the life of the church between the two advents of Christ as it fulfills its mission to the world under the guidance of the Spirit. Bishop Stephen of Liège (reigned 902–920) instituted a Feast of the Holy Trinity on the first Sunday after Pentecost, which spread through northern Europe. It was taken up in the Use of Sarum and was accepted at Rome in 1334 by Pope John XXII. It became common to date the Sundays after this feast, instead of after Pentecost, as in the Roman liturgy, and this practice was followed by the Carthusians and the Dominicans and in the Lutheran and Anglican churches.

**Saints' days and other holy days.** The celebration of days in honour of the saints or "heroes of the faith" is an extension of the devotion paid to Christ, since they are commemorated for the virtues in life and death that derive from his grace and holiness. Originally each local church had its own calendar. Standardization came with the fixation of the rites of the great patriarchal sees, which began in the 4th century and was completed for the Byzantine churches in the 9th century. The Roman calendar of the Gregorian Sacramentary became the basis of the Western Church's observances with the liturgical reform of Charlemagne (*c.* 800), but it was constantly supplemented throughout the Middle Ages by new additions from diocesan or provincial areas. It was not until 1634 that the Roman see gained complete control over the veneration and canonization of saints in the Roman Catholic churches subject to its jurisdiction.

Before the toleration of the Christian Church under Constantine (AD 312), the several churches commemorated only their martyrs, on the anniversaries of their deaths, commonly called their *natale,* or birthdays, with rites similar to those of Easter. By giving up life for their faith, often after cruel tortures, the martyrs were the supreme examples of the imitation of Christ. The earliest attested institution of such an anniversary is recorded in the *Martyrdom of Polycarp* of Smyrna (*c.* 155). The oldest Roman calendar of the martyrs reaches only to the beginning of the 3rd century and includes the joint martyrdom of the church's apostolic founders saints Peter and Paul (June 29), a feast apparently instituted in the year 258.

Veneration of martyrs and the relics of saints

After the age of the martyrs, the calendars continued to be enriched by entries of eminent bishops, teachers, ascetics, and missionaries. Other new feasts were associated with the transfer of the relics of saints to sumptuous shrines or churches dedicated in their honour. A precedent of great influence was the feast of dedication of the Church of the Holy Sepulchre (or Anastasis, "resurrection") at Jerusalem, on Sept. 14, 335, where the discovered tomb and cross of Christ were enshrined on the supposed site of his victory over death. The feast is popularly called Holy Cross Day. From the 4th to the 6th century many "inventions" or discoveries of relics were produced and fictitious "Acts" written to promote the cults of apostles, evangelists, and hitherto unknown martyrs of earlier times.

In the late 4th century a feast of All Martyrs was observed by the East Syrians on May 13 and by the West Syrians and Byzantines on the Sunday after Pentecost. Pope Boniface IV received from the emperor Phocas (reigned 602–610) the Pantheon at Rome, which he dedicated on May 13 to St. Mary and All Martyrs. The Feast of All Saints at Rome on November 1 was promulgated by Pope Gregory IV in 835, in place of the May festival. Some authorities believe this festival to be of Irish origin; others relate it to a chapel of All Saints in St. Peter's Basilica established by Pope Gregory III (reigned 731–741).

Veneration of Mary

Liturgical feasts in honour of Mary—related to the incarnation cycle—developed in the East after the third ecumenical Council of Ephesus in 431, where she was declared to be Theotokos ("God-bearer"). At Rome the earliest special commemoration was on the Octave of Christmas, but Pope Sergius I (reigned 687–701), an Easterner, introduced to Rome her four major feasts: her Nativity (September 8); Purification of the Blessed Virgin Mary (February 2, with its procession of candles—hence "Candlemas"); Annunciation (March 25); and Assumption (August 15).

#### LITURGICAL COLOURS

The early Christians had no system of colours associated with the seasons, nor do the Eastern Churches to this day have any rules or traditions in this matter. The Roman emperor Constantine gave Bishop Macarius of Jerusalem a "sacred robe . . . fashioned with golden threads" for use at baptisms (Theodoret, *Ecclesiastical History,* Book II, chapter 23). Toward the end of the 4th century, references are made to shining white garments worn by celebrants at the Eucharist (*Apostolic Constitutions,* Book VIII, chapter 12; Jerome, *Dialogi contra Pelagianos,* Book I, chapter 29). Inventories of Frankish churches in the 9th century reveal a variety of colours used for vestments, but without any particular sequence for their use; but the *Ordo* of St. Amand of the same period refers specifically to dark vestments at the major litanies and black ones at the Feast of Purification (February 2).

The modern colour sequence of the Roman Catholic Church was first outlined in Pope Innocent III's treatise *De sacro altaris mysterio* (Book I, chapter 65, written before his election as pope in 1198), though some variations are admitted. White, as a symbol of purity, is used on all feasts of the Lord (including Maundy Thursday and All Saints') and feasts of confessors and virgins. Red is used at Pentecost, recalling the fiery tongues that descended upon the Apostles when they received the Holy Spirit, and also at feasts of the Holy Cross, Apostles, and martyrs, as symbol of their bloody passions (sufferings and deaths). Black is used as a symbol of mourning on days of fasting and penitence and at commemorations of the departed—but violet, symbolizing the mitigation of black, is allowed during Advent and Lent. Green is used on other days, without special significance, as a compromise colour distinguished from white, red, and black. Innocent's symbolism is based upon allegorical (symbolic) interpretations of colours and flowers mentioned in Scripture, especially in the Song of Solomon.

In the later Middle Ages other colours were used in various churches, such as blue for certain feasts of the Virgin Mary, and rose (a mitigation of violet) on the third Sunday in Advent and the fourth Sunday in Lent. The missal of Pope Pius V in 1570 prescribed the sequence of Innocent III, with rose on the two Sundays mentioned. In 1868 the Congregation of Rites allowed the use of gold vestments in place of white, red, and green. Medieval English uses showed much variation, but the predominant principle was use of the finest vestments, of whatever colour, on great feasts, and others on lesser days of importance. In the Use of Sarum, white, red, and blue were the primary colours; but in Lent an unbleached cloth was customary, changing to deep red during the two weeks before Easter.

Anglican and Lutheran churches have in recent times generally followed the Roman sequence, although some Anglican churches have restored the colours of the Use of Sarum. In the liturgical experiments since World War II, the sequences and symbolism inherited from the Middle Ages are being abandoned, and a greater freedom is evident in paraments (vestments and hangings), with increasing variety and combinations of colours, especially on festal occasions. (M.H.S.)

## Canon law

Canon law—which in its wider sense includes precepts of divine law, natural or positive, incorporated in various canonical collections or codes—is in this article defined as that body of rules and regulations (canons) concerning the behaviour and actions of individuals and institutions within certain Christian churches, which have, through proper ecclesiastical authority, defined and codified such rules. Though canon law is historically continuous from the early church to the present, it has, as a result of doctrinal and ecclesiastical schisms, developed differing, though often similar, patterns of codification and norms in the various churches that have incorporated it into their ecclesiastical frameworks. The canon law of the Eastern and Western churches was much the same in form until these two groups of churches separated in the Schism of 1054. In Eastern Christianity, however, because of doctrinal and nationalistic disputes during the 5th to 7th centuries, several church groups (especially non-Greek) separated themselves from the nominal head of Eastern Christianity, the patriarch of Constantinople, and developed their own bodies of canon law, often reflecting nationalistic concerns.

Canon law in the Western churches after 1054 developed without interruption until the Reformation of the 16th century. Though other churches of the Reformation rejected the canon law of the Roman Catholic Church, the Church of England retained the concept of canon law and developed its own type, which has acceptance in the churches of the Anglican Communion.

Canon law has had a long history of development throughout the Christian Era. Not a static body of laws, it reflects social, political, economic, cultural, and ecclesiastical changes that have taken place in the past two millennia. During periods of social and cultural upheaval the church has not remained unaffected by its environment. Thus, canon law may be expected to be involved in the far-reaching changes that have come to be anticipated in the modern world.

#### NATURE AND SIGNIFICANCE

A church is defined as a community founded in a unity of faith, a sacramental fellowship of all members with Christ as Lord, and a unity of government. Many scholars assert that a church cannot exist without authority—*i.e.,* binding rules and organizational structures—and that religion and law are mutually inclusive. Thus the calling of a church leader to office is regarded as important in the organizational structure and, like every other fundamental vocation in the churches that accept the validity of canon law, it is also viewed as sacramental and as linked to the priesthood—which, in turn, involves a calling to leadership in liturgy and preaching. According to Roman Catholic belief, the mission of the college of Apostles (presided over by Peter in the 1st century AD) is continued in the college of bishops, presided over by the pope. Other churches may accept this view, without at the same time accepting the authority of the pope. The validity of canon law thus rests on an acceptance of this sacramental view

Necessity of authority

and of the transmitted mission of the Apostles through the bishops.

**Historical and cultural importance of canon law.** Canon law has functioned in different historical periods in the organization of the church's liturgy, preaching, works of charity, and other activities through which Christianity was established and spread in the Mediterranean area and beyond. Canon law, moreover, had an essential role in the transmission of Greek and Roman jurisprudence and in the reception of Justinian law (Roman law as codified under the sponsorship of the Byzantine emperor Justinian in the 6th century) in Europe during the Middle Ages. Thus it is that the history of the Middle Ages, to the extent that they were dominated by ecclesiastical concerns, cannot be written without knowledge of the ecclesiastical institutions that were governed according to canon law. Medieval canon law also had a lasting influence on the law of the Protestant churches. Numerous institutions and concepts of canon law have influenced the secular law and jurisprudence in lands influenced by Protestantism: *e.g.,* marriage law, the law of obligations, the doctrine of modes of property acquisition, possession, wills, legal persons, the law of criminal procedure, and the law concerning proof or evidence. International law owes its very origin to canonists and theologians, and the modern idea of the state goes back to the ideas developed by medieval canonists regarding the constitution of the church. The history of the legal principles of the relation of *sacerdotium* to *imperium*—*i.e.,* of ecclesiastical to secular authority or of church to state—is a central factor in European history.

**Problems in the study of canon law and its sources.** Because of the discontinuity that has developed between church and state in modern times and the more exclusively spiritual and pastoral function of church organization, scholars in canon law are searching for a recovery of vital contact among canon law and theology, biblical exegesis (critical interpretive principles of the Bible), and church history in their contemporary forms. Canon-law scholars are also seeking a link with the empirical social sciences (*e.g.,* sociology, anthropology, and other such disciplines), which is required for insight into and control of the application of canon law. The study of the history of canon law calls not only for juridical and historical training but also for insight into contemporary theological concepts and social relationships. Many sources, such as the documents of councils and popes, are often uncritical and found only in badly organized publications, and much of the material exists only in manuscripts and archives; frequently the legal sources contain dead law (*i.e.,* law no longer held valid) and say nothing about living law. What does and does not come under canon law, what is or is not a source of canon law, which law is universal and which local, and other such questions must be judged differently for different periods.

The role and function of canon law

The function of canon law in liturgy, preaching, and social activities involves the development and maintenance of those institutions that are considered to be most serviceable for the personal life and faith of members of the church and for their vocation in the world. This function is thus concerned with a continual adaptation of canon law to the circumstances of the time as well as to personal needs.

## HISTORY

**The formative period: origins to Gratian (c. 1140).** The early church was not organized in any centralized structure. Over a long period of time, there developed patriarchates (churches believed to have been founded by Apostles) and bishoprics, the leaders of which—either as monarchical bishops or as bishops with shared authority (*i.e.,* collegiality)—issued decrees and regulatory provisions for the clergy and laity within their particular jurisdictions. After the emperor Constantine granted tolerance to Christians within the Roman Empire, bishops from various sees—especially from the eastern part of the empire—met in councils (*e.g.,* the ecumenical Council of Nicaea). Though these councils are known primarily for their consideration of doctrinal conflicts, they also ruled on practical matters (such as jurisdictional and institutional concerns), which were set down in canons. In the West, there was less imperial interference, and the bishop of Rome (the pope) gradually assumed more jurisdictional authority than his counterpart (the ecumenical patriarch of Constantinople) in the East. Throughout this period there were often conflicting canons, since there were many independently developed canonical collections and no centralized attempt to bring order out of the many collections until the Middle Ages.

*Eastern churches.* In addition to the New Testament, the writings of the Apostolic Fathers (second generation of Christian writers) and the pseudo-apostolic writings (documents attributed to but not written by the Apostles) contain the oldest descriptions of the customs existing in the East from the 2nd century until the 5th. The sources of all the others are the *Doctrina duodecim Apostolorum* (*Doctrine of the Twelve Apostles,* 2nd century?), the *Didascalia Apostolorum* (*Teaching of the Apostles,* 3rd century), and the *Traditio Apostolica* (*Apostolic Tradition*), attributed to Hippolytus, written in Rome about AD 220 but far more widely distributed in the East. From these documents, the *Constitutiones Apostolicae* (*Apostolic Constitutions*), in which 85 *Canones Apostolicae* (*Apostolic Canons*) were included, were composed about AD 400.

During the period that followed Constantine's grant of religious toleration, many synods held in the East legislated, among other things, various disciplinary rules, or *canones.* In addition to and in place of the law of custom, written law entered the scene. An ecumenical Council of Chalcedon (AD 451) possessed a chronological collection of the canons of earlier councils. This *Syntagma canonum* ("Body of Canons"), or *Corpus canonum orientale* ("Eastern Body of Canons"), was subsequently complemented by the canons attributed to other 4th- and 5th-century councils, canonical letters of 12 Greek Fathers and of the 3rd-century Latin bishop of Carthage, Cyprian, and the *Constitutiones Apostolicae.* With the exception of the last, the Trullo (supplementary) Council of Quinisextum, or the fifth and sixth councils (692), accepted this complex, along with its own canons, as the official legal code of the Eastern churches. The canons of the second ecumenical Council of Nicaea (787) and of the two councils (861 and 879–880) under Photius, patriarch of Constantinople, were added to that.

Systematic collections and the science of canon law

The systematic collections—and there were many of them—contained either canons of councils or ecclesiastical laws (*nomoi*) of the emperors or both together (nomocanons). The first known Greek collection of canons that is preserved is the *Collectio 50 titulorum* ("Collection of 50 Titles"), after the model of the 50 titles of the work known as the *Pandecta* ("Accepted by All") composed by the patriarch John Scholasticus about 550. He composed from the Novels (*Novellae constitutiones post Codicem*) of Justinian the *Collectio 87 capitulorum* ("Collection of 87 Chapters"). The *Collectio tripartita* ("Tripartite Collection"), from the end of the 6th century and composed of the entire Justinian ecclesiastical legislation, was the most widely distributed. The nomocanons were expressions of the fusion of imperial and church authority. The *Nomocanon 50 titulorum* ("Canon Law of 50 Titles") from about 580, composed of the works of John Scholasticus, remained in use until the 12th century. The edition of the *Nomocanon 14 titulorum* ("Canon Law of 14 Titles") was completed in 883 and accepted in 920 as law for the entire Eastern Church.

The science of canon law was pursued together with the study of secular law, especially in the schools in Constantinople and Beirut. The *Scholia* (commentaries) on the *Basilica,* a compilation of all of imperial law from the time of Justinian, promulgated by the Byzantine emperor Leo VI (reigned 886–912), influenced the method of commenting on and teaching canon law. The best-known commentators in the 12th century were Joannes Zonaras and Theodore Balsamon. Matthew Blastares composed his *Syntagma alphabeticum* ("Alphabetical Arrangement"), an alphabetic manual of all imperial and church law, in 1335 from their works.

*Independent churches of Eastern Christianity.* The churches of Eastern Christianity that separated from the

patriarchal see of Constantinople over a period of several centuries, but primarily during the 5th and 6th centuries, developed bodies of canon law that reflected their isolated and—after the Arab conquests in the 7th century—secondary social position. Among these churches are the Syrian Orthodox Patriarchate of Antioch (in Syria), the Ancient Church of the East (the Assyrians), the Armenian Apostolic Church, and the Coptic Orthodox Church (in Egypt). Another independent church is the Ethiopian Orthodox Church.

Though these churches developed an extensive body of canon law throughout their histories, Western knowledge of their canon law has been very scant. In the 20th century, however, more than 300 manuscripts dealing with canon law were found in various isolated monasteries and ecclesiastical libraries of the Middle East by Arthur Vööbus, an Estonian-American church historian. These manuscripts cover the period from the 3rd to the 14th century and deal with ecclesiastic regulations of the Syrian churches. Included among these manuscripts are the following: "The Canons of the Godly Monastery of St. Mār Mattai" (630), 26 in number, concerning the jurisdiction of the metropolitan (an archbishop) over the monastery; "The Canons of the Holy Qyriaqos, Which the Patriarch Composed and the Synod of the Saints and Bishops with Him" (794), containing 46 canons dealing with ecclesiastical and moral discipline and with liturgical, cultic, and monastic matters; and "The Canons Which Were Composed by the Holy Synod Which Assembled in Bēt Mār Šīlā [in the region] of Serūg, and Which Consecrated Mār Dionysios as Patriarch of Antioch, the City of God" (896), which originally contained 40 canons, though only 25 remain, dealing with the election and examination of candidates for the hierarchy and clergy, the conduct of priests, marriage, pagan influences, and religious and ecclesiastical duties. These canonical collections come from the West Syrian churches. Other canonical collections of the East Syrian churches were published in the early part of the 20th century.

*Western churches.* From about 300 until about 550, canon law in Western churches had a certain unity through the acceptance of the Eastern and North African councils and the binding factor of the papal decretal law (answers of popes to questions of bishops in matters of discipline), which did not exist in the East. The African canons, like the Eastern canons at Chalcedon, were read out at the councils of Carthage and, if confirmed, included in the Acts, which contained the newly enacted canons. Thus, at the third Council of Carthage (397), the Compendium of the Council of Hippo (393) was included. The collection of the 17th Council of Carthage (419) was soon accepted in all of the East and West. In Spain the canons of Nicaea I (325) and Chalcedon (451) and also African and south Gallican canons and Roman decretals were taken over, as well as their own canons, but the later *Hispana* (Spanish collection) crowded out all earlier collections. The Council of Elvira (295–314) in Spain was the first that set up a more complete legislation, followed by Gaul in the first Council of Arles in 314. Texts from the East, Spain, and Rome, including the *Collectio Quesnelliana* (an early 6th-century canonical collection named for its publisher, the 17th-century Jansenist scholar Pasquier Quesnel), circulated there. Gennadius, a priest from Marseille, in about 480 wrote the *Statuta ecclesiae antiqua* ("Ancient Statutes of the Church"), principally inspired by the *Constitutiones Apostolicae.* A tendency toward the unification of canon law revealed itself most clearly in Italy against the disintegrating situation that existed between the Eastern and Western churches—*i.e.,* the so-called Acacian Schism (484–518), occasioned by the patriarch Acacius of Constantinople and the emperor Zeno's neglect of the legislation of the Council of Chalcedon—and the breakup of the Western Empire soon after the fall of Rome (476), at the time of the 30-year "Gelasian renaissance," beginning during the reign of Pope Gelasius I. There also existed in Rome translations of Eastern councils: *Vetus Romana, Versio Hispana* ("Ancient Roman, Spanish Version"), *Isidoriana, versio Prisca* ("The Isidorian, Priscan Version"), and *Itala* ("Italian"). By far the most important is that of the *Liber canonum* ("Book of Canons") of the 6th-century Roman theologian Dionysius Exiguus, about 500. The first two versions contain 50 *Canones Apostolorum,* Greek canons, and the African canons of the 17th Council of Carthage. Dionysius Exiguus also composed a *Liber decretorum* ("Book of Decretals") from Pope Siricius to Pope Anastasius II. Together, the books form the *Corpus* ("Body") or *Codex canonum* ("Code of Canons").

Development of canon law in western Europe

Until the end of the 7th century a greater decentralization and less mutual contact occurred in the separate German kingdoms. Elements of German law found their way into Roman canon law. The *Collectio Avellana* ("Avellan Collection"), written in Rome about 555, which was a Western nomocanon; the *Collectio Novariensis* ("Novarien Collection"); and the *Epitome Hispanica* ("Spanish Abridgment") entered Italy from Spain. In Africa, the first, albeit primitive, systematic collections appeared. These included the *Breviatio canonum* ("Abridgment of Canons") of (Fulgentius) Ferrandus, deacon of the Church of Carthage (*c.* 546), and the *Concordia canonum Cresconii* ("Harmony of the Canons of Cresconius," a 6th- or 7th-century author), a systematic compilation of the *Dionysiana,* subsequently found in different manuscripts in Gaul. There the collections were local ones: every cathedral and monastery had its own *liber canonum.* The church of Arles, the metropolis of southern Gaul, had the *Liber auctoritatum* ("Book of Authorities"; *i.e.,* legal texts), a nomocanon of its privileges. The first systematic Gallic *Collectio Andegavetis* ("Andegavenan Collection"), from the end of the 7th century, was an attempt to unite the ancient law with the native.

In Spain, after the conversion of King Recared in 587, the church of the Visigothic kingdom became a well-knit national church with a classical provincial structure under metropolitan jurisdiction, closely linked to the crown. The national councils of Toledo preserved the unity of law and respect for the ancient law. The *Capitula* ("Chapters") of Martinus, bishop of Braga (*c.* 563), was included completely in the *Hispana* and was also copied outside Spain. The *Collectio Novariensis* was related to the *Epitome Hispanica,* the code of the hierarchy that was temporarily halted at the fourth Council of Toledo (633). The *Hispana* was recognized by popes Alexander III and Innocent III as the authentic *corpus canonum* of the Spanish church. Shortly before the *Hispana,* systematic indices (called *tabula*) were written and were subsequently expanded into *excerpta* ("excerpts") and finally into complete texts, the *Hispana systematica* ("Systematic Spanish [Code]"). After the 10th century, the *Hispana* was also called the *Isidoriana,* attributed to Isidore of Seville, a Spanish encyclopaedist and theologian, who was the author of the *Etymologiae* ("Etymologies"), a universally distributed early medieval book of doctrine.

The most disparate picture is offered by the church in the British Isles. The church there was concentrated around heavily populated monasteries, and discipline outside them was maintained by means of a new penitential practice. In place of ancient canons about public penance, the clergy and monks used *libri poenitentiales* ("penitential books"), which contained detailed catalogs of misdeeds with appropriate penances. They were private writings without official authority and with very disparate content. From the monasteries founded in Europe by the Irish monk St. Columban and missionaries of Anglo-Saxon background, the *libri poenitentiales* spread throughout the continent, where once again new versions emerged. The *Collectio Hibernensis* ("Hibernian [or Irish] Collection") of about 700, used texts from Scripture, mainly from the Old Testament, for the first time in canonical collections, and texts from the Greek and Latin early Church Fathers in addition to canons. The *Liber ex lege Moysi* ("Book from the Law of Moses"), an Irish work, drew exclusively from the Pentateuch.

The Carolingian reform

The reorganization of the Frankish church began with the Carolingian reform in the middle of the 8th century. The canon law was set down especially in the *Capitularia ecclesiastica* ("Ecclesiastical Articles") of the prince, as well as in the *Capitularia missionum* ("Mission Articles"; *i.e.,* instructions given by the prince to the bishops and

abbots who visited in his name). The *Capitularia* ("Short Articles") of Charlemagne, the founder of the Holy Roman Empire, and his son, the emperor Louis the Pious, were collected in 827 by the abbot Ansegisus. Following this model the bishops composed terse *capitula,* the oldest known diocesan statutes, for their clergy. The penance books were condemned and replaced by new ones that were more closely related to tradition. The reception of the *Dionysiana* and the *Hispana* is of importance for the transmission of the text and for the Carolingian cultural renaissance. In 774 Charlemagne received from Pope Adrian I a completed *Dionysiana,* the *Dionysiana-Hadriana,* which was accepted at a national synod in Aachen in 802 but never was adopted as an official national code. About 800 the *Hadriana* and the *Hispana* were developed into a systematic whole, the *Dacheriana* (canonical collection named for its 17th-century publisher, a French scholar, Jean-Luc d'Achéry)—the principal source of the collections before 850—which was of influence until the Gregorian reform in the 11th century.

After Louis the Pious, the central power among the Franks was increasingly divided among counts and barons. German law—which linked the right to govern with land ownership, without distinction between public and private law—expressed itself in the medieval forms of the system of private churches. This northern law looked upon dioceses, churches, and monasteries—with their rights and privileges—as lucrative possessions that deserved to be confiscated, by fraudulent means if necessary.

Such situations became the occasion in about 850 for the massive falsifications (*i.e.,* forgeries) of the pseudo-Isidorian collections: the *Hispana Augustodunensis* ("Spanish Collection of Autun"), the *Capitula Angilramni* ("Chapters of Angilramnus," bishop of Metz), the *Capitularia Benedicti Levitae* ("Frankish Imperial Laws of Benedict the Levite," a fictitious name), and the *Pseudo-Isidorian Decretals.* The central goal of the anonymous Frankish group of authors of these collections was to strengthen the position of the bishops and to rectify the poor condition of ecclesiastical–state affairs. This was accomplished by means of falsified and forged texts that were attributed to the esteemed authority of the old law (*i.e.,* the popes) and the Carolingian princes. They did not have much influence on the real development of canon law, although later collections drew from them abundantly. Only the Magdeburg Centuriators, authors of the *Centuries,* a 16th-century Lutheran church history, denied the genuineness of all the decretals of pseudo-Isidore; the lack of authenticity of the other three works was discovered later.

Several collections appeared before AD 1000. About 882 decretals were organized in the *Collectio Anselmo dedicata* ("Collection Dedicated to Anselm"), a papally oriented, systematic work from northern Italy. In Germany the *Libri duo de synodalibus causis et disciplinis ecclesiasticis* ("Two Books Concerning Synodical Causes and Church Discipline") of Regino, abbot of Prüm (906), was a bishops' manual for the judicial interrogation of jurymen during a visitation; and in France appeared the collection of Abbon, abbot of Fleury (*c.* 996), which defended the legal position of his monastery against the king and bishop. Intended as a doctrinal book for the young cleric, the Decree of Burchard—bishop of Worms from 1000 to 1025—became the canon-law manual in the cathedral schools and in the curias (administrative bureaucracies) of bishops and abbots in Germany, France, and Italy. Burchard was a promoter of moderate imperial reform. He did not reject the system of private churches; he only rejected the misuses proceeding from it, such as simony (buying or selling church offices) and the violation of celibacy.

The Gregorian reformation

The slogans of the Gregorian reformation, initiated by Pope Gregory VII (reigned 1073–85), were *libertas Ecclesiae* ("liberty of the church") and *puritas Ecclesiae* ("purity of the church"). These slogans advocated freedom from the system of private churches on all levels; freedom from papal dependence on the Roman nobility and emperor; freedom from dependence of the village priest on his *senior* (the beginning of the fight against investiture); and purity from simony and from the total collapse of celibacy (which was exhibited in the practice of hereditary parishes and bishoprics). Fundamental principles of Gregorian canon law included those stipulating that only canon law that is given or approved by the pope is valid; papal legates (representatives) stand above the local hierarchies and preside over synods; for possession of every ecclesiastical office, choice and appointment by church authorities is demanded, along with the exclusion of lay investiture; every form of simony makes the appointment invalid; and the faithful must boycott the services of married priests. New material was sought, especially for the confirmation of papal primacy, in archives and libraries. The principal new sources were the *Breviarium* of Cardinal Atto (*c.* 1075), the *Dictatus Papae* ("Dictates of the Pope") of Gregory VII (*c.* 1075), the *Collectio 74 titulorum,* or "Collection of 74 Titles" (1074–76), the collection of Bishop Anselm of Lucca (*c.* 1083) and that of Cardinal Deusdedit (*c.* 1085), and the *Liber de vita Christiana* ("Book Concerning the Christian Life") of Bonizo, bishop of Sutri (*c.* 1090).

The investiture battle over the conflicting asserted rights of lay or ecclesiastical officials to invest a church official with the symbols of his spiritual office ended in France, England, and Germany (Concordat of Worms, 1122) in compromises. Gregorian law, which now seems too strict, had to be reconciled with the established traditions. Ivo, bishop of Chartres from 1091 to 1116, contributed to the settlement of the investiture problem by his political activities; his extended correspondence; and his three law collections, *Tripartita* ("Tripartite Collection"), *Decretum* ("Decrees"; *i.e.,* collection of decrees or canons), and *Panormia* (collections of "All the Laws"), the last two practically a fusion of Burchard's Decree with Gregorian law. The famous Prologue, written by Ivo for either the *Decretum* or the *Panormia,* indicated for the first time a method by which the bishop must handle the conflicting strict and liberal texts, with *justitia* ("justice") or *misericordia* ("mercy"). Bernold of Constance, in his little tractates, written between 1070 and 1091, listed several criteria for the reconciliation of conflicting texts, including authenticity of the text; identity of the author; difference between law, counsel, and dispensation, between universal and local law, of time and place; and different meanings of a word. A Liège (Belgium) canon lawyer, Alger, in his *Liber de misericordia et justitia,* or "Book Concerning Mercy and Justice" (*c.* 1105), applied Ivo's criteria to the problem of the effect of sacraments administered by heretics and persons guilty of simony. The great medieval theologian Abelard developed the method of reconciling texts that are for or against a theological position in his *Sic et non,* or "Yes and No" (1115–17). The same methods were applied by the first writers of glosses (commentaries or interpretations) at the law school in Bologna on the *Pandecta* of Justinian, which was rediscovered in about 1070.

**The Corpus Juris Canonici (c. 1140–c. 1500).** About 1140 the monk John Gratian completed his *Concordia discordantium canonum* ("Harmony of Contradictory Laws"), later called the *Decretum Gratiani* ("Gratian's Decree"); it became not only the definitive canonical collection of the entire preceding tradition but also a systematic application of the scholastic method to all legal material. The *Decretum* dealt with the sources of the law, ordinations, elections, simony, law of procedure, ecclesiastical property, monks, heretics, schismatics, marriage, penance, and sacraments and sacramentals. Primitive as it was, it provided a foundation for systematic compilation of the legal material by the canonists and for the expansion of decretal law. It provided a basis for the education in canon law that began in the schools of Bologna, Paris, Orléans, Canterbury, Oxford, Padua, and elsewhere. It was accepted everywhere in the ecclesiastical administration of justice and government.

Collections of decretals

From the time that the Gregorian reformation introduced a more centralized ecclesiastical administration, the number of appeals to Rome and the number of papal decisions mounted. New papal laws and decisions, called decretals, first added to Gratian's *Decretum,* were soon gathered into separate collections, of which the best known are the *Quinque compilationes antiquae* ("Five Ancient Compilations"). The first, the *Breviarium extravagantium*

("Compendium of Decretals Circulating Outside"; *i.e.,* not yet collected) of Bernard of Pavia, introduced a system inspired by the codification of Justinian, a division of the material into five books, briefly summarized in the phrase *judex* ("judge"), *judicium* ("trial"), *clerus* ("clergy"), *connubium* ("marriage"), *crimen* ("crime"). Each book was subdivided into titles and these in turn into *capitula,* or canons. This system was taken over by all subsequent collections of decretals. These compilations were the foremost source of the *Liber extra* ("Book Outside"; *i.e.,* of decretals not in Gratian's Decree) or *Liber decretalium Gregorii IX* ("Book of Decretals of Gregory IX"), composed by Raymond of Peñafort, a Spanish canonist, and promulgated on Sept. 5, 1234, as the exclusive codex for all of canon law after Gratian. On March 3, 1298, Pope Boniface VIII promulgated *Liber sextus* ("Book Six"), composed of official collections of Innocent IV, Gregory X, and Nicholas III, and private collections and decretals of his own, as the exclusive codex for the canon law since the *Liber extra.* The *Constitutiones Clementinae* ("Constitutions of Clement") of Pope Clement V, most of which were enacted at the Council of Vienne (1311–12), were promulgated on Oct. 25, 1317, by Pope John XXII, but they were not an exclusive collection. The *Decretum Gratiani,* the *Liber extra, Liber sextus,* and *Constitutiones Clementinae,* with the addition of two private collections, the *Extravagantes* of John XXII and the *Extravagantes communes* ("Decretals Commonly Circulating"), were printed and published together for the first time in Paris in 1500. This entire collection soon received the name *Corpus Juris Canonici* ("Corpus of Canon Law").

The science of canon law was developed by the writers of glosses, the commentators on the Decree of Gratian (decretists), and the commentators on the collections of decretals (decretalists). Their glosses were based on the system used by Gratian: next to the texts of canons parallel texts were noted, then conflicting ones, followed by a *solutio* ("solution"), again with text references. In connection with this the glosses of other canonists were also introduced. In this way the *apparatus glossarum,* continuous commentaries on the entire book, arose. The *glossa ordinaria* ("ordinary explanation") on the different parts of the *Corpus Juris Canonici* was the apparatus that was used universally in the schools. After the classical period of the glossators (12th–14th century), terminated by the work of a lay Italian canonist, John Andreae (*c.* 1348), followed that of the post-glossators. In the absence of new legislation in the time of the Babylonian Captivity (1309–77), when the papacy was situated at Avignon, Fr., and the Great Schism (1378–1417), when there were at least two popes reigning simultaneously, the commentaries on decretals continued, but with a larger production of special tracts; *e.g.,* regarding the laws of benefices and marriage and of *consilia* (advice about concrete legal questions).

**From the Council of Trent (1545–63) to the Codex Juris Canonici (1917).** *The end of decretal law.* Toward the end of the Middle Ages decretal law ceased to govern. Medieval Christian society became politically and ecclesiastically divided, according to the principle of *cujus regio, ejus religio* (Latin: "whose region, his religion"; *i.e.,* the religion of the prince is the religion of the land). In Protestant areas the former Roman Catholic church buildings and benefices were taken over by other churches; and even in the lands that remained Catholic the churches found themselves in an isolated position as secularization forced the churches to reorganize. With the end of feudalism, canon law dealing with benefices, chapters, and monasteries, which were closely bound to the feudal structure, changed. The territorial, material, and economic character of canon law and the decentralization allied with it disappeared. The decision of the reform councils from Pisa (1409) until the fifth Lateran Council (1512–17) affected, in particular, benefices, papal reservations, taxes, and other such ecclesiastical matters. In the same period various concordats (agreements) permitted the princes to intervene in the issue of ecclesiastical benefices and property. Canon law took on a more defensive character, with prohibitions regarding books, mixed marriages, participation of Roman Catholics in Protestant worship and vice versa, education of the clergy in seminaries, and other such areas of concern.

At the Roman Catholic reform Council of Trent (1545–63) a new foundation for the further development of canon law was expressed in the *Capita de reformatione* ("Articles Concerning Reform"), which were discussed and accepted in 10 of the 25 sessions. Papal primacy was not only dogmatically affirmed against conciliarism (the view that councils are more authoritative than the pope) but was also juridically strengthened in the conduct and implementation of the council. The central position of the bishops was recovered, over against the decentralization that had been brought about by the privileges and exemptions of chapters, monasteries, fraternities, and other corporate bodies that sprang from Germanic law, as well as caused by the rights granted to patrons. In practically all matters of reform the bishops received authority *ad instar legati S. Sedis* ("like delegates of the Holy See"). Strict demands were made for admission to ordination and offices; measures were taken against luxurious living, nepotism, and the neglect of the residence obligation; training of the clergy in seminaries was prescribed; prescriptions were given about pastoral care, schools for the young, diocesan and provincial synods, confession, and marriage; the right to benefices was purified of misuse; and the formalistic law of procedure was simplified.

The council gave the duty of execution of the reform to the pope. On Jan. 26, 1564, Pius IV confirmed the decisions, reserved to himself their interpretation and execution, and on Aug. 2, 1564, established the Congregation of the Council for that purpose. The congregations of cardinals, which proceeded from the former permanent commissions of the *consistorium* (the assembly of the pope with the College of Cardinals), were organized by Pope Sixtus V in 1587. Since then the administrative apparatus of the Curia has consisted of congregations of cardinals together with courts and offices. This apparatus made it possible for the Latin Church to acquire a uniform canon law system that was developed in detail.

The congregations of cardinals

*Law for the missions.* Expansion of the church brought with it expansion of the ordinary hierarchical episcopal structure. This was true also for the new colonies under the right of patronage of the Spanish and Portuguese kings. In the other mission areas and in the areas taken over by the Protestants, where the realization of the episcopal structure and the decretal law adopted by Trent was not possible, the organization of mission activity was taken from missionaries and religious orders and given to the Holy See. The Sacred Congregation for Propagation of the Faith (the Propaganda) was established for this purpose in 1622. Missionaries received their mandate from Rome; the administration was given over to apostolic vicars (bishops of territories having no ordinary hierarchy) and prefects (having episcopal powers, but not necessarily bishops) who were directly dependent on the Propaganda, from which they received precisely described faculties. A new, uniform mission law was created, without noteworthy native influence; this sometimes led to conflict, such as the Chinese rites controversy in the 17th and 18th centuries over the compatibility of rites honouring Confucius and ancestors with Christian rites.

The first Vatican Council (1869–70) strengthened the central position of the papacy in the constitutional law of the church by means of its dogmatic definition of papal primacy. Disciplinary canons were not enacted at the council; but the desire expressed by many bishops that canon law be codified did have influence on the emergence and content of the code of canon law.

**The Codex Juris Canonici (1917).** Since the closing off of the *Corpus Juris Canonici* there had been no official or noteworthy private collection of the canon law, except for the constitutions of Pope Benedict XIV (reigned 1740–58). The material was spread out in the collections of the *Corpus Juris Canonici* and in the generally very incomplete private publications of the *acta* of popes, of general and local councils, and the various Roman congregations and legal organs, which made canon law into something unmanageable and uncertain. The need for codification was recognized even more because of the fact that since

the end of the 18th century, secular law had undergone a period of great codification. Several private attempts to do this had met with little success.

On March 19, 1904, Pius X announced his intention to complete the codification, and he named a commission of 16 cardinals, with himself as chairman. Bishops and university faculties were asked to cooperate. The schemata of the five books that were prepared in Rome—universal norms, personal law, law of things, penal law, and procedural law—were proposed in the years 1912–14 to all those who would ordinarily be summoned to an ecumenical council, and with their observations were then reworked in the cardinals' commission. The entire undertaking and all the drafts were under the papal seal of secrecy and were not published. Meanwhile, Pius X introduced various reforms that were to a great degree the results of the commission's work. In July 1916 the preparations for the *Codex Juris Canonici* ("Code of Canon Law") were completed. The code was promulgated on Pentecost Sunday, May 27, 1917, and became effective on Pentecost Sunday, May 19, 1918.

Reorganization of the Code of Canon Law

In contrast to all earlier official collections this code was a complete and exclusive codification of all universal church law then binding in the Latin Church. Out of fear of political difficulties, a systematic handling of public church law, especially what concerned the relations between church and state, was omitted. Its main purpose was to offer a codification of the law, and only incidentally adaptation, and so it introduced relatively little that was new legislation. The 2,414 canons were divided into five books that no longer followed the system of the collections of decretals but did follow that of the Perugian canonist Paul Lancelotti's *Institutiones juris canonici* (1563; "Institutions of Canon Law"), which in turn went back to the division of the 2nd-century Roman lawyer Gaius' *Institutiones*—one section on persons, two sections on things, and one section on actions—and was based on the fundamental idea of Roman law; *i.e.,* subjective right. In some editions the sources that were used by the editors were indicated at the individual canons. With the publication of the codex these sources belonged to the history of the law. Older general and particular law, in conflict with the codex, was given up and, insofar as it was not in conflict with it, served only as a means for interpreting the code. The old law of custom in conflict with the code and expressly reprobated by it was rendered null; when not reprobated and 100 years old or immemorial it could be allowed by ordinaries for pressing reasons. Acquired rights and concordats in force remained in force. With this change, an independent science of the history of canon law became necessary, in addition to the dogmatic canonical science of canon law on the basis of the code.

In order to ensure the unity of the codification and the law, a commission of cardinals was established on Sept. 15, 1917, for the authentic interpretation of the new code. At the same time it was decided that the cardinals' congregations should no longer make new general decrees but only instructions for the carrying out of the prescriptions of the code. Should a general decree appear necessary, it was determined, the commission would formulate new canons and insert them into the code. Neither of these decisions was carried out. Only two canons were altered and congregations promulgated numerous general decrees. New papal legislation complemented and altered the law of the code.

**The Eastern churches in union with Rome.** Catholic Eastern churches (churches in union with the Roman Catholic Church) retain their own traditions in liturgy and church order, insofar as these are not considered to be in conflict with the norms taken by Rome to be divine law. In 1929 Pius XI set up a commission of cardinals for the codification of canon law valid for all Uniate churches in the East. In the following year a commission was established for the preparation of the codification and one for the collection of the sources of Eastern law, in which experts of all rites were involved. These collections were published in three series, begun respectively in 1930, 1935, and 1942.

In 1935 the preparatory commission became the Pontifical Commission for the Redaction of the *Codex Juris Canonici Orientalis* ("Code of Oriental Canon Law"). The cooperation of all Eastern ordinaries (bishops, patriarchs, and others having jurisdictions) was requested, and the drafts of the various documents were sent to them. Thereafter four parts were published: in 1949, on marriage law; in 1952, on the law for monks and other religious, on ecclesiastical properties, and a title *De Verborum Significatione* ("Concerning the Meaning of Words"), a series of definitions of legal terms used in the canons; and in 1957, on constitutional law, especially of the clergy. The still incomplete codification followed the Latin code with the assimilation of the authentic interpretation and with textual corrections, and also with the insertion of the general law proper to the Eastern churches, including the Orthodox churches, regarding the patriarchs and their synods, marriage law, the law of religious, and other matters. The promulgation was made only in Latin in the *Acta Apostolicae Sedis,* the official organ of the Holy See. The Catholic Eastern churches came under the Congregation for the Eastern Churches that was established on Jan. 6, 1862, by Pius IX as part of the *Propaganda Fide,* and was made independent by Benedict XV on May 1, 1917, and expanded considerably by Pius XI on March 25, 1938. Roman legislation as well as the jurisdiction of a congregation of the Roman Curia was criticized as being incompatible with the traditional autonomy of the Eastern churches in legislation and administration.

## THE SECOND VATICAN COUNCIL AND POST-CONCILIAR CANON LAW

**Vatican II.** Fundamental to the development of canon law in the Roman Catholic Church is the second Vatican Council's (Oct. 11, 1962–Dec. 8, 1965) vision of the church as the people of God. In this connection the former concept of the church as *societas perfecta* (the "perfect society"), founded by Christ through the mission of the Apostles and their successors, to which one belongs through subjection to the hierarchy, is replaced by a vision of the church as a community in which all possess the sacramental mission to live and proclaim the Gospel, and all have a function in the service of the whole. The legislative and administrative functions remain related to the hierarchy, but this is much more expressly seen as a service for the religious life of the community. The idea of collegiality, resting on the recognition of the vocation received by each one from the Lord, works itself out in the relationship existing among the bishops and with the pope, of the bishops with the clergy, and of the clergy with the laity. Related to this is a tendency to co-responsibility and the democratization of the church structure and also an autonomy for the laity to exercise individually and collectively the Christian mission proper to them; namely, to bring the spirit of Christ into the secular life of mankind. The right of clergy and laity to a share in the leadership of bishops and pope is recognized. The vision of the people of God as *sacramentum mundi,* a sign of redemption for all mankind, gave a new insight into the relationships with the Protestant churches, the other world religions, and the nonreligious atheistic and humanistic movements. In this view, freedom of religion and philosophy became the most fundamental right of humanity.

The church and collegiality

**Post-conciliar legislation.** From a schematically chronological survey of the principal conciliar and post-conciliar legislation a new era apparently began for canon law. In 1960, the Secretariat for Promoting Christian Unity was established. Three years later various faculties, previously reserved to Rome, were given to the bishops; and in 1964, actions were undertaken for the reorganization of the papal commission for communications media, establishment of the Secretariat for Non-Christians, and lifting of the prohibition against cremation. Other legislative changes indicating a new era included several regulations that could not have been proposed with any possibility of their being accepted prior to Vatican II. In 1965, for example, preeminence in the College of Cardinals was given to Eastern patriarchs, after deacon and subdeacon and after the cardinals of the dioceses of the province of Rome; in the same year the Secretariat for Non-Believers

Canon law after 1960

was established and the Holy Office (formerly the Inquisition) became the Congregation for the Doctrine of the Faith, with emphasis on the positive fostering of theological research. In 1966, greatly reduced prescriptions for fasting and abstinence were adopted, the Index of prohibited books became a moral guide instead of obligatory law, and in implementation of the conciliar degree on the episcopal office, the principle according to which ordinaries (*e.g.,* bishops) dispense from universal laws only when this is allowed by law or special faculties was replaced by the principle that ordinaries can always dispense unless it is explicitly reserved to Rome—and such reserved dispensations in question are indicated.

In addition to these changes, further canonical regulations were accepted. New regulations for mixed marriages were adopted in 1966. Norms were established for the implementation of the conciliar decrees on the office of bishops and priests; missionary activity; personal and material aid to needy churches; introduction of priests' councils and pastoral councils of priests, religious (*i.e.,* monks and nuns), and laity as advisory groups for bishops; international episcopal conferences and their mutual relationships; and other concerns. From 1967 to 1970 more changes were made in canonical regulations—*e.g.,* in 1967, total revision of the norms for indulgences, establishment in the Roman Curia of the council of laymen and the study commission Justitia et Pax ("Justice and Peace"), new dispensation rights for Eastern bishops, directory for ecumenical cooperation with Christian churches, regulation of the office of the diaconate to include married men, and reorganization of the Roman Curia; in 1970, a mandate to the secretary of state to discuss with the world episcopacy the question of celibacy and ordination of married men in areas that need priests.

Characteristics of the new regulations included a searching for structures to allow all members of the church to have a voice in ecclesiastical decision making and decentralization and autonomy of local churches. Regulations from Rome were kept to the general, with ample room for local adaptation. In addition, new regulations were to be enacted only after extensive and open inquiry and test by experience, with possibilities for experimentation. In place of regulations of religious behaviour, canon law was becoming an ordering of the cooperation of all members of the Roman Catholic Church for the realization of its mission in the world.

**Revision of the Code of Canon Law.** On Jan. 25, 1959, John XXIII announced the revision of the church's code. On March 28, 1963, he set up a commission of cardinals for that purpose. On April 17, 1964, Paul VI named the first consultants. No publicity was given to the commission's work, but the first episcopal synod (Sept. 30–Oct. 4, 1967) gave its approval to a document in which several principles for the revision were indicated (*Principia quae codicis juris canonici recognitionem dirigant,* or "Principles Which Guide the Recognition of the Code of Canon Law"): the juridical character of the code ought to be preserved and not, as some wished, be limited to a rule for faith and morals; canon law for the area of each one's personal conscience should be maintained, but conflicts between law for conscience and public law ought to be avoided, especially in marriage and penal law; as a means to stimulate pastoral work it was recommended that the laws be expressed in a spirit of love, fairness, and humanity; no binding prescriptions were to be given where admonition and counsel suffice; pastoral workers were to be given more discretionary powers, and greater freedom was to be given to bishops, especially in mission areas; laws were to be such that ample possibility is given for local adaptation, carrying through the principle of subsidiarity (*i.e.,* that nothing should be committed to higher organs that can be accomplished by individuals or lesser or subordinate bodies), however, with care to retain the unity of law and jurisdiction; regulation of administrative jurisdiction and in principle public jurisdiction; distinction of legislative, administrative, and judicial functions; limitation of punishments, in particular limitation of punishments incurred automatically upon commission of the offense to very few and very serious crimes. On May 28, 1968, the commission approved a preliminary division of the new codex. (P.Hu./Ed.)

Principles for revision of the code

As the drafts of the various parts of the new code became available, a vast process of consultation was initiated. The departments of the Roman Curia, the local bishops and their regional conferences, the heads of religious institutes, and university faculties of canon law were invited to evaluate the schemata and offer suggestions for their improvement. This lengthy procedure was completed in 1982.

**The new Code of Canon Law.** The second *Codex Juris Canonici* in history for the Catholics of the Latin rite was promulgated by Pope John Paul II on Jan. 25, 1983, and entered into effect on Nov. 27, 1983. It contains 1,752 canons divided among seven books. The books are: (1) "General Norms," concerning the operating principles of canon law, definitions of juridical persons, and ecclesiastical offices; (2) "The People of God," describing the rights and duties of the faithful in general and of clerics and lay persons in particular, as well as the organizational structures of the church, papacy, episcopal college, Roman Curia, particular churches, and institutes of consecrated life; (3) "The Teaching Office of the Church," concerning catechetical and missionary activities, schools, and media of communication; (4) "The Sanctifying Office of the Church," describing sacraments and worship in all its forms; (5) "The Temporal Goods of the Church," defining ownership and administration of property, contracts, and charitable foundations; (6) "Sanctions in the Church," describing various crimes, delicts, and penalties; and (7) "Procedures," outlining the administration of justice by ecclesiastical courts, various quasi-judicial actions, and remedies.

The declared intention of the drafters of the new code was to give practical effect to the theological insights of the second Vatican Council. The emphasis in the new law is on the universal people of God, and the governing power of the hierarchy is presented as a call to serve. The fundamental rights of the faithful are clearly asserted, and their active participation in the life of the church is encouraged. An effort was made toward decentralization, with local bishops enjoying more autonomy. Despite criticism from some scholars and clerics that the new code remains conservative on certain issues, it is recognized that the body of the law is permeated by an ecumenical spirit and displays a respect for the freedom of conscience and religious conviction of every human being. With the new code the hermeneutics of canon law have changed significantly. Apart from the strictly legal transactions, creating enforceable rights and duties (as in matters of property), the application of the laws must be guided and moderated according to pastoral needs.

**The Eastern churches.** A process similar to that used for the preparation of the new code for the Latin church was in progress during the 1980s for all the Eastern Catholic churches. The first draft of the new, unified code of laws was completed in 1986. It consists of 1,561 canons, organized into 30 titles. The institutions and structures of the Uniate churches are supported; the right to worship according to their own liturgical traditions is confirmed. The dignity and power of the patriarchs and of the major archbishops are recognized; the importance of synodal government at different levels is affirmed. Overall, the major themes found in this draft are the same as the ones in the Latin code (although arranged in a different order); in matters of common interest many canons are taken word-by-word from the Latin code. The process of consultation over this schema, once completed, is likely to bring about many changes in the proposed canons. The task of codifying in a single volume the laws of so many churches—having different historical memories, rooted in various cultures, and without a common language—is a daunting one, even if they all profess the same faith and are in communion with Rome. (L.M.Ö.)

The proposed Oriental Code

### ANGLICAN CANON LAW

The Anglican Communion embraces the Church of England and its affiliated churches. Since the submission of the clergy demanded by King Henry VIII and the Act of Supremacy in 1534, in which the Parliament recognized

him as supreme head of the Church of England and which was renewed by Queen Elizabeth I, the law of the English Church rests on the supremacy of the prince or of the Parliament. It is theoretically accepted that, outside the law determined by the English synods in the ancient independent national churches, only the principles of the *jus ecclesiasticum commune* ("common ecclesiastical law") are binding, but other norms, promulgated by popes and councils, are accepted only to the extent that they were accepted by English ecclesiastical or secular courts. For practical purposes the development of church law in the English Church is held by some canonists (usually Roman Catholic) to be not canon law but the ecclesiastical law of the state. The hierarchy has the power to ordain by virtue of the apostolic succession, which was preserved—according to the Anglican view—by the consecration of Matthew Parker as archbishop of Canterbury (1559), but it does not possess legislative authority. The ecclesiastical provinces are administered by convocations of Canterbury and York, consisting of an upper house of bishops and a lower house of clergy. In 1919 a Church Assembly was established by the Enabling Act; the assembly consists of three houses (of bishops, members of the convocations, and laity) with the authority to make proposals relating to any matter concerning the Church of England—with the exception of dogmas of faith—and to present these proposals to the ecclesiastical committee of Parliament. If the committee agrees on a positive report, then the Parliament can approve or reject the proposal but not amend it; if both houses of Parliament accept it, then it acquires the force of law by royal approval. Lambeth Conferences, which have been held approximately every 10 years since 1867 at the London palace of the archbishop of Canterbury and which involve all Anglican bishops from throughout the world, do not have legislative authority.

Ecclesiastical law of the state

(P.Hu./Ed.)

# ASPECTS OF THE CHRISTIAN RELIGION

## Patristic literature

Patristic literature is generally identified today with the entire Christian literature of the early Christian centuries, irrespective of its orthodoxy or the reverse. Taken literally, however, patristic literature should denote the literature emanating from the Fathers of the Christian Church, the Fathers being those respected bishops and other teachers of exemplary life who witnessed to and expounded the orthodox faith in the early centuries. This would be in line with the ancient practice of designating as "the Fathers" prominent church teachers of past generations who had taken part in ecumenical councils or whose writings were appealed to as authoritative. Almost everywhere, however, this restrictive definition has been abandoned. There are several reasons why a more elastic usage is to be welcomed. One is that some of the most exciting Christian authors, such as Origen, were of questionable orthodoxy, and others—Tertullian, for example—deliberately left the church. Another is that the undoubtedly orthodox Fathers themselves cannot be properly understood in isolation from their doctrinally unorthodox contemporaries. Most decisive is the consideration that early Christian literature exists, and deserves to be studied, as a whole and that much will be lost if any sector is neglected because of supposed doctrinal shortcomings.

Reasons for a broad concept of patristic literature

### THE ANTE-NICENE PERIOD

During the first three centuries of its existence the Christian Church had first to emerge from the Jewish environment that had cradled it and then come to terms with the predominantly Hellenistic (Greek) culture surrounding it. Its legal position at best precarious, it was exposed to outbursts of persecution at the very time when it was working out its distinctive system of beliefs, defining its position vis-à-vis Judaism on the one hand and Gnosticism (a heretical movement that upheld the dualistic view that matter is evil and the spirit good) on the other, and constructing its characteristic organization and ethic. It was a period of flux and experiment, but also one of consolidation and growing self-confidence, and these are all mirrored in its literature.

**The Apostolic Fathers.** According to conventional reckoning, the earliest examples of patristic literature are the writings of the so-called Apostolic Fathers; the name derives from their supposed contacts with the Apostles or the apostolic community. These writings include the church order called the *Didachē,* or *Teaching of the Twelve Apostles* (dealing with church practices and morals), the *Letter of Barnabas,* and the *Shepherd of Hermas,* all of which hovered at times on the fringe of the New Testament canon in that they were used as sacred scripture by some local churches; the *First Letter of Clement,* the seven letters that Ignatius of Antioch (d. *c.* 110) wrote when being escorted to Rome for his martyrdom, the related *Letter to the Philippians* by Polycarp of Smyrna (d. *c.* 156 or 168), and the narrative report of Polycarp's martyrdom; some fragmentary accounts of the origins of the Gospels by Papias (fl. late 1st or early 2nd century AD), bishop of Hierapolis in Phrygia, Asia Minor; and an ancient homily (sermon) known as the *Second Letter of Clement.* They all belong to the late 1st or early 2nd century and were all to a greater or lesser extent influenced (sometimes by way of reaction) by the profoundly Jewish atmosphere that pervaded Christian thinking and practice at this primitive stage. For this reason alone, modern scholars tend to regard them as a somewhat arbitrarily selected group. A more scientific assessment would place them in the context of a much wider contemporary Jewish-Christian literature that has largely disappeared but whose character can be judged from pseudepigraphal (or noncanonical) works such as the *Ascension of Isaiah,* the *Odes of Solomon,* and certain extracanonical texts modeled on the New Testament.

Earliest patristic writings

Even with this qualification the Apostolic Fathers, with their rich variety of provenance and genre (types), illustrate the difficult doctrinal and organizational problems with which the church grappled in those transitional generations. Important among these problems were the creation of a ministerial hierarchy and of an accepted structure of ecclesiastical authority. The *Didachē,* which is Syrian in background and possibly the oldest of these documents, suggests a phase when Apostles and prophets were still active but when the routine ministry of bishops and deacons was already winning recognition. The *First Letter of Clement,* an official letter from the Roman to the Corinthian Church, reflects the more advanced state of a collegiate episcopate, with its shared authority among an assembly of bishops. This view of authority was supported by an emergent theory of apostolic succession in which bishops were regarded as jurisdictional heirs of the early Apostles. The *First Letter of Clement* is also instructive in showing that the Roman Church, even in the late 1st century, was asserting its right to intervene in the affairs of other churches. The letters of Ignatius, bishop of Antioch at the beginning of the 2nd century, depict the position of the monarchical bishop, flanked by subordinate presbyters (priests) and deacons (personal assistants to the bishop), which had been securely established in Asia Minor.

Almost more urgent was the question of the relation of Christianity to Judaism, and in particular of the Christian attitude toward the Old Testament. In the *Didachē* there is little sign of embarrassment; Jewish ethical material is taken over with suitable adaptations, and the Jewish basis of the liturgical elements is palpable. But with *Barnabas* the tension becomes acute; violently anti-Jewish, the Alexandrian author substitutes allegorism (use of symbolism) for Jewish literalism and thus enables himself to wrest a Christian meaning from the Old Testament. The same tension is underlined by Ignatius' polemic against Judaizing tendencies in the church. At the same time all these writings—especially those of Ignatius, Polycarp, and Papias—testify to the growing awareness of a specifically

Relationship of Judaism and Christianity

Christian tradition embodied in the teaching transmitted from the Apostles.

Almost all the Apostolic Fathers throw light on primitive doctrine and practice. The *Didachē,* for example, presents the Eucharist as a sacrifice, and *I Clement* incorporates contemporary prayers. *II Clement* invites its readers to think of Christ as of God, and of the church as a preexistent reality. The *Shepherd of Hermas* seeks to modify the rigorist view that sin committed after baptism cannot be forgiven. But the real key to the theology of the Apostolic Fathers, which also explains its often curious imagery, is that it is Jewish-Christian through and through, expressing itself in categories derived from latter-day Judaism and apocalyptic literature (depicting the intervention of God in history in the last times), which were soon to become unfashionable and be discarded.

**The Gnostic writers.** Hardly had the church thrown off its early Jewish-Christian idiosyncrasies when it found itself confronted by the amorphous but pervasive philosophical-religious movement known as Gnosticism. This movement made a strong bid to absorb Christianity in the 2nd century, and a number of Christian Gnostic sects flourished and contributed richly to Christian literature. Although the church eventually maintained its identity intact, the confrontation forced it to clarify its ideas on vital issues on which it differed sharply from the Gnostics.

Gnostic emphasis

Chief among these were the Gnostics' distinction between the unknown supreme God and the Demiurge (identified with the God of the Old Testament) who created this world; their dualist disparagement of the material order and insistence that the Redeemer became incarnate in appearance only; their belief in salvation by esoteric knowledge; and their division of humanity into a spiritual elite able to achieve salvation and, below this elite, "psychics" capable of a modified form of salvation and "material" people cut off from salvation.

Among the leading 2nd-century Christian Gnostics were Saturninus and Basilides, reputedly pupils of Menander, a disciple of Simon Magus (late 1st century), the alleged founder of the movement; they worked at both Antioch and Alexandria. Most famous and influential was the Egyptian Valentinus, who acquired a great reputation at Rome (*c.* 150) and founded an influential school of thought. Basilides and Valentinus are reported to have written extensively, and their systems can be reconstructed from hostile accounts by Irenaeus, Clement of Alexandria, and other orthodox critics. The Gnostics generally seem to have been prolific writers, and as they needed their own distinctive scriptures they soon created a body of apocryphal books patterned on the New Testament. It was a Syrian Gnostic convert, Tatian, who compiled (late 2nd century) the first harmony of the four Gospels (the *Diatessaron*)—a single gospel using the material from the Gospels; and an Italian Gnostic, Heracleon (2nd century), who prepared the earliest commentary on the Gospel According to John (extracts from it were preserved by Origen). Epiphanius (*c.* 315–403) preserved a *Letter to Flora,* by the Valentinian Gnostic Ptolemaeus (late 2nd century), supplying rules for interpreting the Mosaic Law (the Torah) in a Christian sense; and another disciple of Valentinus, Theodotus (2nd century), published an account of his master's system that was excerpted by Clement of Alexandria.

Almost the entire vast literature of Gnosticism has perished, and until recently the only original documents available to scholars (apart from extracts such as those already mentioned, which were preserved by orthodox critics) were a handful of treatises in Coptic contained in three codices (manuscript books) that were discovered in the 18th and late 19th centuries. The most interesting of these are *Pistis Sophia* and the *Apocryphon of John,* the former consisting of conversations of the risen Jesus with his disciples about the fall and redemption of the aeon (emanation from the Godhead) called Pistis Sophia, the latter of revelations made by Jesus to St. John explaining the presence of evil in the cosmos and showing how mankind can be rescued from it.

The Naj' Ḥammādī codices

Since 1945, however, this meagre store has been richly supplemented by the discovery near Naj' Ḥammādī, in Egypt on the Nile about 78 miles northwest of Luxor, of 13 codices containing Christian Gnostic treatises in Coptic translations. Among these, the Jung Codex (named in honour of the psychoanalyst Carl Jung by those who purchased it for his library) includes five important items: a *Prayer of the Apostle Paul;* an *Apocryphon of James,* recording revelations imparted by the risen Christ to the Apostles; the *Gospel of Truth,* perhaps to be identified with the work of this name attributed by Irenaeus to Valentinus; the *Epistle to Rheginos,* a Valentinian work, possibly by Valentinus himself, on the Resurrection; and a *Tripartite Treatise,* probably written by Heracleon, of the school of Valentinianism. The other documents from the Naj' Ḥammādī library include the *Gospel of Thomas,* a collection of sayings and parables that are ascribed to Jesus; the *Apocryphon of John,* which represents the first chapter of Genesis in mythological terms; and writings ascribed to Philip, Mary Magdalene, Adam, Peter, and Paul.

A figure of immense significance who is often, though perhaps mistakenly, counted among the Gnostics was Marcion, who after breaking with the Roman Church in 144 set up a successful organization of his own. Teaching that there is a radical opposition between the Law and the Gospel, he refused to identify the God of love revealed in the New Testament with the wrathful Creator God of the Old Testament. He set forth these contrasts in his *Antitheses,* and his adoption of a reduced New Testament consisting of the Gospel According to Luke and certain Pauline epistles, all purged of presumed Jewish interpolations, had an important bearing on the church's formation of its own fuller canon.

**The Apologists.** The orthodox literature of the 2nd and early 3rd centuries tends to have a distinctly defensive or polemical colouring. It was the age of Apologists, and these Apologists engaged in battle on two fronts. First, there was the hostility and criticism of pagan society. Because of its very aloofness the church was popularly suspected of sheltering all sorts of immoralities and thus of threatening the established order. At a higher level, Christianity, as it became better known, was being increasingly exposed to intellectual attack. The physician Galen of Pergamum (129–*c.* 199) and the Middle Platonist thinker Celsus, who followed the religiously inclined form of Platonism that flourished from the 3rd century BC to the 3rd century AD (compare his devastating *Alēthēs logos,* or *True Word,* written *c.* 178), were only two among many "cultured despisers." But, second, orthodoxy had to take issue with distorting tendencies within, whether these took the form of Gnosticism or of other heresies, such as the so-called semi-Gnostic Marcion's rejection of the Old Testament revelation or the claim of the ecstatic prophet from Phrygia, Montanus, to be the vehicle of a new outpouring of the Holy Spirit. Christianity had also to define exactly where it stood in relation to Hellenistic culture.

Characteristics of apologetic literature

Strictly speaking, the term Apologists denotes the 2nd-century writers who defended Christianity against external critics, pagan and Jewish. The earliest of this group was Quadratus, who in about 124 addressed an apology for the faith to the emperor Hadrian; apart from a single fragment it is now lost. Other early Apologists who are mere names known to scholars are Aristo of Pella, the first to prepare an apology to counter Jewish objections, and Apollinaris, bishop of Hierapolis, said to be the author of numerous apologetic works and also of a critique of Montanism. An early apology that has survived intact is that of Aristides, addressed about 140 to the emperor Antoninus Pius; after being completely lost, the text was rediscovered in the 19th century. The most famous Apologist, however, was Justin, who was converted to Christianity after trying various philosophical schools, paid lengthy visits to Rome, and was martyred there (*c.* 165). Justin's two *Apologies* are skillful presentations of the Christian case to the pagan critics; and his *Dialogue with Trypho* is an elaborate defense of Christianity against Judaism.

Relationship to philosophy

Justin's attitude to pagan philosophy was positive, but his pupil Tatian could see nothing but evil in the Greco-Roman civilization. Indeed, Tatian's *Discourse to the Greeks* is less a positive vindication of Christianity than a sharp attack on paganism. His contemporary Athenagoras of Athens, author of the apologetic work *Embassy for*

*the Christians* and a treatise *On the Resurrection of the Dead,* is as friendly as Justin to Greek culture and philosophy. Two others who deserve mention are Theophilus of Antioch, a prolific publicist whose only surviving work is *To Autolycus,* prepared for his pagan friend Autolycus; and the anonymous author of the *Letter to Diognetus,* an attractive and persuasive exposition of the Christian way of life that is often included among the Apostolic Fathers.

As stylists the Apologists reach only a passable level; even Athenagoras scarcely achieves the elegance at which he obviously aimed. But they had little difficulty in refuting the spurious charges popularly brought against Christians, including atheism, cannibalism, and promiscuity, or in mounting a counterattack against the debasements of paganism. More positively, they strove to vindicate the Christian understanding of God and specific doctrines such as the divinity of Christ and the resurrection of the body. In so doing, most of them exploited current philosophical conceptions, in particular that of the Logos (Word), or rational principle underlying and permeating reality, which they regarded as the divine reason, become incarnate in Jesus. They have been accused of Hellenizing Christianity (making it Greek in form and method), but they were in fact attempting to formulate it in intellectual categories congenial to their age. In a real sense they were the first Christian theologians. But the same tension between the Gospel and philosophy was to persist throughout the patristic period, with results that were sometimes positive, as in Augustine and Gregory of Nyssa, and sometimes negative, as in the radical Arians Aëtius and Eunomius.

As the 2nd century advanced, a more confident, aggressive spirit came over Christian Apologists, and their intellectual and literary stature increased greatly. Clement of Alexandria, for example, while insisting on the supremacy of faith, freely drew on Platonism and Stoicism to clarify Christian teaching. In his *Protreptikos* ("Exhortation") and *Paidagōgos* ("Instructor") he urged pagans to abandon their futile beliefs, accept the Logos as guide, and allow their souls to be trained by him. In interpreting scripture he used an allegorizing method derived from the Jewish philosopher Philo, and against Gnosticism he argued that the baptized believer who studies the Scriptures is the true Gnostic, faith being at once superior to knowledge and the beginning of knowledge.

The critique of Gnosticism was much more systematically developed by Clement's older contemporary, Irenaeus of Lyon, in his voluminous *Against Heresies.* While countering the Valentinian dualism that asserted that spirit was good and matter evil, this treatise makes clear the church's growing reliance on its creed or "rule of faith," on the New Testament canon, and on the succession of bishops as guarantors of the true apostolic tradition. Irenaeus was also a constructive theologian, expounding ideas about God as Creator, about the Son and the Spirit as his "two hands," about Christ as the New Adam who reconciles fallen humanity with God, and about the worldwide church with its apostolic faith and ministry, a concept that theology was later to take up eagerly.

More brilliant as a stylist and controversialist, the North African lawyer Tertullian was also the first Latin theologian of considerable importance. Unlike Clement, he reacted with hostility to pagan culture, scornfully asking, "What has Athens to do with Jerusalem?" His *Apology* remains a classic of ancient Christian literature, and his numerous moral and practical works reveal an uncompromisingly rigid moral view. Although later becoming a Montanist himself (a follower of the morally rigorous and prophetic sect founded by Montanus), he wrote several antiheretical tracts, full of abuse and biting sarcasm. Yet, in castigating heresy he was able to formulate the terminology, and to some extent the theory, of later Trinitarian and Christological orthodoxy; his teaching on the Fall of Man, aimed against Gnostic dualism, in part anticipates Augustine.

Roughly contemporary with Tertullian, and like him an intellectual and a rigorist, was Hippolytus, a Greek-speaking Roman theologian and antipope. He, too, had a vast literary output, and although some of the surviving works attributed to him are disputed, it is probable that he wrote the comprehensive *Refutation of All Heresies,* attacking Gnosticism, as well as treatises denouncing specifically Christian heresies. He was also the author both of numerous commentaries on scripture and (probably) of the *Apostolic Tradition,* an invaluable source of knowledge about the primitive Roman liturgy. His *Commentary on Daniel* (*c.* 204) is the oldest Christian biblical commentary to survive in its entirety. His exegesis (interpretive method) is primarily typological—*i.e.,* treating the Old Testament figures, events, and other aspects as "types" of the new order that was inaugurated by Christ.

**Late 2nd to early 4th century.** Meanwhile, a brilliant and distinctive phase of Christian literature was opening at Alexandria, the chief cultural centre of the empire and the meeting ground of the best in Hellenistic Judaism, Gnosticism, and Neoplatonism. Marked by the desire to present Christianity in intellectually satisfying terms, this literature has usually been connected with the catechetical school, which, according to tradition, flourished at Alexandria from the end of the 2nd through the 4th century. Except for the brief period, however, when Origen was in charge of it, it may be doubted whether the school was ever itself a focus of higher Christian studies. When speaking of the school of Alexandria, some scholars claim that it is better to think of a distinguished succession of like-minded thinkers and teachers who worked there and whose highly sophisticated interpretation of Christianity exercised for generations a formative impact on large sectors of eastern Christendom.

The catechetical school of Alexandria

The real founder of this theology, with its Platonist leaning, its readiness to exploit the metaphysical implications of revelation, and its allegorical understanding of scripture, was Clement (*c.* 150–*c.* 215), the Christian humanist whose welcoming attitude to Hellenism and critique of Gnosticism were noted above. His major work, the *Strōmateis* ("Miscellanies"), untidy and deliberately unsystematic, brings together the inheritance of Jewish Christianity and Middle Platonism in what aspires to be a summary of Christian gnosis (knowledge). All his reasoning is dominated by the idea of the Logos who created the universe and who manifests the ineffable Father alike in the Old Testament Law, the philosophy of the Greeks, and finally the incarnation of Christ. Clement was also a mystic for whom the higher life of the soul is a continuous moral and spiritual ascent.

Significance and influence of Origen

But it is Origen (*c.* 185–*c.* 254) whose achievement stamps the Alexandrian school. First and foremost, he was an exegete (critical interpreter), as determined to establish the text of scripture scientifically (compare his *Hexapla*) as to wrest its spiritual import from it. In homilies, scholia (annotated works), and continuous commentaries he covered the whole Bible, deploying a subtle, strongly allegorical exegesis designed to bring out several levels of significance. As an apologist, in his *Contra Celsum,* he refuted the pagan philosopher Celsus' damaging onslaught on Christianity. In all his writings, but especially his *On First Principles,* Origen shows himself to be one of the most original and profound of speculative theologians. Neoplatonist in background, his system embraces both the notion of the preexistence of souls, with their fall and final restoration, and a deeply subordinationist doctrine of the Trinity—*i.e.,* one in which the Son is subordinate to the Father. For his spiritual teaching, with its emphasis on the battle against sin, on freedom from passions, and on the soul's mystical marriage with the Logos, his *Commentary on Canticles* provides an attractive introduction.

Origen's influence on Christian doctrine and spirituality was to be immense and many-sided; the orthodox Fathers and the leading heretics of the 4th century alike reflect it. Meanwhile, the Alexandrian tradition was maintained by several remarkable disciples. Two of these whose works have been entirely lost but who are reported to have been polished writers were Theognostus (fl. 250–280) and Pierius (fl. 280–300), both heads of the catechetical school and apparently propagators of Origen's ideas. But there are two others of note, Dionysius of Alexandria (*c.* 200–*c.* 265) and Gregory Thaumaturgus (*c.* 213–*c.* 270), of whose works some fragments have survived. Dionysius of Alexandria wrote on natural philosophy and the Christian doctrine of creation but is chiefly remembered for his dis-

pute with Pope Dionysius (reigned 259–268) of Rome on the correct understanding of the Trinity. In this Dionysius of Alexandria is revealed as a faithful exponent of Origen's pluralism and subordinationism. Gregory Thaumaturgus left a fascinating *Panegyric to Origen,* giving a graphic description of Origen's method of instruction, as well as a dogmatically important *Symbol* and a *Canonical Epistle* that is in effect one of the most ancient treatises of casuistry (*i.e.,* the application of moral principles to practical questions).

If Origen inspired admiration, his daring speculations also provoked criticism. At Alexandria itself, Peter, who became bishop in about 300 and composed theological essays of which only fragments remain, attacked Origen's doctrines of the preexistence of souls and their return into the condition of pure spirits. But the acutest of his critics was Methodius of Olympus (d. 311), of whose treatises *The Banquet,* exalting virginity, survives in Greek and others mainly in Slavonic translations. Although indebted to Alexandrian allegorism, Methodius remained faithful to the Asiatic tradition (literal and historical) of Irenaeus—who had come to France from Asia Minor—and his realism and castigated Origen's ideas on the preexistence of souls, the flesh as the spirit's prison, and the spiritual nature of the resurrected body. As a writer he strove after literary effect, and Jerome, writing a century later, praised the excellence of his style.

Origin and gradual emergence of Latin patristic literature

Latin Christian literature was slow in getting started, and North Africa has often been claimed as its birthplace. Tertullian, admittedly, was the first Christian Latinist of genius, but he evidently had humbler predecessors. Latin versions of the Bible, recoverable in part from manuscripts, were appearing in Africa, Gaul, and Italy during the 2nd century. In that century, too, admired works such as *I Clement, Barnabas,* and the *Shepherd of Hermas* were translated into Latin. The oldest original Latin texts are probably the Muratorian Canon, a late 2nd-century Roman canon, or list of works accepted as scripture, and the *Acts of the Scillitan Martyrs* (180) of Africa.

The first noteworthy Roman Christian to use Latin was Novatian, the leader of a rigorist schismatic group. His surviving works reveal him as an elegant stylist, trained in rhetoric and philosophy, and a competent theologian. His doctrinally influential *De trinitate* ("Concerning the Trinity") is basically apologetic: against Gnostics it defends the oneness and creative role of Almighty God, against Marcion it argues that Christ is the Son of God the Creator, against Docetism (the heresy claiming that Jesus only seemed the Christ) that Christ is truly man, and against Sabellianism (the denial of real distinctions in the Godhead, viewing the Father, Son, and Holy Spirit as three successive modes of revelation) that in spite of Christ's being fully divine there is but one God. His rigorous moralism comes out in his *On Public Shows* and *On the Excellence of Chastity* (both once attributed to Cyprian); in *On Jewish Foods* he maintains that the Old Testament food laws no longer apply to Christians, the animals that were classified as unclean having been intended to symbolize vices.

Works of Cyprian

A much greater writer than Novatian was his contemporary and correspondent, Cyprian, the statesmanlike bishop of Carthage. A highly educated convert to Christianity, Cyprian left a large corpus of writings, including 65 letters and a number of moral, practical, and theological treatises. As an admirer of Tertullian, he continued some of his fellow North African's tendencies, but his style is more classical, though much less brilliant and individual. Cyprian's letters are a mine of information about a fascinating juncture in church history. His collections of *Three Books of Testimonies to Quirinus,* or authoritative scripture texts, illustrate the church's reliance on these in defending its theological and ethical positions. A work that has been of exceptional importance historically is *On the Unity of the Catholic Church,* in which Cyprian contends that there is no salvation outside the church and defines the role of the Roman see. His *To Demetrianus* is an original, powerful essay refuting the allegation of pagans that Christianity was responsible for the calamities afflicting society.

Three writers from the later portion of this period deserve mention. Victorinus of Pettau was the first known Latin biblical exegete; of his numerous commentaries the only one that remains is the commentary on Revelation, which maintained a millenarian outlook—predicting the 1,000-year reign of Christ at the end of history—and was clumsy in style. Arnobius the Elder (converted by 300) sought in his *Adversus nationes* ("Against the Pagans"), like Tertullian and Cyprian before him, to free Christianity from the charge of having caused all the evils plaguing the empire, but ended up by launching a violent attack on the contemporary pagan cults. A surprising feature of this ill-constructed, verbose apology is Arnobius' apparent ignorance concerning several cardinal points of Christian doctrine, combined with his great enthusiasm for his newfound faith.

By contrast, his much abler pupil Lactantius (*c.* AD 240–*c.* 320), like him a native of North Africa, was a polished writer and the leading Latin rhetorician of the day. His most ambitious work, the *Divine Institutes,* attempted, against increasingly formidable pagan attacks, to portray Christianity as the true form of religion and life and is in effect the first systematic presentation of Christian teaching in Latin. The later *On the Death of Persecutors,* now generally recognized as his, describes the grim fates of persecuting emperors; it is a primary source for the history of the early 4th century and also represents a crude attempt at a Christian philosophy of history.

### THE POST-NICENE PERIOD

The 4th and early 5th centuries witnessed an extraordinary flowering of Christian literature, the result partly of the freedom and privileged status now enjoyed by the church, partly of the diversification of its own inner life (compare the rise of monasticism), but chiefly of the controversies in which it hammered out its fundamental doctrines.

Arianism, which denied Christ's essential divinity, aroused an all-pervasive reaction in the 4th century; the task of the first two ecumenical councils, at Nicaea (325) and Constantinople (381), was to affirm the orthodox doctrine of the Trinity. In the 5th century the Christological question moved to the fore, and the Council of Chalcedon (451), completing that of Ephesus (431), defined Christ as one person in two natures. The Christological controversies of the 5th century were extremely complex, involving not only theological issues but also issues of national concerns—especially in the Syrian-influenced East, where the national churches were called non-Chalcedonian because they rejected the doctrinal formulas of the Council of Chalcedon.

Involved in the 5th-century Christological controversy were many persons and movements: Nestorius, consecrated patriarch of Constantinople in 428, and his followers, the Nestorians, who were concerned with preserving the humanity of Christ as well as his divinity; Cyril, patriarch of Alexandria, and his followers, who were devoted to maintaining a balanced emphasis on both of the natures of Christ, divine and human; Eutyches (*c.* 378–after 451), a muddleheaded archimandrite (head of a monastery) who affirmed two natures before and one nature after the incarnation; the Monophysites, who (following Eutyches) stressed the one unified nature of Christ; and the moderates and those who sought theological, ecclesiastical, and even political solutions to this highly complex doctrinal dispute, such as Pope Leo I. It was a time when the Alexandrian and Antiochene theological schools vied with each other for the control of the theology of the church. In the Syrian East the Antiochene tradition continued in the schools of Edessa and Nisibis, which became centres of a non-Greek national renaissance. The issues of grace, free will, and the Fall of Man concerned the West mainly. Meanwhile, old literary forms were developing along more mature lines, and new ones were emerging, including historiography, lives of saints, set piece (fixed-form) oratory, mystical writings, and hymnody.

Literature of the Arian controversy

**The Nicene Fathers.** A seesaw struggle between Arians and orthodox Christians dominated the immediate post-Nicene period. Arius himself, Eusebius of Nicomedia, and other radicals occupied the extreme left wing, carrying Origen's views on the subordination of the Son to what

became dangerous lengths. Apart from a few precious letters and fragments, their writings have perished. On the extreme right Athanasius, Eustathius of Antioch, and Marcellus of Ancyra (strongly anti-Origenist) tenaciously upheld the Nicene decision that the Son was of the same substance with the Father. Again, the writings of the two latter figures, except for scattered but illuminating fragments, have disappeared. Most churchmen preferred the middle ground; loyal to the Origenist tradition, they suspected the Nicene Creed of opening the door to Sabellianism but were equally shocked by Arianism in its more uncompromising forms. Eusebius of Caesarea (*c.* 260–*c.* 340) was their spokesman, and for decades the eastern emperors supported his mediating line.

Eusebius is chiefly known as a historian; his *Ecclesiastical History,* with its scholarly use of documents and guiding idea that the victory of Christianity is the proof of its divine origin, introduced something novel and epoch-making. But he also wrote voluminous apologetic treatises, biblical and exegetical works, and polemical tracts against Marcellus of Ancyra. From these can be gathered his theology of the Word, which was Origenist in inspiration and profoundly subordinationist and which made the strict Nicenes suspect him as an ally of Arius. Such suspicions were unjust, for he upheld Origen's doctrine of eternal generation (*i.e.,* that the Word is generated outside the category of time) and rejected the extreme Arian theses. His influence can be studied in the works of Cyril of Jerusalem (*c.* 315–386?), whose *Catecheses,* or introductory lectures on Christian doctrine for candidates for baptism, exemplify a pastoral type of Christian literature. Though critical of the Arian positions, Cyril remained reserved in his attitude toward the Nicene theology and at several other points showed affinities with Eusebius.

Significance of Athanasius

Athanasius (*c.* 293–373) bestrides the 4th century as the inflexible champion of the Nicene dogma. He had been present at the council, defending Alexander, the theologian-bishop of Alexandria from 313 to 328, who had exposed Arius; and after succeeding Alexander in 328 he spent the rest of his stormy life defending, expounding, and drawing out the implications of the Nicene theology. His most thorough and effective exposition of the Son's eternal origin in the Father and essential unity with him is contained in his *Four Orations Against the Arians;* but in addition he produced a whole series of treatises, historical or dogmatic or both, as well as letters, covering different aspects of the controversy.

It would be misleading, however, to delineate Athanasius exclusively as a polemicist. First, even in his polemical writings he was working out a positive doctrine of the triune God that anticipated later formal definitions. His *Letters to Bishop Serapion,* with their persuasive presentation of the Holy Spirit as a consubstantial (of the same substance) person in the Godhead, are an admirable illustration. Also his noncontroversial works, such as the relatively early but brilliant apologies *Discourse Against the Pagans* and *The Incarnation of the Word of God;* the attractive and influential *Life of St. Antony,* which was to give a powerful impulse to monasticism (especially in the West); and his numerous exegetical and ascetic essays, which survive largely in fragments, sometimes in Coptic or Syriac translations, should not be overlooked.

**The Cappadocian Fathers.** Although Athanasius prepared the ground, constructive agreement on the central doctrine of the Trinity was not reached in his lifetime, either between the divided parties in the East or between East and West with their divergent traditions. The decisive contribution to the Trinitarian argument was made by a remarkable group of philosophically minded theologians from Cappadocia—Basil of Caesarea, his younger brother Gregory of Nyssa, and his lifelong friend Gregory of Nazianzus. Of aristocratic birth and consummate culture, all three were drawn to the monastic ideal, and Basil and Gregory of Nazianzus achieved literary distinction of the highest order. While their joint accomplishments in doctrinal definition were indeed outstanding, each made a noteworthy mark in other fields as well.

Contributions to theology and monasticism

So far as Trinitarian dogma is concerned, the Cappadocians succeeded, negatively, in overthrowing Arianism in the radical form in which two acute thinkers, Aëtius (d. *c.* 366) and Eunomius (d. *c.* 394), had revived it in their day, and, positively, in formulating a conception of God as three Persons in one essence that eventually proved generally acceptable. The oldest of Basil's dogmatic writings is his only partially successful *Against Eunomius,* the most mature his essay *On the Holy Spirit.* Gregory of Nyssa continued the attack on Eunomius in four massive treatises and published several more positive dogmatic essays, the most successful of which is the *Great Catechetical Oration,* a systematic theology in miniature. The output of Gregory of Nazianzus was much smaller, but his 45 *Orations,* as well as being masterpieces of eloquence, contain his classic statement of Trinitarian orthodoxy. Basil's vast correspondence testifies to his practical efforts to reconcile divergent movements in Trinitarian thinking.

Basil is famous as a letter writer and preacher and for his views on the appropriate attitude of Christians toward Hellenistic culture; but his achievement was not less significant as a monastic legislator. His two monastic rules, used by St. Benedict and still authoritative in the Greek Orthodox Church, are tokens of this. Gregory of Nazianzus, too, was an accomplished letter writer, but his numerous, often lengthy poems have a special interest. Dogmatic, historical, and autobiographical, they are often intensely personal and lay bare his sensitive soul. On the other hand, Gregory of Nyssa, much the most speculative of the three, was an Origenist both in his allegorical interpretation of scripture and his eschatology. But he is chiefly remarkable as a pioneer of Christian mysticism, and in his *Life of Moses, Homilies on Canticles,* and other books he describes how the soul, in virtue of having been created in the divine image, is able to ascend, by successive stages of purification, to a vision of God.

Epiphanius of Salamis

A figure who stood in sharp contrast, intellectually and in temperament, to the Cappadocians was their contemporary, Epiphanius of Salamis, in Cyprus. A fanatical defender of the Nicene solution, he was in no sense a constructive theologian like them, but an uncritical traditionalist who rejected every kind of speculation. He was an indefatigable hammer against heretics, and his principal work, the *Panarion* ("Medicine Chest"), is a detailed examination of 80 heresies (20 of them pre-Christian); it is invaluable for the mass of otherwise unobtainable documents it excerpts. Conformably with Epiphanius' contempt for classical learning, the work is written in Greek without any pretension to elegance. His particular bête noire was Origen, to whose speculations and allegorism he traced virtually all heresies.

**Monastic literature.** From the end of the 3rd century onward, monasticism was one of the most significant manifestations of the Christian spirit. Originating in Egypt and spreading thence to Palestine, Syria, and the whole Mediterranean world, it fostered a literature that illuminates the life of the ancient church.

Both Anthony (*c.* 250–355), the founder of eremitical, or solitary, monasticism in the Egyptian desert, and Ammonas (fl. *c.* 350), his successor as leader of his colony of anchorites (hermits), wrote numerous letters; a handful from the pen of each is extant, almost entirely in Greek or Latin translation of the Coptic originals. Those of Ammonas are particularly valuable for the history of the movement and as reflecting the uncomplicated mysticism that inspired it. The founder of monastic community life, also in Egypt, was Pachomius (*c.* 290–346), and the extremely influential rule that he drew up has been preserved, mainly in a Latin translation made by Jerome.

Influence of monastic literature on ethics and mysticism

Though these and other early pioneers were simple, practical men, monasticism received a highly cultivated convert in 382 in Evagrius Ponticus. He was the first monk to write extensively and was in the habit of arranging his material in groups of a hundred aphorisms, or "centuries," a literary form that he invented and that was to have a great vogue in Byzantine times. A master of the spiritual life, he classified the eight sins that undermine the monk's resolution and also the ascending levels by which the soul rises to wordless contemplation. Later condemned as an Origenist, he was deeply influential in the East, and, through John Cassian, in the West as well.

Side by side with works composed by monks there sprang up a literature concerned with them and the monastic movement. Much of it was biographical, the classic example being Athanasius' *Life of St. Antony.* Sulpicius Severus (*c.* 363–*c.* 420) took this work as his model when early in the 5th century he wrote his *Life of St. Martin of Tours,* the first Western biography of a monastic hero and the pattern of a long line of medieval lives of saints. But it was Palladius (*c.* 363–before 431), a pupil of Evagrius Ponticus, who proved to be the principal historian of primitive monasticism. His *Lausiac History* (so called after Lausus, the court chamberlain to whom he dedicated it), composed about 419/420, describes the movement in Egypt, Palestine, Syria, and Asia Minor. Since much of the work is based on personal reminiscences or information received from observers, it is, despite the legendary character of many of its narratives, an invaluable source book.

Finally, no work so authentically conveys the spirit of Egyptian monasticism as the *Apophthegmata Patrum* ("Sayings of the Fathers"). Compiled toward the end of the 5th century, but using much older material, it is a collection of pronouncements of the famous desert personalities and anecdotes about them. The existing text is in Greek, but it probably derives from an oral tradition in Coptic.

**The school of Antioch.** Antioch, like Alexandria, was a renowned intellectual centre, and a distinctive school of Christian theology flourished there and in the surrounding region throughout the 4th and the first half of the 5th century. In contrast to the Alexandrian school, it was characterized by a literalist exegesis and a concern for the completeness of Christ's manhood. Little is known of its traditional founder, the martyr-priest Lucian (d. 312), except that he was a learned biblical scholar who revised the texts of the Septuagint and the New Testament. His strictly theological views, though a mystery, must have been heterodox, for Arius, Eusebius of Nicomedia, and other Arians claimed to be his disciples ("fellow Lucianists"), and Bishop Alexander of Alexandria, who denounced them, lists Lucian among those who influenced them. But Eustathius of Antioch, the champion of Nicene orthodoxy, is probably more representative of the school, with his antipathy to what he regarded as Origen's excessive allegorism and his recognition, as against the Arians, of the presence of a human soul in the incarnate Christ.

Literalistic views of the School of Antioch

It was, however, much later in the 4th century, in the person of Diodore of Tarsus (*c.* 330–*c.* 390), that the School of Antioch began to reach the height of its fame. Diodore courageously defended Christ's divinity against Julian the Apostate, the Roman emperor who attempted to revive paganism, and in his lifetime was regarded as a pillar of orthodoxy. Later critics detected anticipations of Nestorianism (the heresy upholding the division of Christ's Person) in his teaching, and as a result his works, apart from some meagre fragments, have perished. They were evidently voluminous and wide-ranging, covering exegesis, apologetics, polemics, and even astronomy; and he not only strenuously opposed Alexandrian allegorism but also expounded the Antiochene *theoria,* or principle for discovering the deeper intention of scripture and at the same time remaining loyal to its literal sense.

Significant Antiochene theologians

In stature and intellectual power Diodore was overshadowed by his two brilliant pupils, Theodore of Mopsuestia (*c.* 350–428/429) and John Chrysostom (*c.* 347–407). Both had also studied under the famous pagan Sophist rhetorician Libanius (314–393), thereby illustrating the cross-fertilization of pagan and Christian cultures at this period. Like Diodore, Theodore later fell under the imputation of Nestorianism, and the bulk of his enormous literary output—comprising dogmatic as well as exegetical works—was lost. Fortunately, the 20th century has seen the recovery of a few important texts in Syriac translations (notably his *Commentary on St. John* and his *Catechetical Homilies*), as well as the reconstruction of the greater part of his *Commentary on the Psalms.* This fresh evidence confirms that Theodore was not only the most acute of the Antiochene exegetes, deploying the hermeneutics (critical interpretive principles) of his school in a thoroughly scientific manner, but also an original theologian who, despite dangerous tendencies, made a unique contribution to the advancement of Christology. His *Catechetical Homilies* are immensely valuable both for understanding his ideas and for the light they throw on sacramental doctrine and liturgical practice.

In contrast to Theodore, John was primarily a preacher; indeed he was one of the most accomplished of Christian orators and amply merited his title "Golden-Mouthed" (*Chrysostomos*). With the exception of a few practical treatises and a large dossier of letters, his writings consist entirely of addresses, the majority being expository of the Bible. There he shows himself a strict exponent of Antiochene literalism, reserved in exploiting even the traditional typology (*i.e.,* treatment of Old Testament events and so forth as prefigurative of the new Christian order) but alert to the moral and pastoral lessons of his texts. This interest, combined with his graphic descriptive powers, makes his sermons a mirror of the social, cultural, and ecclesiastical conditions in contemporary Antioch and Constantinople, as well as of his own compassionate concern as a pastor. Indefatigable in denouncing heresy, he was not an original thinker; on the other hand, he was outstanding as a writer, and connoisseurs of rhetoric have always admired the grace and simplicity of his style in some moods, its splendour and pathos in others.

The last noteworthy Antiochene, Theodoret of Cyrrhus (*c.* 393–*c.* 458), in Syria, was also an elegant stylist. His writings were encyclopaedic in range, but the most memorable perhaps are his *Remedy for Greek Maladies,* the last of ancient apologies against paganism; and his *Ecclesiastical History,* continuing Eusebius' work down to 428. His controversial treatises are also important, for he skillfully defended the Antiochene Christology against the orthodox Bishop Cyril of Alexandria and was instrumental in getting its more valuable features recognized at the Council of Chalcedon. He was a scholar with a comprehensive and eclectic mind, and his large correspondence testifies to his learning and mastery of Greek prose as well as illustrating the history and intellectual life of the age.

**The schools of Edessa and Nisibis.** Parallel with its richer and better-known Greek and Latin counterparts, an independent Syriac Christian literature flourished inside, and later outside (in Persia), the frontiers of the Roman Empire from the early 4th century onward. Aphraates, an ascetic cleric under whose name 23 treatises written between 336 and 345 have survived, is considered the first Syriac Father. Deeply Christian in tone, these tracts present a primitive theology, with no trace of Hellenistic influence but a firm grasp and skillful use of scripture. Edessa and Nisibis (now Urfa and Nusaybin in southeast Turkey) were the creative centres of this literature. Edessa had been a focus of Christian culture well before 200; the old Syriac version of the New Testament and Tatian's *Diatessaron,* as well as a mass of Syriac apocryphal writings, probably originated there.

Syriac Christian literature

The chief glory of Edessene Christianity was Ephraem Syrus (*c.* 306–373), the classic writer of the Syrian Church who established his school of theology there when Nisibis, its original home and his own birthplace, was ceded to Persia under the peace treaty of 363, after the death of Julian the Apostate. In his lifetime Ephraem had a reputation as a brilliant preacher, commentator, controversialist, and above all, sacred poet. His exegesis shows Antiochene tendencies, but as a theologian he championed Nicene orthodoxy and attacked Arianism. His hymns, many in his favourite seven-syllable metre, deal with such themes as the Nativity, the Epiphany, and the Crucifixion or else are directed against skeptics and heretics. His *Carmina Nisibena* ("Songs of Nisibis") make a valuable source book for historians, especially for information about the frontier wars.

After Ephraem's death in 373, the school at Edessa developed his lively interest in exegesis and became increasingly identified with the Antiochene line in theology. Among those responsible for this was one of its leading instructors, Ibas (d. 457), who worked energetically translating Theodore of Mopsuestia's commentaries and disseminating his Christological views. His own stance on the now urgent Christological issue was akin to that of Theodoret of Cyrrhus—roughly midway between Nestorius' dualism

Influence of the Antiochene school of thought in the Syrian East

and the Alexandrian doctrine of one nature—and he bluntly criticized Cyril's position in his famous letter to Maris (433), the sole survivor (in a Greek translation) of his abundant works; it was one of the Three Chapters anathematized by the second Council of Constantinople (553).

The frankly Antiochene posture typified by Ibas brought the school into collision with Rabbula, bishop of Edessa from 412 to 435, an uncompromising supporter of Cyril and the Alexandrian Christology. As well as writing numerous letters, hymns, and a sermon against Nestorius, Rabbula translated Cyril's *De recta fide* (*Concerning the Correct Faith*) into Syriac and also probably compiled the revised Syriac version of the four Gospels (contained in the Peshitta) in order to oust Tatian's *Diatessaron.* On his death he was succeeded by Ibas, who predictably exerted his influence in an Antiochene direction.

Another eminent Edessene writer was Narses (d. *c.* 503), who became one of the formative theologians of the Nestorian Church. He was the author of extensive commentaries, now lost, and of metrical homilies, dialogue songs, and liturgical hymns. In 447, when a Monophysite reaction set in, he was expelled from Edessa along with Barsumas, the head of the school, but they promptly set up a new school at Nisibis on Persian territory. The school at Edessa was finally closed, because of its Nestorian leanings, by the emperor Zeno in 489, but its offshoot at Nisibis flourished for more than 200 years and became the principal seat of Nestorian culture. At one time it had as many as 800 students and was able to ensure that the then prosperous church in Persia was Nestorian. On the other hand, Philoxenus of Mabbug, who had studied at Edessa in the second half of the 5th century and was one of the most learned of Syrian theologians, was a vehement advocate of Monophysitism. His 13 homilies on the Christian life and his letters reveal him as a fine prose writer; but he is chiefly remembered for the revision of the Syriac translation of the Bible (the so-called Philoxenian version) for which he was responsible and which was used by Syrian Monophysites in the 6th century.

**The Chalcedonian Fathers.** From about 428 onward Christology became an increasingly urgent subject of debate in the East and excited interest in the West as well. Two broad positions had defined themselves in the 4th century. Among Alexandrian theologians the "Word-flesh" approach was preferred, according to which the Word had assumed human flesh at the Incarnation; Christ's possession of a human soul or mind was either denied or ignored. Antiochene theologians, on the other hand, consistently upheld the "Word-man" approach, according to which the Word had united himself to a complete man; this position ran the risk, unless carefully handled, of so separating the divinity and the humanity as to imperil Christ's personal unity.

Apollinarius the Younger (*c.* 310–*c.* 390) had brilliantly exposed the logical implications of the Alexandrian view; although condemned as a heretic, he had forced churchmen of all schools to recognize, though with varying degrees of practical realism, a human mind in the Redeemer. His writings were systematically destroyed, but the remaining fragments confirm his intellectual acuteness as well as his literary skill. The crisis of the 5th century was precipitated by the proclamation by Nestorius, patriarch of Constantinople—pushing Antiochene tendencies to extremes—of a Christology that seemed to many to imply two Sons. Nestorius held that Mary was not only *Theotokos* ("God-bearing") but also *anthropotokos* ("man-bearing"), though he preferred the term *Christotokos* ("Christ-bearing"). In essence, he was attempting to protect the concept of the humanity of Christ. The controversy raged with extraordinary violence from 428 to 451, when the Council of Chalcedon hammered out a formula that at the time seemed acceptable to most and that attempted to do justice to the valuable insights of both traditions.

The rise of the Christological crisis

A number of theologians and ecclesiastics either prepared the way for or contributed to the Chalcedonian solution. Three who deserve mention are Theodoret of Cyrrhus, Proclus of Constantinople, and John Cassian. The first was probably responsible for drafting the Formula of Union (433) that became the basis of the Chalcedonian Definition. Proclus was an outstanding pulpit orator, and several of his sermons as well as seven letters concerned with the controversy have been preserved; he worked indefatigably to reconcile the warring factions. Cassian prepared the West for the controversy by producing in 430, at the request of the deacon (later pope) Leo, a weighty treatise against Nestorius.

But much the most important, not least because they approached the debate from different standpoints, were Cyril of Alexandria and Pope Leo the Great. Cyril had been the first to denounce Nestorius, and in a whole series of letters and dogmatic treatises he drove home his critique and expounded his own positive theory of hypostatic (substantive, or essential) union. He secured the condemnation of Nestorius at the Council of Ephesus (431), and his own letters were canonically approved at Chalcedon. A convinced adherent of the Alexandrian Word-flesh Christology, he deepened his understanding of the problem as the debate progressed; but his preferred expression for the unity of the Redeemer remained "one incarnate nature of the Word," which he mistakenly believed to derive from Athanasius. Leo provided the necessary balance to this with his famous *Dogmatic Letter,* also endorsed at Chalcedon, which affirmed the coexistence of two complete natures, united without confusion, in the one Person of the Incarnate Word, or Christ.

Contributions of Cyril of Alexandria and Leo the Great

In patristic literature, however, the interest of both Cyril and Leo extends far beyond Christology. Cyril published essays on the Trinitarian issue against the Arians and also commentaries on Old and New Testament books. If the former show little originality, his exegesis marked a reaction against the more fanciful Alexandrian allegorism and a concentration on the strictly typological significance of the text. Leo, for his part, was a notable preacher and one of the greatest of popes. His short, pithy sermons, clear and elegant in style, set a fine model for pulpit oratory in the West; and his numerous letters give an impressive picture of his continuous struggle to promote orthodoxy and the interests of the Roman see.

**Non-Chalcedonian Fathers.** The Chalcedonian settlement was not achieved without some of the leading participants in the debate that preceded it being branded as heretics because their positions fell outside the limits accepted as permissible. It also left to subsequent generations a legacy of misunderstanding and division.

5th-century heretical literature

The outstanding personalities in the former category were Nestorius and Eutyches. It was Nestorius whose imprudent brandishing of extremist Antiochene theses—particularly his reluctance to grant the title of *Theotokos* to Mary, mother of Jesus—had touched off the controversy. Only fragments of his works remain, for after his condemnation their destruction was ordered by the Byzantine government, but these have been supplemented by the discovery, in a Syriac translation, of his *Book of Heraclides of Damascus.* Written late in his life, when Monophysitism had become the bogey, this is a prolix apology in which Nestorius pleads that his own beliefs are identical with those of Leo and the new orthodoxy. Eutyches, on the other hand, an over-enthusiastic follower of Cyril, was led by his antipathy to Nestorianism into the opposite error of confusing the natures. He contended that there was only one nature after the union of divinity and humanity in the Incarnate Word, and he was thus the father of Monophysitism in the strict, and not merely verbal, sense.

After the Council of Ephesus in 431 the eastern bishops of Nestorian sympathies gradually formed a separate Nestorian Church on Persian soil, with the see of its patriarch at Ctesiphon on the Tigris. Edessa and then Nisibis were its theological and literary centres. But a much wider body of eastern Christians, particularly from Egypt and Palestine, found the Chalcedonian dogma of "two natures" a betrayal of the truth as stated by their hero Cyril. For the next two centuries the struggle between these Monophysites and strict Chalcedonians to secure the upper hand convulsed the Eastern Church. Among the Monophysites it produced theologians of high calibre and literary distinction, notably the moderate Severus of Anti-

och (*c.* 465–538), who while contending stoutly for "one nature after the union" was equally insistent on the reality of Christ's humanity. His contemporary Julian of Halicarnassus taught the more radical doctrine that, through union with the Word, Christ's body had been incorruptible and immortal from the moment of the Incarnation.

In the 7th century, inspired by the need for unity in the face of successive Persian and Arab attacks, an attempt was made to reconcile the Monophysite dissenters with the orthodox Chalcedonians. The formula, which it was thought might prove acceptable to both, asserted that, though Christ had two natures, he had only one activity—*i.e.,* one divine will. This doctrine, Monothelitism, stimulated an intense theological controversy but was subjected to profound and far-reaching criticism by Maximus the Confessor, who perceived that, if Christians are to find in Christ the model for their freedom and individuality, his human nature must be complete and therefore equipped with a human will. The formula was condemned as heretical at the third Council of Constantinople of 680–681.

**The post-Nicene Latin Fathers.** Latin Christian literature in this period was slower than Greek in getting started, and it always remained sparser. Indeed, the first half of the 4th century produced only Julius Firmicus Maternus, author not only of the most complete treatise on astrology bequeathed by antiquity to the modern world but also of a fierce diatribe against paganism that has the added interest of appealing to the state to employ force to repress it and its immoralities. From Africa, rent asunder by Donatism, the heretical movement that rejected the efficacy of sacraments administered by priests who had denied their faith under persecution, came the measured anti-Donatist polemic of Optatus of Milevis, writing in 366 or 367, whose line of argument anticipates Augustine's later attack against the Donatists.

Much more significant than either, however, was Gaius Marius Victorinus, the brilliant professor whose conversion in 355 caused a sensation at Rome. Obscure but strikingly original in his writings, he was an effective critic of Arianism and sought to present orthodox Trinitarianism in uncompromisingly Neoplatonic terms. His speculations about the inner life of the triune Godhead were to be taken up by Augustine.

Three remarkable figures, all different, dominate the second half of the century. The first, Hilary of Poitiers, was a considerable theologian, next to Augustine the finest produced by the West in the patristic epoch. For years he deployed his exceptional gifts in persuading the anti-Arian groups to abandon their traditional catchwords and rally round the Nicene formula, which they had tended to view with suspicion. Often unfairly described as a popularizer of Eastern ideas, he was an original thinker whose scriptural commentaries and perceptive Trinitarian studies brought fresh insights. The second, Ambrose of Milan, was an outstanding ecclesiastical statesman, equally vigilant for orthodoxy against Arianism as for the rights of the church against the state. Both in his dogmatic treatises and in his largely allegorical, pastorally oriented exegetical works he relied heavily on Greek models. One of the pioneers of Catholic moral theology, he also wrote hymns that are still sung in the liturgy.

Hilary, Ambrose, and Jerome

The third, Jerome, was primarily a biblical scholar. His enormous commentaries are erudite but unequal in quality; the earlier ones were greatly influenced by Origen's allegorism, but the ones written later, when he had turned against Origen, were more literalist and historical in their exegesis. Jerome's crowning gift to the Western Church and Western culture was the Vulgate translation of the Bible. Prompted by Pope Damasus, he thoroughly revised the existing Latin versions of the Gospels; the Old Testament he translated afresh from the Hebrew. His historical and polemical writings (the latter full of sarcasm and invective) are all interesting, and his rich correspondence supremely so. As a stylist he wrote with a verve and brilliance unmatched in Latin patristic literature.

The two foremost Christian Latin poets of ancient times, Prudentius and Paulinus of Nola, also belong to this half-century. Both used the old classical forms with considerable skill, filling them with a fresh Christian spirit. Prudentius' work is both the finer in quality and the more wide-ranging; in his *Psychomachia* ("The Contest of the Soul"), he introduced an allegorical form that made an enormous appeal to the Middle Ages. Paulinus is also interesting for his extensive correspondence, much admired in his own day, which kept him in close touch with many leading Christian contemporaries.

All these figures are overshadowed by the towering genius of Augustine (354–430). The range of his writings was enormous: they comprise profound discussions of Christian doctrine (notably his *De Trinitate,* or *On the Trinity*); sustained and carefully argued polemics against heresies (Manichaeism, a dualistic religion; Donatism; and Pelagianism, a view that emphasized free will); exegesis, homilies, and ordinary sermons; and a vast collection of letters. His two best-known works, the *Confessions* and *The City of God,* broke entirely fresh ground, the one being both an autobiography and an interior colloquy between the soul and God, the other perhaps the most searching study ever made of the theology of history and of the fundamental contrast between Christianity and the world. On almost every issue he handled—the problem of evil, creation, grace and free will, the nature of the church—Augustine opened up lines of thought that are still debated. The prose style he used matched the level of his argument, having a rich texture, subtle assonance, and grave beauty that were new in Latin.

Augustine as the major theologian of the West

In part recovered in recent years, the works of Pelagius (fl. 405–418) show him to have been a writer and thinker of high quality. Early in the 5th century, when the monasteries of southern Gaul became active intellectual centres, Vincent of Lérins and John Cassian published critiques of Augustine's extreme positions on grace and free will, proposing the alternative doctrine called Semi-Pelagianism, which held that humans by their own free will could desire life with God. This in turn was criticized by able writers like Prosper of Aquitaine (*c.* 390–*c.* 463) and the celebrated preacher Caesarius of Arles (470–542) and was condemned at the Council of Orange (529). Cassian, however, a firsthand student of Eastern monasticism, is chiefly important for his studies of the monastic life, based on material collected in the East. The rules he formulated were freely drawn upon a century later by St. Benedict of Nursia, the reformer of Western monasticism, when Benedict composed his famous and immensely influential rule at Monte Cassino.

The 6th century marks the final phase of Latin patristic literature, which includes several notable figures, of whom Boethius (480–524), philosopher and statesman, is the most distinguished. His *Consolation of Philosophy* was widely studied in the Middle Ages, but he also composed technically philosophical works, including translations of, and commentaries on, Aristotle. Beside him should be set his longer-lived contemporary, Cassiodorus (*c.* 490–*c.* 585), who, as well as encouraging the study of Greek and Latin classics and the copying of manuscripts in monasteries, was himself the author of theological, historical, and encyclopaedic treatises. Also notable is Venantius Fortunatus (*c.* 540–*c.* 600), an accomplished poet whose hymns, such as "Vexilla regis" ("The royal banners forward go") and "Pange lingua" ("Sing, my tongue, the glorious battle"), are still sung. Finally, Gregory the Great (*c.* 540–604) was so prolific and successful an author as to earn the title of Fourth Doctor of the Latin Church. Although unoriginal theologically and reflecting the credulity of the age, his works (which include the earliest life of St. Benedict) made an enormous appeal to the medieval mind.

The decline of Latin and Greek patristic literature

**Later Greek Fathers.** The closing phase of patristic literature lasted longer in the Greek East than in the Latin West, where the decline of culture was hastened by barbarian inroads. But even in the East a slackening of effort and originality was becoming perceptible in the latter half of the 5th century. A clear illustration of this is provided by the practice of substituting chain commentaries composed of excerpts from earlier exegetes and anthologies of opinions of respected past theologians for independent exposition and speculation.

Yet the picture was not altogether dim. In the strictly theological field, Leontius of Byzantium (d. *c.* 545) showed

ability and originality in reinterpreting the Chalcedonian Christology along the lines of St. Cyril with the aid of the increasingly favoured Aristotelian philosophy. Two other writers, very different from him and from each other, revived in the late 5th and early 6th centuries the brilliance of past generations. One was the figure who called himself Dionysius the Areopagite (*c.* 500), the unidentified author of theological and mystical treatises that were destined to have an enormous influence. Based on a synthesis of Christian dogma and Neoplatonism, his work exalts the negative theology (God is understood by what he is not) and traces the soul's ascent from a dialectical knowledge of God to mystical union with him. The other is Romanos Melodos (fl. 6th century), greatest hymnist of the Eastern Church, who invented the kontakion, an acrostic verse sermon in many stanzas with a recurring refrain. The sweep, pathos, and grandeur of his compositions give him a high place of honour among religious poets.

With Maximus the Confessor and John of Damascus the end of the patristic epoch is reached. Maximus was a major critic of Monothelitism; he was also a remarkable constructive thinker whose speculative and mystical doctrines were held in unity by his vision of the incarnation as the goal of history. Writing early in the 8th century, John was chiefly influential through his comprehensive presentation of the teaching of the Greek Fathers on the principal Christian doctrines. But in constructing his synthesis he added at many points a finishing touch of his own; his writings in defense of images, prepared to counter the Iconoclasts (those who advocated destruction of religious images, or icons), were original and important; and he was the author of striking poems, some of which found a place in the Greek liturgy.

#### THE CHARACTER OF THE HERITAGE

The vitality of patristic literature

For 400 or 500 years, when secular culture was slowly but steadily in decline, the patristic writers breathed new life into the Greek and Latin languages and created Syriac as a literary medium. Even when the period came to an end, the halt was really only a temporary pause until the impulses behind it could force other outlets. The literature of the later Byzantine Empire looked back to and drew nourishment from the golden centuries of the Fathers, while Latin Christian letters experienced more than one renascence in the Middle Ages.

The range and variety, too, of the literature are impressive. Its overwhelmingly theological concern necessarily imposed understandable but serious limitations, but, when these have been allowed for, the Christian writers must be acknowledged to have been remarkably successful at molding the traditional literary forms to their new purposes and also at improvising fresh ones adapted to their special situations. Aesthetically considered, patristic literature contains much that is mediocre and even shoddy, but also a great deal that by any standards reaches the heights. And it has a unique interest as the creation of an immensely dynamic and far-reachingly important religious movement during the centuries when it could dominate the whole of life and society. (J.N.D.K.)

## Christian philosophy

It has been debated whether there is anything that is properly called Christian philosophy. Christianity is not a system of ideas but a religion, a way of salvation. But as a religion becomes a distinguishable strand of human history, it inevitably absorbs philosophical assumptions from its environment and generates new philosophical constructions and arguments both in the formation of doctrines and in their defense against philosophical objections. These two topics cannot be kept entirely separate, however, for philosophical criticism from both within and without the Christian community has influenced the development of its beliefs.

#### HISTORY OF THE INTERACTIONS OF PHILOSOPHY AND THEOLOGY

As the Christian movement expanded beyond its original Jewish nucleus into the Greco-Roman world, it had to understand, explain, and defend itself in terms that were intelligible in an intellectual milieu largely structured by Greek philosophical thought. By the 2nd century AD several competing streams of Greek and Roman philosophy—Middle Platonism, Neoplatonism, Epicureanism, Stoicism—had to a great extent flowed together into a common worldview that was basically Neoplatonic, though enriched by the ethical outlook of the Stoics. This constituted the broad intellectual background for most educated people throughout the Roman Empire, functioning in a way comparable, for example, to the pervasive contemporary Western secular view of the universe as an autonomous system within which everything can in principle be understood scientifically.

Influence of Neoplatonism

Some of the Neoplatonic themes that provided intellectual material for Christian and non-Christian thinkers alike in the early centuries of the Common Era were a hierarchical conception of the universe, with the spiritual on a higher level than the physical; the eternal reality of such values as goodness, truth, and beauty and of the various universals that give specific form to matter; and the tendency of everything to return to its origin in the divine reality. The early Christian Apologists were at home in this thought-world, and many of them used its ideas and assumptions both in propagating the Gospel and in defending it as a coherent and intellectually tenable system of belief. Their most common attitude was to accept the prevailing Neoplatonic worldview as basically valid and to present Christianity as its fulfillment, correcting and completing rather than replacing it. Philosophy, they thought, was to the Greeks what the Law was to the Jews—a preparation for the Gospel; and several Apologists agreed with the Jewish writer Philo that Greek philosophy must have received much of its wisdom from Moses. Tertullian (*c.* 155/160–after 220) and Tatian (*c.* 120–173), on the other hand, rejected pagan learning and philosophy as inimical to the Gospel; and the question has been intermittently discussed by theologians ever since whether the Gospel completes and fulfills the findings of human reason or whether reason is itself so distorted by sin as to be incapable of leading toward the truth.

Greek philosophy, then, provided the organizing principles by which the central Christian doctrines were formulated. It is possible to distinguish between, on the one hand, first-order religious expressions, directly reflecting primary religious experience, and, on the other, the interpretations of these in philosophically formulated doctrines whose articulation both contributes to and is reciprocally conditioned by a comprehensive belief-system. Thus the primitive Christian confession of faith, "Jesus is Lord," expressed the Disciples' perception of Jesus as the one through whom God was transformingly present to them and to whom their lives were accordingly oriented in complete trust and commitment. The interpretive process whereby the original experience developed a comprehensive doctrinal superstructure began with the application to Jesus of the two distinctively Jewish concepts of the expected messiah and the Son of man who was to come on the last day and also of the son of God metaphor, which was commonly applied in the ancient world to individuals, whether kings or holy men, who were believed to be close to God. It continued on a more philosophical level with the use, in the Gospel According to John, of the idea of the Logos, drawn both from the Hebraic notions of the Wisdom and the Word of God and from the Greek notion of the Logos as the universal principle of rationality and self-expression. As Jesus, son of God, became Christ, God the Son, the second Person of the Trinity, he was identified with the Logos.

Emergence of official doctrine

For several generations there was great variety and experimentation in Christian thinking. But as Christianity was legally recognized under Constantine in 313 and then became the sole official religion of the Roman Empire under Theodosius, its doctrines had to be formalized and agreed throughout the church. This pressure for uniformity provoked intense debates, which lasted for several generations before the great ecumenical councils (principally Nicaea, 325; Constantinople, 381; and Chalcedon, 451) finally established the official versions of the doctrines of Christ and

the Trinity; to differ from these was heresy. The key ideas in terms of which these Christological and Trinitarian debates were conducted and their conclusions formulated were the Greek concepts of ousia (nature or essence) and hypostasis (entity, used as virtually equivalent to *prosōpon,* person). (In Latin these terms became *substantia* and *persona.*) Christ was said to have two natures, one of which was of the same nature (*homoousios*) as the Father, whereas the other was of the same nature as humanity; and the Trinity was said to consist of one ousia in three hypostases. The Platonic origin of this conceptuality is clear in the explanation of the Cappadocian Fathers that the Father, Son, and Holy Spirit share the same divine ousia in the way Peter, James, and John shared the same humanity. (On the doctrines of Incarnation and the Trinity, see above *Christian doctrine.*)

Another prime example of the influence of Neoplatonism on Christian thought occurred in the response of the greatest of the early Christian thinkers, St. Augustine (354–430), to the perennially challenging question of how it is that evil exists in a world created by an all-good and all-powerful God. Augustine's answer (which, as refined by later thinkers, remained the standard Christian answer until modern times) includes both theological aspects (the ideas of the fall of angels and then of humans, of the redemption of some by the cross of Christ, and of the ultimate disposal of souls in eternities of bliss and torment) and philosophical aspects. The basic philosophical theme, drawn directly from Neoplatonism, is one that the American philosopher Arthur Lovejoy, in *The Great Chain of Being* (1936), called the principle of plenitude. This is the idea that the best possible universe does not consist only of the highest kind of creature, the archangels, but contains a maximum richness of variety of modes of being, thus realizing every possible kind of existence from the highest to the lowest. The result is a hierarchy of degrees both of being and of goodness, for the identity of being and goodness was another fundamental idea received by Augustine from Neoplatonism and in particular from Plotinus (205–270). God, as absolute being and goodness, stands at the summit, with the great chain of being descending through the many forms of spiritual, animal, and plant life down to lifeless matter. This conception explains why there are lower forms of existence—dogs, snakes, insects, viruses—as well as higher. Each embodies being and is therefore good on its own level; and together they constitute a universe whose rich variety is beautiful in the sight of God. Evil only comes about when creatures at any level forfeit the distinctive goodness with which the Creator had endowed them. Evil is thus negative or privative, a lack of proper good rather than anything having substance in its own right. This, too, was a theme that had been taken over from Neoplatonism by a number of earlier Christian writers. And if evil is not an entity, or substance, it follows that it was not a part of God's original creation. It consists instead in the going wrong of something that is in itself good, though (because made out of nothing) also mutable. Augustine locates the origin of this going-wrong in the sinful misuse of freedom by some of the angels and then by the first humans. His theodicy is thus a blend of Neoplatonic and biblical themes and shows clearly the immense influence of Neoplatonism upon Christian thought during its early formative period.

The origin of evil

Augustine himself, together with Christian thinking as a whole, departed from Neoplatonism at one crucial point. Neoplatonism saw the world as continuous in being with the ultimate divine reality, the One. The One, in its limitless plenitude of being, overflows into the surrounding void, and the descending and attenuating degrees of being constitute the many-leveled universe. In contrast to this emanationist conception Augustine held that the universe is a created realm, brought into existence by God out of nothing (ex nihilo). It has no independent power of being, or aseity, but is through and through contingent, absolutely dependent upon the creative divine power. Further, Augustine was clear that the *nihil* out of which God created was not any kind of preexistent matter or chaos, but that "out of nothing" simply means "not out of anything" (*De natura boni*). This understanding of creation, entailing the universe's total emptiness of independent self-existence and yet its ultimate goodness as the free expression of God's creative love, is perhaps the most distinctively Christian contribution to metaphysical thought. It goes beyond the earlier Hebraic understanding in making explicit the ex nihilo character of creation in contrast to the emanationism of the Neoplatonic thought-world. This basic Christian idea entails the value of creaturely life and of the material world itself, its dependence upon God, and the meaningfulness of the whole temporal process as fulfilling an ultimate divine purpose.

Augustine's view of creation ex nihilo

Modern Christian treatments of the idea of creation ex nihilo have detached it from a literal use of the Genesis creation myth. The idea of the total dependence of the universe upon God is neutral as to whether it had a temporal beginning; nor does it in any way preclude the development of the universe in its present phase from the "big bang" onward, including the evolution of the forms of life on Earth. Although creation ex nihilo (a term apparently first introduced into Christian discourse by Irenaeus in the 2nd century) remains the general Christian conception of the relation between God and the physical universe, some recent Christian thinkers have substituted the view (derived from Alfred North Whitehead and developed by Charles Hartshorne) that God, instead of being its transcendent Creator, is an aspect of the universe itself, being either the inherent creativity in virtue of which it is a living process or a deity of finite power who seeks to lure the world into ever more valuable forms.

Aristotelian themes in Christian philosophy

Although Neoplatonism was the major philosophical influence on Christian thought in its early period and has never ceased to be an important element within it, Aristotle was also always known, though at first only as a logician. But in the 12th and 13th centuries his writings on physics, metaphysics, and ethics became available in Latin, translated either from the Greek or from Arabic sources, and they were crucial for the greatest of the medieval Christian thinkers, St. Thomas Aquinas (*c.* 1225–74). One of the Aristotelian themes that influenced Thomas was that knowledge is not innate but is gained from the reports of the senses and from logical inference from self-evident truths. (Thomas, however, in distinction from Aristotle, added divinely revealed propositions to self-evident truths in forming his basis for inference.) Thomas also received from Aristotle the conception of metaphysics as the science of being. His doctrine of analogy, according to which statements about God are true analogically rather than univocally, was likewise inspired by Aristotle, as were his distinctions between act and potency, essence and existence, substance and accidents, and the active and passive intellect and his view of the soul as the "form" of the body.

Thomas Aquinas' system, however, was by no means simply Aristotle Christianized. He did not hesitate to differ from "the Philosopher," as he called him, when the Christian tradition required this; for whereas Aristotle had been concerned to understand how the world functions, Thomas was also concerned, more fundamentally, to explain why it exists.

With the gradual breakdown of the medieval worldview—its assumptions undermined by the Renaissance, the Reformation, the rise of modern science, and the spread of a new spirit of exploration and free inquiry—the nature of the philosophical enterprise began to change. The French thinker René Descartes (1596–1650) is generally regarded as the father of modern philosophy, and in the new movements of thought that began with him philosophy became less a matter of building and defending comprehensive metaphysical systems, or imagined pictures of the universe, and more a critical probing of presuppositions, categories of thought, and modes of reasoning and an inquiry into what it is to know, how knowledge and belief are arrived at in different areas of life, how well various kinds of beliefs are grounded, and how thought is related to language. There has long ceased to be a generally accepted philosophical framework, comparable with Neoplatonism, in terms of which Christianity can appropriately be expressed and defended. There is instead a plurality of philosophical perspectives and methods—analytic, positivist, phenomenological, idealist, pragmatist,

and existentialist. Thus modern Christianity, having inherited a body of doctrines developed in the framework of ancient worldviews that are now virtually defunct, lacks any philosophy of comparable status in terms of which to rethink its beliefs. In this situation some theologians have turned to existentialism, which is not so much a philosophical system as a hard-to-define point of view and style of thinking. Indeed, the earlier existentialists, such as the Danish philosopher Søren Kierkegaard (1813–55), vehemently rejected the idea of a metaphysical system—in particular, for 19th-century existentialists, the Hegelian system—though some later ones, such as the German philosopher Martin Heidegger (1889–1976), have developed their own systems. Existentialists are identified by the appearance in their writings of one or more of a number of loosely related themes. These include the significance of the concrete individual in contrast to abstractions and general principles; a stress upon human freedom and choice and the centrality of decision, and hence a view of religion as ultimate commitment; a preference for paradox rather than rational explanation; and the highlighting of certain special modes of experience that cut across ordinary consciousness, particularly a generalized anxiety or dread and the haunting awareness of mortality. Existentialists have been both atheists (*e.g.,* Friedrich Nietzsche and Jean-Paul Sartre) and Christians (*e.g.,* Kierkegaard, the Protestant Rudolf Bultmann, and the Roman Catholic Gabriel Marcel). It would be difficult to identify any doctrines that are common to all these thinkers. Existentialist themes have also been incorporated into systematic Christian theologies (*e.g.,* by John Macquarrie).

Influence of existentialism

Others have sought to construct theologies in the mold of 19th-century German idealism (*e.g.,* Paul Tillich); some, as process theologians, in that of the early 20th-century British mathematician and metaphysician Alfred North Whitehead (*e.g.,* Charles Hartshorne on the doctrine of God, John Cobb on Christology); some, the liberation theologians, in highly pragmatic and political terms (*e.g.,* Juan Luis Segundo, Gustavo Gutiérrez); and some, as feminist theologians, in terms of the newly awakened self-consciousness of women and the awareness of a distorting patriarchial influence on all past forms of Christian thought (*e.g.,* Rosemary Ruether, Elizabeth Fiorenza). Most theologians, however, have continued to accept the traditional structure of Christian beliefs. The more liberal among them have sought to detach these from the older conceptualities and to reformulate them so as to connect with modern consciousness (*e.g.,* Friedrich Schleiermacher, Albrecht Ritschl, Adolf von Harnack, Karl Rahner, Gordon Kaufman); while the more conservative have sought to defend the traditional formulations within an increasingly alien intellectual environment (*e.g.,* B.B. Warfield, Charles Hodge, Karl Barth, Cornelis Berkouwer).

Effect of the sciences on modern Christian thought

Of the factors forming the intellectual environment of Christian thought in the modern period, perhaps the most powerful have been the physical and human sciences. The former have compelled the rethinking of certain Christian doctrines, as astronomy undermined the assumption of the centrality of the Earth in the universe, as geologic evidence concerning its age rendered implausible the biblical chronology, and as biology located humanity within the larger evolution of the forms of life on Earth. The human sciences of anthropology, psychology, sociology, and historical research have suggested possible naturalistic explanations of religion itself based, for example, upon the projection of desire for a cosmic father figure, the need for socially cohesive symbols, or the power of royal and priestly classes. Such naturalistic interpretations of religion, together with the ever-widening scientific understanding of the physical universe, have prompted some Christian philosophers to think of the religious ambiguity of the universe as a totality that can, from the human standpoint within it, be interpreted in both naturalistic and religious ways, thus providing scope for the exercise of faith as a free response to the mystery of existence.

## FAITH AND REASON

Different conceptions of faith cohere with different views of its relation to reason or rationality. The classic medieval understanding of faith, set forth by Thomas Aquinas, saw it as the belief in revealed truths on the authority of God as their ultimate source and guarantor. Thus, though the ultimate object of faith is God, their revealer, its immediate object is the body of propositions articulating the basic Christian dogmas. Such faith is to be distinguished from knowledge. Whereas the propositions that are the objects of scientia, or knowledge, compel belief by their self-evidence or their demonstrability from self-evident premises, the propositions accepted by faith do not thus compel assent but require a voluntary act of trusting acceptance. As unforced belief, faith is "an act of the intellect assenting to the truth at the command of the will" (*Summa theologiae,* II/II, Q. 4, art. 5); and it is because this is a free and responsible act that faith is one of the virtues. It follows that one cannot have knowledge and faith at the same time in relation to the same proposition; faith can only arise in the absence of knowledge. Faith also differs from mere opinion, which is inherently changeable. Opinions are not matters of absolute commitment but allow in principle for the possibility of doubt and change. Faith, as the wholehearted acceptance of revealed truth, excludes doubt.

Faith distinguished from knowledge

In the wider context of his philosophy Thomas Aquinas held that human reason, without supernatural aid, can establish the existence of God and the immortality of the soul; though these are also revealed, for acceptance by faith, for the benefit of those who cannot or do not engage in such strenuous intellectual activity. Faith, however, extends beyond the findings of reason in accepting such further truths as the triune nature of God and the divinity of Christ. Thomas thus supported the general (though not universal) Christian view that revelation supplements, rather than cancels or replaces, the findings of sound philosophy.

From a skeptical point of view, which does not acknowledge divine revelation, this Thomist conception amounts to faith as belief that is unevidenced or that is stronger than the evidence warrants, the gap being filled by the believer's own will to believe. As such it attracts the charge that belief upon insufficient evidence is always wrong.

Faith as a rational wager

In response to this kind of attack the French philosopher Blaise Pascal (1623–62) proposed a voluntarist defense of faith as a rational wager. Pascal assumed, in disagreement with Thomas Aquinas but in agreement with much modern thinking, that divine existence can neither be proved nor disproved; and he reasoned that if one decides to believe in God and to act on this basis, one gains eternal life if right but loses little if wrong, whereas if one decides not to believe, one gains little if right but may lose eternal life if wrong. In these circumstances, he concluded, the rational course is to believe. The argument has been criticized theologically for presupposing an unacceptable image of God as rewarding such calculating worship and also on the philosophical ground that it is too permissive in that it could justify belief in the claims, however fantastic, of any person or group who threatened nonbelievers with damnation or other dangerous consequences.

The American philosopher William James (1842–1910) refined this approach by limiting it, among matters that cannot be determined by proof or evidence, to belief-options that one has some real inclination or desire to accept, carry momentous implications, and are such that a failure to choose constitutes a negative choice. Theistic belief is for many people such an option, and James claimed that they have the right to make the positive decision to believe and to proceed in their lives on that basis. Either choice involves unavoidable risks: on the one hand the risk of being importantly deluded and on the other the risk of missing a limitlessly valuable truth. In this situation each individual is entitled to decide which risk to run. This argument has also been criticized as being too permissive and as constituting in effect a license for wishful believing, but its basic principle can perhaps be validly used in the different context of opting to base beliefs upon one's religious experience.

The element of risk in faith as a free cognitive choice was emphasized, to the exclusion of all else, by Kierkegaard in his idea of the leap of faith. He believed that without

Kierkegaard's leap of faith

risk there is no faith, and that the greater the risk the greater the faith. Faith is thus a passionate commitment, not based upon reason but inwardly necessitated, to that which can be grasped in no other way.

The epistemological character of faith as assent to propositions, basic to the Thomist account, is less pronounced in the "betting one's life" conceptions of Pascal and James in that these accept not a system of doctrines but only the thought of God as existing—though of course that thought itself has conceptual and hence implicitly propositional content. Kierkegaard's self-constituting leap of faith likewise only implicitly involves conceptual and propositional thought. The same is true of the account of faith based upon Ludwig Wittgenstein's concept of seeing-as (*Philosophical Investigations,* 1953). Wittgenstein pointed to the epistemological significance of puzzle pictures, such as the ambiguous duck-rabbit that can be seen either as a duck's head facing one way or a rabbit's head facing another way. The enlarged concept of experiencing-as (developed by the British philosopher John Hick) refers to the way in which an object, event, or situation is experienced as having a particular character or meaning such that to experience it in this manner involves being in a dispositional state to behave in relation to the object or event, or within the situation, in ways that are appropriate to its having that particular character. All conscious experience is in this sense experiencing-as. The application of this idea to religion suggests that the total environment is religiously ambiguous, capable of being experienced in both religious and naturalistic ways. Religious faith is the element of uncompelled interpretation within the distinctively religious ways of experiencing—for theism, experiencing the world or events in history or in one's own life as mediating the presence and activity of God. In ancient Hebrew history, for example, events that are described by secular historians as the effects of political and economic forces were experienced by some of the great prophets as occasions in which God was saving or punishing, rewarding or testing, the Israelites. In such cases religious does not replace secular experiencing-as but supervenes upon it, revealing a further order of meaning in the events of the world. And the often unconscious cognitive choice whereby someone experiences religiously constitutes, on this view, faith in its most epistemologically basic sense.

For these voluntarist, existentialist, and experiential conceptions of faith the place of reason in religion, although important, is secondary. Reason cannot directly establish the truth of religious propositions. Its function is rather to defend the rational propriety of trusting one's deeper intuitions or one's religious experience and basing one's beliefs and life upon them. These schools of thought assume that the philosophical arguments for and against the existence of God are inconclusive, and that the universe is capable of being consistently thought of and experienced in both religious and naturalistic ways. This assumption, however, runs counter to the long tradition of natural theology.

## CHRISTIAN PHILOSOPHY AS NATURAL THEOLOGY

Natural theology is generally characterized as the project of establishing religious truths by rational argument and without reliance upon alleged revelations, its two traditional topics being the existence of God and the immortality of the soul.

Evidence of God in nature

**Arguments for the existence of God.** *The design (or teleological) argument.* St. Paul, with many others in the Greco-Roman world, believed that the existence of God is evident from the appearances of nature: "Ever since the creation of the world his invisible nature, namely, his eternal power and deity, has been clearly perceived in the things that have been made" (Romans 1:20). The most popular, because the most accessible, of the theistic arguments is that which identifies evidences of design in nature, inferring from them a divine designer. The argument was propounded by medieval Christian thinkers and was developed in great detail in 17th- and 18th-century Europe by such writers as Robert Boyle, John Ray, Samuel Clarke, and William Derham and at the beginning of the 19th century by William Paley. Such writers asked: Is not the eye as manifestly designed for seeing, and the ear for hearing, as a pen for writing or a clock for telling the time; and does not such design imply a designer? The fact that the universe as a whole is a coherent and efficiently functioning system likewise, in this view, indicates a divine intelligence behind it.

This kind of argument was powerfully criticized by the Scottish philosopher David Hume in his *Dialogues Concerning Natural Religion* (1779). Hume granted that the world constitutes a more or less smoothly functioning system; indeed, he points out, it could not exist otherwise. He suggests, however, that this may have come about as a result of the chance permutations of particles falling into a temporary or permanent self-sustaining order, which thus has the appearance of design. A century later the idea of order without design was rendered more plausible by Charles Darwin's discovery that the adaptations of the forms of life are a result of the natural selection of inherited characteristics having positive, and the elimination of those having negative, survival value within a changing environment. Hume also pointed out that, even if one could infer an intelligent designer of the world, one would not thereby be entitled to claim that such a designer is the infinitely good and powerful Creator who is the object of Christian faith. For the world is apparently imperfect, containing many inbuilt occasions of animal pain and human suffering, and one cannot legitimately infer a greater perfection in the cause than is observed in the effect.

Probability arguments

In the 20th century, however, the design argument has been reformulated in more comprehensive ways, particularly by the British philosophers Frederick R. Tennant (*Philosophical Theology,* 1928–30) and Richard Swinburne (using Bayes's probability theorem in *The Existence of God,* 1979), taking account not only of the order and functioning of nature but also of the "fit" between human intelligence and the universe, whereby one can understand its workings, as well as human aesthetic, moral, and religious experience. There are also attempts to show that the evolution of the universe, from the "big bang" of some 15,000,000,000 years ago to the present state that includes conscious life, required the conjunction of so many individually improbable factors as to be inexplicable except as the result of a deliberate coordinating control. If, for example, the initial heat of the expanding universe, or its total mass, or the strength of the force of gravity, or the mass of neutrinos, or the strength of the strong nuclear force, had been different by a small margin, there would have been no galaxies, no stars, no planets, and hence no life. Surely, it is argued, all this must be the work of God creating the conditions for human existence.

These probability arguments have, however, been strongly criticized. A basic consideration relevant to them all is that there is by definition only one universe, and it is difficult to see how its existence, either with or without God, can be assessed as having a specific degree of probability in any objective sense. It can of course be said that any form in which the universe might be is statistically enormously improbable as it is only one of a virtual infinity of possible forms. But its actual form is no more improbable, in this sense, than innumerable others. It is only the fact that humans are part of it that makes it seem so special, requiring a transcendent explanation. The design argument is thus an area in which debate continues.

*The cosmological argument.* St. Thomas Aquinas gave the first-cause argument and the argument from contingency—both forms of cosmological reasoning—a central place for many centuries in the Christian enterprise of natural theology. (Similar arguments were also used in parallel strands of Islāmic philosophy.) Thomas' formulations (*Summa theologiae,* I, Q. 2, art. 3) have been refined in modern neo-Thomist discussions and continue to be topics of Christian philosophical reflection.

Change and causation

The first-cause argument begins with the fact that there is change in the world. A change is always the effect of some cause or causes. Each cause is itself the effect of a further cause or set of causes; this chain moves in a series that either never ends or is completed by a first cause, which must be of a radically different nature in that it is not itself caused. Such a first cause is an important aspect, though not the entirety, of what Christianity means by God.

The argument from contingency follows by another route the same basic movement of thought from the nature of the world to its ultimate ground. It starts with the fact that everything in the world is contingent for its existence upon other factors. Its presence is thus not self-explanatory but can only be understood by reference beyond itself to prior or wider circumstances that have brought it about. These other circumstances are likewise contingent; they too point beyond themselves for the ground of their intelligibility. If this explanatory regress is unending, explanation is perpetually postponed and nothing is finally explained. The existence of anything and everything thus remains ultimately unintelligible. But rational beings are committed to the search for intelligibility and cannot rest content until it is found. The universe can only finally be intelligible as the creation of an ontologically necessary being who is eternal and whose existence is not contingent upon anything else. This is also part of what Christianity has meant by God.

Criticism of these arguments points to the possibility that there is no first cause because the universe had no beginning, having existed throughout time, and is thus itself the necessary being that has existed eternally and without dependence upon anything else. Proponents of the cosmological argument reply that the existence of such a universe, as a procession of contingent events without beginning, would still be ultimately unintelligible. On the other hand, a personal consciousness and will, constituting a self-existent Creator of the universe, would be intrinsically intelligible; for human beings have experience in themselves of intelligence and free will as creative. Critics respond that insofar as the argument is sound it leaves one with the choice between believing that the universe is ultimately intelligible, because created by a self-existent personal will, or accepting that it is finally unintelligible, simply the ultimate given brute fact. The cosmological argument does not, however, compel one to choose the first alternative; logically, the second remains equally possible.

*The ontological argument.* The ontological argument, which proceeds not from the world to its Creator but from the idea of God to the reality of God, was first clearly formulated by St. Anselm (1033/34–1109) in his *Proslogion* (1077–78). Anselm began with the concept of God as that than which nothing greater can be conceived (*aliquid quo nihil majus cogitari possit*). To think of such a being as existing only in thought and not also in reality involves a contradiction. For an X that lacks real existence is not that than which no greater can be conceived. A yet greater being would be X with the further attribute of existence. Thus the unsurpassably perfect being must exist—otherwise it would not be unsurpassably perfect.

The necessity of God's existence

This argument has intrigued philosophers ever since. After some discussion in the 13th century it was reformulated for the modern world by Descartes in his *Meditations* (1641). Descartes made explicit the assumption, implicit in Anselm's reasoning, that existence is an attribute that a given X can have or fail to have. It follows from this—together with the assumption that existence is an attribute that is better to have than to lack—that God, as unsurpassably perfect, cannot lack the attribute of existence.

It was the assumption that existence is a predicate that has, in the view of most subsequent philosophers, proved fatal to the argument. The criticism was first made by Descartes's contemporary Pierre Gassendi and later and more prominently by the German philosopher Immanuel Kant (1724–1804) in his *Critique of Pure Reason* (1781). Putting their point as it has come to be further clarified by Bertrand Russell and others in the 20th century, to say that something with stated properties (whether it be a triangle, defined as a three-sided plane figure, or God, defined as an unsurpassably perfect being) exists, is not to attribute to it a further property, namely existence, but is to assert that the concept (of a triangle, or of God) is instantiated—that there actually are instances of that concept. But whether or not a given concept is instantiated is a question of fact. It cannot be determined a priori but only by whatever is the appropriate method for discovering a fact of that kind. This need for, in the broadest sense, observation cannot be circumvented by writing existence into the definition of the concept ("an existing three-sided plane figure," "an existing unsurpassably perfect being"), for the need arises again as the question of whether this enlarged concept is instantiated.

In the 20th century several Christian philosophers (notably Charles Hartshorne, Norman Malcolm, and Alvin Plantinga) have rediscovered and claimed validity for a second form of Anselm's argument. This hinges upon "necessary existence," a property with even higher value than "existence." A being that necessarily exists cannot coherently be thought not to exist. And so God, as the unsurpassably perfect being, must have necessary existence—and therefore must exist. This argument, however, has been criticized as failing to observe the distinction between logical and ontological, or factual, necessity. Logically necessary existence, it is said, is an incoherent idea, for logical necessity applies to the relations between concepts, not to their instantiation. God's necessity, then, must be an ontologically, or factually, rather than a logically, necessary existence: God exists as the ultimate fact, without beginning or end and without depending upon anything else for existence. But whether this concept of an ontologically necessary being is instantiated cannot be determined a priori. It cannot be validly inferred from the idea of an eternal and independent being that there actually is such a being.

Logical and ontological necessity

*Moral arguments.* Moral theistic argument belongs primarily to the modern world and perhaps reflects the modern lack of confidence in metaphysical constructions. Kant, having rejected the cosmological, ontological, and design proofs, argued in the *Critique of Practical Reason* (1788) that the existence of God, though not directly provable, is a necessary postulate of the moral life. To take seriously the awareness of a categorical imperative to act rightly is to commit oneself to work for an ideal state of affairs in which perfect goodness and happiness coincide. But as this universal apportioning of happiness to virtue is beyond human power, a divine agent capable of bringing it about must be assumed.

Other Christian thinkers, particularly during the 19th and early 20th centuries, have developed the theme that to accept the absolute demands of ethical obligation is to presuppose that this is a morally structured universe; and that this in turn implies a personal God whose commands are reflected in the human conscience. It cannot be proved that this is such a universe, it is said, but it is inevitably assumed in acknowledging the claims of morality.

The basic criticism of all attempts to trace ethical obligation to a transcendent divine source has been that it is possible to account for morality without going beyond the human realm. It is argued that the exigencies of communal life require agreed codes of behaviour, which become internalized in the process of socialization as moral laws; and the natural affection that develops among humans produces the more occasional sense of a call to heroic self-sacrifice on behalf of others. It seems, then, that the moral arguments for divine existence do not rise to the level of strict proofs.

*Arguments from religious experience and miracles.* Religious experience is used in Christian apologetics in two ways—in the argument from religious experiences to God as their cause and in the claim that it is (in the absence of contrary indications) as reasonable to trust religious as it is to trust nonreligious experience in forming beliefs about the total environment. (The first use is considered here among the traditional theistic arguments; for the second, see below *Contemporary discussions.*)

The argument maintains that special episodes, such as seeing visions of Christ or Mary or hearing a voice speaking with apparently divine authority, as well as the more pervasive experience of "living in God's presence" or of "absolute dependence upon a higher power," constitute evidence of God as their source. The criticism of this reasoning is that although such experiences may be accepted as having occurred, their cause might be purely natural. To establish that the experiences are real, as experiences, is not to establish that they are caused by an infinite, omnipotent, omniscient, divine being. As Thomas Hobbes succinctly put it, when someone says that God has spoken

to him in a dream, this "is no more than to say he dreamed that God spake to him" (*Leviathan,* Pt. III, ch. 32).

God as the agent of miracles

The analogous argument, from miracles to God as their cause, is more complex, involving two sets of problems. The argument may assert that the children of Israel were miraculously rescued from Egypt or Jesus was miraculously raised from the dead and therefore that God must exist as the agent of these miracles. The first problem concerns the reports. Whereas in the case of private religious experiences the skeptic (to whom the argument is addressed) may well be willing to grant that such experiences occurred, in the case of public miracles the skeptic will require adequate evidence for the described event; and this is not forthcoming for the classic miracle stories referring to alleged extraordinary events of many centuries ago. There are, however, well-evidenced contemporary and recent accounts of "miraculous" healings and other remarkable happenings. On the assumption that some of these, and also some of the classic miracle stories, are historically accurate, the second problem arises. How can it be established that these events were caused by supernatural divine intervention rather than by the operation of natural psychic laws, such as seem to be indicated by the phenomena of telepathy and telekinesis?

Once again, any kind of strict proof seems to be lacking. These arguments can, however, still be seen as displaying aspects of the explanatory power of the idea of God. Divine activity is not the only possible way of understanding the ordered and developing character of the universe, its contingent existence, the unconditional claims of morality, or the occurrence of religious experiences and "miracles." Nevertheless, the concept of deity offers a possible, satisfying answer to the fundamental questions to which these various factors point. They may thus be said to open the door to rational theistic belief—but still leaving the nonbeliever waiting for a positive impetus to go through that door. Some of the contemporary work by Christian philosophers has been in search of such a positive impetus.

**The immortality of the soul.** Human beings seem always to have had some notion of a shadowy double that survives the death of the body. But the idea of the soul as a mental entity, with intellectual and moral qualities, interacting with a physical organism but capable of continuing after its dissolution, derives in Western thought from Plato and entered into Judaism during approximately the last century before the Common Era and thence into Christianity. In Jewish and Christian thinking it has existed in tension with the idea of the resurrection of the person conceived as an indissoluble psychophysical unity. Christian thought gradually settled into a pattern that required both of these apparently divergent ideas. At death the soul is separated from the body and exists in a conscious or unconscious disembodied state. But on the future Day of Judgment souls will be re-embodied (whether in their former but now transfigured earthly bodies or in new resurrection bodies) and will live eternally in the heavenly kingdom.

The indestructibility of the soul

Within this framework philosophical discussion has centred mainly on the idea of the immaterial soul and its capacity to survive bodily death. Plato, in the *Phaedo,* argued that the soul is inherently indestructible. To destroy something, including the body, is to disintegrate it into its constituent elements; but the soul, as a mental entity, is not composed of parts and is thus an indissoluble unity. Although Thomas Aquinas' concept of the soul, as the "form" of the body, was derived from Aristotle rather than Plato, he too argued for its indestructibility (*Summa theologiae,* I, Q. 76, art. 6). The French philosopher Jacques Maritain (1882–1973), a modern Thomist, summarized the conclusion as follows: "A spiritual soul cannot be corrupted, since it possesses no matter; it cannot be disintegrated, since it has no substantial parts; it cannot lose its individual unity, since it is self-subsisting, nor its internal energy since it contains within itself all the sources of its energies" (*The Range of Reason,* 1952). But though it is possible to define the soul in such a way that it is incorruptible, indissoluble, and self-subsisting, critics have asked whether there is any good reason to think that souls as thus defined exist. If, on the other hand, the soul means the conscious mind or personality—something whose immortality would be of great interest to human beings—this does not seem to be an indissoluble unity. On the contrary, it seems to have a kind of organic unity that can vary in degree but that is also capable of fragmentation and dissolution.

Much modern philosophical analysis of the concept of mind is inhospitable to the idea of immortality, for it equates mental life with the functioning of the physical brain (see MIND, THE PHILOSOPHY OF). Impressed by evidence of the dependence of mind on brain, some Christian thinkers have been willing to accept the view—corresponding to the ancient Hebrew understanding—of the human being as an indissoluble psychophysical unity, but these thinkers have still maintained a belief in immortality, not as the mind surviving the body, but as a divine resurrection or re-creation of the living body-mind totality. Such resurrection persons would presumably be located in a space different from that which they now inhabit and would presumably undergo a development from the condition of a dying person to that of a viable inhabitant of the resurrection world. But all theories in this area carry with them their own difficulties, and discussion continues.

Kant offered a different kind of argument for immortality—as a postulate of the moral life. The claim of the moral law demands that human beings become perfect. This is something that can never be finally achieved but only asymptotically approached, and such an unending approach requires the unending existence of the soul. This argument also is open to criticism. Are humans indeed subject to a strict obligation to attain moral perfection? Might not their obligation, as finite creatures, be to do the best they can? But this does not seem to entail immortality.

It should be noted that in the case of all these arguments, both for the immortality of the soul and for the existence of God, the debate has been as much among Christian philosophers as between them and non-Christian skeptics. It is by no means the case that Christian thinkers have all regarded the project of natural theology as viable. There have indeed been, and are, many who hold that divine existence can be definitively proved or shown to be objectively probable. But there are many others who not only hold that the attempted proofs all require premises that a disbeliever is under no rational obligation to accept but who also question the evidentialist assumption that the only route to rational theistic belief is by inference from previously accepted evidence-stating premises.

### CONTEMPORARY DISCUSSIONS

Contemporary discussion among Christian philosophers is predominantly epistemological. Among Roman Catholic thinkers it includes the original work of Bernard Lonergan in *Insight* (1957), which has stimulated considerable discussion. Lonergan argued that the act of understanding, or insight, is pivotal for the apprehension of reality, and that it implies in the long run that the universe is itself due to the fiat of an "unrestricted act of understanding," which is God. Other Roman Catholic thinkers have continued to refine and extend the Thomistic approach, particularly the idea of analogical predication in statements about God. Others, in common with non-Catholic philosophers, have discussed the traditional divine attributes—omniscience, omnipotence, eternity, immutability, personality, goodness. The concept of a finite deity developing through time has also been proposed (*e.g.,* by Charles Hartshorne) to meet objections to some of these concepts: If God is immutable, how can God be aware of successive events in time? If God has absolute self-existence, how can God respond with sympathy to the pains of creaturely life? Others have defended the traditional attributes as logically coherent, both individually and in their relationship to one another, and as allowing for divine awareness of the created universe, God's activity in history, and divine sympathy with human suffering.

Influence of logical positivism

Perhaps the largest body of work, however, has been generated in dialogue with the new linguistic turn of philosophy in the English-speaking world, particularly since World War II, concentrating on the analysis of language in its various uses. The logical positivist movement originated in the 1920s with the Vienna Circle. Although

mainly concerned with the philosophy of science, it posed by implication a major challenge to the logical meaningfulness of religious language. The positivist position, in its developed form, was that a statement has factual meaning only if it is capable in principle of being verified or falsified, or at least in some degree confirmed or disconfirmed, within human experience; otherwise it is meaningless, or cognitively vacuous. In the years immediately after World War II this account of factual meaning was applied (*e.g.,* by Antony Flew) to theological statements, raising such questions as: What observable difference does it make whether it is true or false that "God loves us"? Whatever tragedies occur, do not the faithful still maintain their belief, adding perhaps that the divine love is beyond human comprehension? But if it is not possible to conceive of circumstances in which "God loves us" would have to be judged false, is not the statement factually empty, or meaningless?

This challenge evoked three kinds of response. Some Christian philosophers have declared it to be a non-challenge, on the ground that the positivists never succeeded in finding a precise formulation of the verification criterion that was fully satisfactory even to themselves. Others have held that this does not block the central thrust of the positivist challenge. Does it really make no difference within actual or possible human experience whether or not God exists and loves us; and if so, is not the significance of the belief thereby fatally damaged? Among those who felt it necessary to face this challenge, one group granted that theological statements lack factual meaning and suggested that their proper use lies elsewhere, as expressing a way of looking at the world (*e.g.,* Richard M. Hare) or a moral point of view and commitment (*e.g.,* R.B. Braithwaite). The other group claimed that theism is ultimately open to experiential confirmation. The theory of eschatological verification (developed by John Hick) holds that the belief in future postmortem experiences will be verified if true (though not falsified if false), and that in a divinely governed universe such experiences will take forms confirming theistic faith. Thus although the believer and the disbeliever do not have different expectations about the course of earthly history, they do expect the total course of the universe to be radically different.

In the late 20th century, under the stimulus of Wittgenstein's posthumously published works, attention has been directed to the multiple legitimate uses of language in the various language games developed within different human activities and forms of life; and it has been urged (*e.g.,* by D.Z. Phillips) that religious belief has its own autonomous validity, not subject to verificationist or scientific or other extraneous criteria. Statements about God and eternal life do not make true-or-false factual claims but express, in religious language, a distinctive attitude to life and way of engaging in it. This suggestion forms part of the broader non-realist interpretation of religion, holding that its beliefs do not refer to putative transcendent realities but are instead expressive of human ideals, desires, hopes, attitudes, and intentions. Such thinking goes back to the German philosopher Ludwig Feuerbach (*The Essence of Christianity,* 1841) in the 19th century and to George Santayana, John Dewey, and J.H. Randall, Jr., in the early 20th century and is advocated today by some Christian writers, notably D.Z. Phillips and Don Cupitt. According to them, true Christianity consists in the inner purity of an unself-centred attitude to life and does not involve belief in the objective reality of God or of a life after death. The criticism this view inevitably attracts is that to deny the transcendent reference of religious language empties it of any substantial meaning. The issue is the focus of considerable contemporary discussion.

In addition to this and other work concerning religious language there has been a renewal of fundamental discussion of Christian, and more broadly religious, epistemology. The natural theology tradition held that, in order to be rational, religious belief must be supported by adequate evidences or arguments. It was assumed that God's existence must be validly inferred from generally acceptable premises. This evidentialist principle has been questioned, however, not only by such earlier thinkers as Pascal and William James but also by a number of contemporary Christian philosophers. Evidentialist thinking was foundationalist in granting that there are some beliefs that can be reasonably held directly and not by inference from other evidence-stating beliefs. Thomas Aquinas, for example, recognized self-evident truths and the reports of the senses as basic in the sense that they do not need support from other beliefs. They thus provide the foundations on which a belief structure can properly be built. Belief in the existence of God was not regarded as basic or foundational in this way but was thought to require adequate evidence or arguments. It has been argued (by Alvin Plantinga) that the range of properly basic beliefs is wider than classic foundationalism had recognized. It can include not only beliefs about the past and the existence of other persons but also belief in the reality of God. Such beliefs can be basic (*i.e.,* not inferred), and they are properly basic if held in appropriate circumstances. Thus, the belief that "There is a tree before me" is properly basic for one who is having the experience of seeing a tree; and the belief that "God exists" is properly basic for one who experiences God's judgment, forgiveness, love, claim, providential care, or some other mode of divine presence.

Challenges to the evidentialist principle

Discussion of this proposal centres upon the criteria for proper basicality: In what circumstances is it appropriate, and in what circumstances not, to hold the basic belief in God or the basic beliefs of other religions or of the naturalistic worldviews?

A related contemporary development, pursued by William Alston and others, is the claim that religious experience constitutes an entirely proper basis for religious beliefs. The claim is not that one can validly infer God as the cause of theistic religious experience, but that one who participates in such experience is entitled to trust it as a ground for belief. It is argued that human beings all normally operate with a "principle of credulity" whereby they take what seems to be so as indeed so, unless they have some positive reason to doubt it. Accordingly, one who has the experience of living in the presence of God can properly proceed in both thought and life on the basis that God is real. Such belief inevitably involves epistemic risk—the risk of error versus the risk of missing the truth. But perhaps the right to believe that was defended by William James applies in this situation.

The principle of credulity

The discussion focuses on the analogies between religious forms of experience and the kinds of sensory experience in relation to which the principle of credulity is virtually universally accepted. It is uncontroversially proper to hold beliefs reflecting sense experience, but what of beliefs reflecting religious experience? Whereas all human beings hold the former and could not survive without doing so, the latter type of belief seems to be optional. Although beliefs regarding physical objects can be empirically confirmed or disconfirmed, religious beliefs cannot. Acknowledging these significant differences, some Christian philosophers have argued that they are the kinds of differences that are to be expected, given the difference between the human relationship to the world and to God. It is necessary to human existence as physical organisms that a consciousness of the material environment should be forced upon human beings. On the other hand, it is necessary for existence as relatively autonomous and responsible beings that consciousness of God should not be forced upon them, for to be compulsorily aware of God's universal presence as limitless goodness and power, making a total claim upon human life, would deprive them of creaturely freedom. Humans are accordingly set at an epistemic distance from God that is overcome only by faith, which can be identified with the voluntary interpretive element within the experience of God's presence.

The central Christian doctrine of the divinity of Christ is another topic of current discussion. Philosophical questions concerning this were debated intensively in the 3rd to 5th centuries, as noted above, in terms of the key notion of ousia/substantia. The concept of substance, however, although confidently used throughout the medieval period, has been widely questioned within modern thought and no longer figures in the distinctively 20th-century streams of philosophy. There have consequently

been a variety of attempts, in which theology and philosophy mingle inextricably, to find an interpretation that is intelligible today. Instead of the basically static notion of substance—Jesus qua human being of human substance and qua divine of God's substance—many have preferred the more dynamic idea of divine action. From this point of view Jesus was divine in the sense that God was acting redemptively through him; or, instead of a *homo-ousion,* identity of substance, between Jesus and the heavenly Father, there was a *homo-agapion,* an identity of divine loving. Others, however, have criticized such alternatives to the older substance language, often on the ground that, whereas "being of the same substance as" is an all-or-nothing concept, divine activity in and through a human life is capable of degrees, so that the divinity of Christ may in principle be de-absolutized. The intertwining theological and philosophical issues continue to be strongly debated.

Questions of religious pluralism

The problems of religious pluralism are increasingly being seen as requiring the attention of Christian philosophers. One reason arises from the kind of apologetic described above, hinging upon the reasonableness of basing beliefs upon religious experience. It is evident that there are many forms of religious experience, giving rise to many forms of religious belief. There is considerable variety within the Christian tradition itself, and in the world as a whole Muslim forms of religious experience give rise to and justify Islāmic beliefs, Jewish forms of experience to Jewish beliefs, Hindu to Hindu beliefs, Buddhist to Buddhist beliefs, and so on. These different belief systems include mutually incompatible doctrines. Thus the experiential solution to the problem of justifying Christian beliefs has given rise to a new problem constituted by the conflicting truth-claims of the different religious traditions.

The other reason the great world faiths provide new issues for Christian philosophy is that some of their belief systems challenge long-standing Christian assumptions. Whereas Judaism and Islām raise theological questions, the most challenging philosophical issues are raised by Buddhism. The belief in God as the personal ultimate is challenged by the idea of the ultimacy of the nonpersonal *dharma-kāya.* The idea of the immortal soul is challenged by the *anattā* ("no soul") doctrine, with its claim that the personal mind or soul is not an enduring substance but a succession of fleeting moments of consciousness. And yet Buddhism, teaching as it does doctrines that are radically different from those of the Christian faith, also challenges Christianity by the centrality within it of compassion, peaceableness, and a respect for all life.

These and other issues raised by the fact of religious plurality are ones that Christian philosophers have only begun to face but that suggest the possibility of major developments in Christian thinking. (J.Hi.)

## Christian mysticism

Christian mysticism refers to the human being's direct experience or consciousness of ultimate reality, understood as God within the context of Christian faith. The essence of mysticism is the sense of some form of contact with the divine or transcendent, frequently understood in its higher forms as involving union with God. Mysticism has played an important role in the history of Christian religion, and it has once again become a noticeably living influence in recent times.

In the modern period mysticism has been studied from many perspectives: psychological, comparativist, philosophical, and theological, to name only the most vital. The mystical text has been the subject of new attention sparked by hermeneutical and deconstructionist philosophies. Among the theoretical questions that have been much debated are such issues as whether mysticism constitutes the core or essence of personal religion or whether it is better viewed as one element interacting with others in the formation of concrete religions. Those who emphasize a strong distinction between mystical experience and subsequent interpretation tend to search for a common core of all mysticism; others insist that experience and interpretation cannot be so easily sundered and that mysticism is in most cases tied to a specific religion and contingent upon its teachings. Both those who search for the common core, such as the British philosopher Walter T. Stace, and those who emphasize the differences among forms of mysticism, such as the British historian of religion Robert C. Zaehner, have made use of typologies of mysticism, often based on the contrast between introvertive and extrovertive mysticism developed by the comparativist Rudolf Otto. Studies have criticized the typological approach, but many scholars still find it useful.

The cognitive status of mystical knowing and its clash with the mystics' claims about the ineffability of their experiences have also been topics of interest for modern students of mysticism. Among the most important investigations of mystical knowing are those of the Belgian Jesuit Joseph Maréchal and the French philosophers Henri Bergson and Jacques Maritain.

The relation between mysticism and morality has been a topic of scholarly debate since the time of William James, but certain questions have concerned Christian mystics for centuries. Does mystical experience always confirm traditional religious ideas about right and wrong, or is mysticism totally independent of moral issues? The problems regarding mysticism are fairly easy to identify; definitive solutions seem far off.

Mysticism's role in Christianity

The role of mysticism in Christianity has been variously evaluated by modern theologians. Many Protestant thinkers, from Albrecht Ritschl and Adolf von Harnack through Karl Barth and Rudolf Bultmann, have denied mysticism an integral role in the Christian religion, claiming that mystical union was a Greek import incompatible with saving faith in the Gospel word. Other Protestant theologians, such as Ernst Troeltsch in *The Social Teaching of the Christian Churches* (trans. 1931) and Albert Schweitzer in *The Mysticism of Paul the Apostle* (trans. 1931), were more sympathetic. Anglican thinkers, especially William R. Inge, Evelyn Underhill, and Kenneth E. Kirk, championed the importance of mysticism in Christian history. Orthodox Christianity has given mysticism so central a role in Christian life that all theology in the Christian East by definition is mystical theology, as the Russian emigré thinker Vladimir Lossky showed in *The Mystical Theology of the Eastern Church* (trans. 1957).

The most extensive theological discussions of mysticism in Christianity have been found in modern Roman Catholicism. In the first half of the 20th century Neoscholastic authors—invoking the authority of Thomas Aquinas and the Spanish mystics Teresa of Ávila and John of the Cross—debated whether mystical contemplation was the goal of all Christians or a special grace offered only to a few. The discrimination of the various forms of prayer and the distinction between acquired contemplation, for which the believer could strive with the help of grace, and infused contemplation, which was a pure and unmerited gift, framed much of this discussion. Other Roman Catholic theologians, such as Cuthbert Butler in *Western Mysticism* (1922) and Anselm Stolz in *Theologie der Mystik* (1936), broke with the narrow framework of Neoscholasticism to consider the wider scriptural and patristic tradition. In the second half of the century Roman Catholic theologians including Karl Rahner and Hans Urs von Balthasar addressed key theological issues in mysticism, such as the relation of mystical experience to the universal offer of grace and the status of non-Christian mysticism.

### HISTORY OF CHRISTIAN MYSTICISM

**Early church.** Although the essence of mysticism is the sense of contact with the transcendent, mysticism in the history of Christianity should not be understood merely in terms of special ecstatic experiences but as part of a religious process lived out within the context of the Christian community. From this perspective mysticism played a vital part in the early church. Early Christianity was a religion of the spirit that expressed itself in the heightening and enlargement of human consciousness. It is clear from the Synoptic Gospels (*e.g.,* Matthew 11:25–27) that Jesus was thought to have enjoyed a sense of special contact with God. In the primitive church an active part was played by prophets, who were believed to be recipients of a revelation coming directly from the Holy Spirit.

The mystical aspect of early Christianity finds its fullest expression in the letters of Paul and the Gospel According to John. For Paul and John mystical experience and aspiration are always for union with Christ. It was Paul's supreme desire to know Christ and to be united with him. The recurring phrase, "in Christ," implies personal union, a participation in Christ's death and Resurrection. The Christ with whom Paul is united is not the man Jesus who is known "after the flesh." He has been exalted and glorified, so that he is one with the Spirit.

Christ-mysticism

Christ-mysticism finds renewed embodiment in the Gospel According to John, particularly in the farewell discourse (chapters 14–16), where Jesus speaks of his impending death and of his return in the Spirit to unite himself with his followers. In the prayer of Jesus in chapter 17 there is a vision of an interpenetrating union of souls in which all who are one with Christ share his perfect union with the Father.

In the early Christian centuries the mystical trend found expression not only in the stream of Pauline and Johannine Christianity (as in the writings of Ignatius of Antioch and Irenaeus of Lyon) but also in the Gnostics (early Christian heretics who viewed matter as evil and the spirit as good). Scholars still debate the origins of Gnosticism, but most Gnostics thought of themselves as followers of Christ, albeit a Christ who was pure spirit. The mysticism of the Gnostics can be seen in the religion of Valentinus, who was excommunicated in about AD 150. He believed that human beings are alienated from God because of their spiritual ignorance; Christ brings them into the gnosis (esoteric revelatory knowledge) that is union with God. Valentinus held that all human beings come from God and that all will in the end return to God. Other Gnostic groups held that there were three types of people—"spiritual," "psychic," and "material"—and that only the first two can be saved. The *Pistis Sophia* (3rd century) is preoccupied with the question of who finally will be saved. Those who are saved must renounce the world completely and follow the pure ethic of love and compassion. They will then be identified with Jesus and become rays of the divine Light.

**Eastern Christianity.** The classic forms of Eastern Christian mysticism appeared toward the end of the 2nd century, when the mysticism of the early church began to be expressed in categories of thought explicitly dependent on the Greek philosophical tradition of Plato and his followers. This intermingling of primitive Christian themes with Greek speculative thought has been variously judged by later Christians, but contemporaries had no difficulty in seeing it as proof of the new religion's ability to adapt and transform all that was good in the world. The philosophical emphasis on the unknowability of God found an echo in many texts of the Old and New Testaments, affirming that the God of Abraham and the Father of Jesus could never be fully known. The understanding of the role of the preexistent Logos, or Word, of the Gospel According to John in the creation and restoration of the universe was clarified by locating the Platonic conception of Ideas in the Logos. Greek emphasis on the vision or contemplation (*theōria*) of God as the goal of human blessedness found a scriptural warrant in the sixth Beatitude: "Blessed are the pure in heart, for they shall see God" (Matthew 5:8). The notion of deification (*theiosis*) fit with the New Testament emphasis on becoming sons of God and such texts as 2 Peter 1:4, which talked about sharing in the divine nature. These ecumenical adaptations later provided an entry for the language of union with God, especially after the notion of union became more explicit in Neoplatonism, the last great pagan form of philosophical mysticism. Many of these themes are already present in germ in the works of Clement of Alexandria, written in about 200. They are richly developed in the thought of Origen, the greatest Christian writer of the pre-Constantinian period and the earliest major speculative mystic in Christian history.

Origen and monasticism

Origen's mystical theology, however, required a social matrix in which it could take on life as formative and expressive of Christian ideals. This was the achievement of early Christian monasticism, the movement into the desert that began to transform ideals of Christian perfection at the beginning of the 4th century. The combination of the religious experience of the desert Christians and the generally Origenist theology that helped shape their views created the first great strand of Christian mysticism, one that remains central to the East and that was to dominate in the West until the end of the 12th century. Though not all the Eastern Christian mystical texts were deeply imbued with Platonism, all were marked by the monastic experience.

The first great mystical writer of the desert was Evagrius Ponticus (346–399), whose works were influenced by Origen. His writings show a clear distinction between the ascetic, or "practical," life and the contemplative, or "theoretical," life, a distinction that was to become classic in Christian history. His disciple, John Cassian, conveyed Evagrian mysticism to the monks of western Europe, especially in the exposition of the "degrees of prayer" in his *Collations of the Fathers,* or *Conferences.* Gregory of Nyssa, the younger brother of Basil, sketched out a model for progress in the mystical path in his *Life of Moses* and, following the example of Origen, devoted a number of homilies to a mystical interpretation of the Song of Solomon, showing how the book speaks both of Christ's love for the church and of the love between the soul and the Divine Bridegroom.

Perhaps the most influential of all Eastern Christian mystics wrote in the 5th or 6th century in the name of Dionysius the Areopagite, Paul's convert at Athens. He was probably a Syrian monk. In the chief works of this Pseudo-Dionysius, *Mystical Theology* and *On the Divine Names,* the main emphasis was on the ineffability of God ("the Divine Dark") and hence on the "apophatic" or "negative" approach to God. Through a gradual process of ascension from material things to spiritual realities and an eventual stripping away of all created beings in "unknowing," the soul arrives at "union with Him who transcends all being and all knowledge" (*Mystical Theology,* chapter 1). The writings of the Pseudo-Dionysius also popularized the threefold division of the mystical life into purgative, illuminative, and unitive stages. Later Eastern mystical theologians, especially Maximus the Confessor in the 7th century, adopted much of this thought but corrected it with greater Christological emphasis, showing that union with God is possible only through the action of the God-man.

Eastern mystics distinguish between the essence of God and divine attributes, which they regard as energies that penetrate the universe. Creation is a process of emanation, whereby the divine Being is "transported outside of Himself . . . to dwell within the heart of all things . . . " (Pseudo-Dionysius the Areopagite, *On the Divine Names,* iv. 13). The divinization of humanity is fundamental to Eastern mysticism.

Divinization of humanity

Divinization comes through contemplative prayer, and especially through the method of Hesychasm (from *hesychia,* "stillness") adopted widely by the Eastern monks. The method consisted in the concentration of the mind on the divine Presence, induced by the repetition of the "Jesus-prayer" (later formalized as "Lord Jesus Christ, Son of God, have mercy on me a sinner"). This culminated in the ecstatic vision of the divine Light and was held to divinize the soul through the divine energy implicit in the name of Jesus. Much of this program can already be found in the writings of Symeon the New Theologian (*c.* 949–1022), a monk of Constantinople. It reached its theologically most evolved form in Gregory Palamas (1296–1359), who defended the Hesychast tradition against its opponents. This rich form of Christian mysticism found a new centre in the Slavic lands after the conquest of the Greek East by the Turks. It experienced a flowering in Russia, beginning with the *Philokalia,* an anthology of ascetical and mystical texts first published in 1782, and continuing to the Revolution of 1917. Eastern Christian mysticism is best known in the West through translations of the anonymous 19th-century Russian text *The Way of the Pilgrim,* but noted Russian mystics, such as Seraphimof Sarov (1759–1833) and John of Kronshtadt (1829–1909), are gradually becoming better known in the West.

In the Eastern as in the Western Church mystical reli-

gion received at times heretical expressions. These trends begin with the Messalians (Syriac for "praying people") of the 4th century, who were accused of neglecting the sacraments for ceaseless prayer and of teaching a materialistic vision of God. Later mystics, both orthodox and suspect, have been accused of Messalianism. Other mystic sects grew up in Russia. The Dukhobors, who originated in the 18th century among the peasants, resemble the Quakers in their indifference to outer forms, standing for the final authority of the Inner Light. They were severely persecuted in Russia and migrated to Canada early in the 20th century. (S.Sp./B.J.McG.)

**Western Catholic Christianity.** The founder of Latin Christian mysticism is Augustine, bishop of Hippo (354–430). In his *Confessions* Augustine mentions two experiences of "touching" or "attaining" God. Later, in the *Literal Commentary on Genesis,* he introduced a triple classification of visions—corporeal, spiritual (*i.e.,* imaginative), and intellectual—that influenced later mystics for centuries. Although he was influenced by Neoplatonist philosophers such as Plotinus, Augustine did not speak of personal union with God in this life. His teaching, like that of the Eastern Fathers, emphasized the ecclesial context of Christian mysticism and the role of Christ as mediator in attaining deification, or the restoration of the image of the Trinity in the depths of the soul. The basic elements of Augustine's teaching on the vision of God, the relation of the active and contemplative lives, and the sacramental dimension of Christian mysticism were summarized by Pope Gregory I the Great in the 6th century and conveyed to the medieval West by many monastic authors.

Two factors were important in the development of this classic Augustinian form of Western mysticism. The first was the translation of the writings of Pseudo-Dionysius the Areopagite and other Eastern mystics by the 9th-century thinker Johannes Scotus Erigena. In combining the Eastern and Western mystical traditions, Erigena created the earliest version of a highly speculative negative mysticism that was later often revived. The other new moment began in the 12th century when new forms of religious life burst on the scene, especially among monks and those priests who endeavoured to live like monks (the canons). The major schools of 12th-century mysticism were inspired by new trends in monastic piety, especially those introduced by Anselm of Canterbury, but they developed these in a systematic fashion unknown to previous centuries. The great figures of the era, especially Bernard of Clairvaux among the Cistercians and Richard of Saint-Victor among the canons, have remained the supreme teachers of mystical theology in Catholic Christianity, along with the Spanish mystics of the 16th century.

Monastic piety

What the Cistercian and Victorine authors contributed to the development of Catholic mysticism was, first, a detailed study of the stages of the ascent of the soul to God on the basis of a profound understanding of the human being as the image and likeness of God (Genesis 1:26) and, second, a new emphasis on the role of love as the power that unites the soul to God. Building on both Origen and Augustine, Bernard and his contemporaries made affective, or marital, union with God in oneness of spirit (1 Corinthians 6:17) a central theme in Western mysticism, though along with Gregory the Great they insisted that "love itself is a form of knowing," that is, of vision or contemplation of God.

The great mystics of the 12th century contributed to an important expansion of mysticism in the following century. For the first time mysticism passed beyond the confines of the monastic life, male writers, and the Latin language. This major shift is evident not only in the life of Francis of Assisi, who emphasized the practical following of Jesus and came to be identified with him in a new form of Christ-mysticism, manifested in his reception of the stigmata, or wounds of the crucified Christ, but also in the remarkable proliferation of new forms of religious life and mystical writing in the vernacular on the part of women. Though female mystics such as Hildegard von Bingen were not unknown in the 12th century, the 13th century witnessed a flowering of interest in mysticism among women, evident in the Flemish Hadewijch of Brabant, the German Mechthild von Magdeburg, the French Marguerite Porete, and the Italians Clare of Assisi and Ângela da Foligno.

Among the important themes of the new mysticism of the 13th century was a form of Dionysian theology in which the stage of divine darkness surpassing all understanding was given a strong affective emphasis, as well as the emergence of an understanding of union with God that insisted upon a union of indistinction in which God and the soul become one without any medium. The first of these tendencies is evident in the writings of Bonaventure, the supreme master of Franciscan mysticism; the second is present in some of the women mystics but finds its greatest proponent in the Dominican Meister Eckhart, who was condemned for heresy in 1329.

Eckhart's mysticism

Eckhart taught that "God's ground and the soul's ground is one ground," and the way to the realization of the soul's identity with God lay less in the customary practices of the religious life than in a new state of awareness achieved through radical detachment from all created things and a breakthrough to the God beyond God. Though Eckhart's thought remained Christological in its emphasis on the necessity for the "birth of Son in the soul," his expressions of the identity between the soul that had undergone this birth and the Son of God seemed heretical to many. Without denying the importance of the basic structures of the Christian religion, and while insisting that his radical preaching to the laity was capable of an orthodox interpretation, Eckhart and the new mystics of the 13th century were a real challenge to traditional Western ideas of mysticism. Their teaching seemed to imply an autotheism in which the soul became identical with God, and many feared that this might lead to a disregard of the structures and sacraments of the church as the means to salvation and even to an antinomianism that would view the mystic as exempt from the moral law. The Council of Vienne condemned such errors in 1311, shortly after Marguerite Porete was burned as a heretic for continuing to disseminate her book, *The Mirror of Simple Souls.* The council associated these views with the Beguines, groups of religious women who did not live in cloister or follow a recognized rule of life. In the centuries that followed, some mystics were condemned and others executed on this basis, though evidence for a widespread "mystical heresy" is lacking.

The great mystical writers of the late Middle Ages, however, took pains to prove their orthodoxy. Eckhart's followers among the Rhineland mystics, especially Heinrich Suso and Johann Tauler, defended his memory but qualified his daring language. Texts such as the anonymous *Theologia Germanica* of the late 14th century, which reflects the ideas of the loose groups of mystics who called themselves the Friends of God, conveyed this German mysticism to the Reformers. In the Low Countries, the rich mystical literature that developed reached its culmination in writings of Jan van Ruysbroeck (1293–1381). In Italy two remarkable women, Catherine of Siena in the 14th century and Catherine of Genoa in the 15th, made important contributions to the theory and practice of mysticism. The 14th century also saw the "Golden Age" of English mysticism, as conveyed in the writings of the hermit Richard Rolle; the canon Walter Hilton, who wrote *The Scale* (or *Ladder*) *of Perfection;* the anonymous author of *The Cloud of Unknowing;* and his contemporary, the visionary recluse Mother Julian of Norwich, whose *Revelations of Divine Love* is unsurpassed in English mystical literature. Julian's meditations on the inner meaning of her revelations of the crucified Christ express the mystical solidarity of all humanity in the Redeemer, who is conceived of as a nurturing mother.

16th-century Spanish mystics

In the 16th century the centre of Roman Catholic mysticism shifted to Spain, the great Roman Catholic power at the time of the Reformation. Important mystics came both from the traditional religious orders, such as Francis de Osuna among the Franciscans, Luis de León among the Augustinians, and Luis de Grenada among the Dominicans, and from the new orders, as with Ignatius of Loyola, the founder of the Jesuits. The two pillars of Spanish mysticism, however, were Teresa of Avila (1515–

82) and her friend John of the Cross (1542–91), both members of the reform movement in the Carmelite order. Teresa's *Life* is one of the richest and most convincing accounts of visionary and unitive experiences in Christian mystical literature; her subsequent synthesis of the seven stages on the mystical path, *The Interior Castle,* has been used for centuries as a basic handbook. John of the Cross was perhaps the most profound and systematic of all Roman Catholic mystical thinkers. His four major works, *The Dark Night of the Soul, The Ascent of Mount Carmel, The Spiritual Canticle,* and *The Living Flame of Love,* constitute a full theological treatment of the active and passive purgations of the sense and the spirit, the role of illumination, and the unification of the soul with God in spiritual marriage.

In the 17th century France took the lead with figures such as Francis of Sales, Pierre de Bérulle, Brother Lawrence (the author of *The Practice of the Presence of God*), and Marie Guyard. At this time concentration on the personal experience of the mystic as the source for "mystical theology" (as against the common scriptural faith and sacramental life of the church) led to the creation of mysticism as a category and the description of its adherents as mystics. This century also saw renewed conflict over mysticism with the rise of the Quietist controversy. A Spaniard resident in Rome, Miguel de Molinos, author of the popular *Spiritual Guide* (1675), was condemned for his doctrine of the "One Act," that is, the teaching that the will, once fixed on God in contemplative prayer, cannot lose its union with the divine. In France Mme Guyon and her adviser, François Fénelon, archbishop of Cambrai, were also condemned for Quietist tendencies emphasizing the role of pure love to the detriment of ecclesiastical practice. These debates cast a pall over the role of mysticism in Roman Catholicism into the 20th century, though important mystics continued to be found. (B.J.McG.)

**Protestant Christianity.** The chief representatives of Protestant mysticism are the continental "Spirituals," among whom Sebastian Franck (*c.* 1499–*c.* 1542), Valentin Weigel (1533–88), and Jakob Böhme (1575–1624) are especially noteworthy. Among traditional Lutherans Johann Arndt (1555–1621) in his *Four Books on True Christianity* took up many of the themes of medieval mysticism in the context of Reformation theology and prepared the way for the spiritual revival known as Pietism, within which such mystics as Count von Zinzendorf flourished. In England the Anglican divines known as the Cambridge Platonists, the Quakers headed by George Fox (1624–91), and William Law (1686–1761) were important. In Holland a mystical group known as Collegiants, similar to the Quakers, broke away from the Remonstrant (Calvinist) Church. Other mystical bodies were the Schwenckfeldians, founded by Kaspar Schwenckfeld, and the Family of Love, founded in Holland by Hendrik Niclaes early in the 16th century before moving to England about 1550. The religion of the Ranters and other radical Puritans in 17th-century England had mystical aspects.

The cardinal feature of Protestant mysticism is the emphasis laid on the divine element in humanity variously known as the "spark" or "ground" of the soul, the "divine image" or "holy self," the "Inner Light," or the "Christ within." This was one of the essential elements of Rhineland mysticism and shows the connection between medieval and Reformation mysticism. For Böhme and the Spirituals, essential reality lies in the ideal world, which Böhme described as "the uncreated Heaven." Böhme took over the Gnostic belief that the physical world arose from a primeval fall, renewed with the Fall of Adam. His teaching was the main formative influence on the developed outlook of William Law and William Blake (1757–1827).

The mystical view of sin

For Protestant as well as for Roman Catholic mystics, sin is essentially the assertion of the self in its separation from God. The divine life is embodied in "the true holy self that lies within the other" (Böhme, *First Epistle*). When that self is manifested, there is a birth of God (or of Christ) in the soul. Protestant mystics rejected the Lutheran and Calvinist doctrine of the total corruption of human nature. William Law remarked: "the eternal Word of God lies hid in thee, as a spark of the divine nature" (*The Spirit of Prayer,* I.2.). "The eternal Word of God" is the inner Christ, incarnate whenever people rise into union with God. By the Spirituals Christ was viewed as the ideal humanity born in God from all eternity. This conception received its greatest emphasis with Kaspar Schwenckfeld, who, unlike Protestant mystics generally, taught that humans as created beings are totally corrupt; salvation means deliverance from the creaturely nature and union with the heavenly Christ.

Protestant mystics explicitly recognize that the divine Light or Spark is a universal principle. Hans Denck in the early 16th century spoke of the witness of the Spirit in "heathens and Jews." Sebastian Franck, like the Cambridge Platonists, found divine revelation in the work of the sages of Greece and Rome. George Fox appealed to the conscience of the American Indians as a proof of the universality of the Inner Light. William Law described non-Christian saints as "apostles of a Christ within." Protestant mystics stated plainly that, for the mystic, supreme authority lies of necessity not in the written word of Scripture but in the Word of God in the self. Fox said: "I saw, in that Light and Spirit that was before the Scriptures were given forth" (*Journal,* chapter 2). It was especially on this ground that the mystics came into conflict with the established church, whether Roman Catholic or Protestant.

The Ranters provide a good example of the conflict between mysticism and established religion. They held, with Fox and Hendrik Niclaes, that perfection is possible in this life. Puritan leaders under the Commonwealth denounced them for their "blasphemous and execrable opinions," and there was, no doubt, an antinomian tendency among them that rejected the principle of moral law. Some rejected the very notion of sin and believed in the universal restoration of all things in God.

### STAGES OF CHRISTIAN MYSTICISM

Christian mystics have described the stages of the return of the soul to God in a variety of ways. Following the Belgian Jesuit Joseph Maréchal, it can be suggested that Christian mysticism includes three broadly defined stages: (1) the gradual integration of the ego under the mastery of the idea of a personal God and according to a program of prayer and asceticism, (2) a transcendent revelation of God to the soul experienced as ecstatic contact or union, frequently with a suspension of the faculties, and (3) "a kind of readjustment of the soul's faculties" by which it regains contact with creatures "under the immediate and perceptible influence of God present and acting in the soul" (Maréchal, *Studies in the Psychology of the Mystics*). It is this final stage, which almost all of the greatest Christian mystics have insisted upon, that belies the usual claim that mysticism is a selfish flight from the world and an avoidance of moral responsibility.

**The dying to self.** The mystics agree on the necessity of dying to the false self dominated by forgetfulness of God. In order to attain the goal, it is necessary to follow the way of purgation: the soul must be purified of all those feelings, desires, and attitudes that separate it from God. This dying to the self implies the "dark night of the soul" in which God gradually and sometimes painfully purifies the soul to ready it for the divine manifestation.

Christ as the model of the purified soul

Christian mystics have always taken Christ, especially the crucified Christ, as the model for this process. According to the *Theologia Germanica,* "Christ's human nature was so utterly bereft of self, and apart from all creatures, as no man's ever was, and was nothing but a 'house and habitation of God' " (chapter 15). The following of Christ involves a dying to self, a giving up of oneself wholly to God, so that one may be possessed by the divine Love. Such detachment and purgation were frequently expressed in extreme terms that imply the renunciation of all human ties. Paradoxically, those who insist upon the most absolute detachment also emphasize that purifying the self is more a matter of internal attitude than of flight from the world and external penance. In the words of William Law: "The one true way of dying to self wants no cells, monasteries or pilgrimages. It is the way of patience, humility and resignation to God" (*The Spirit of Love,* Part 1).

The practice of meditation and contemplative prayer,

leading to ecstasy, is typical of Christian and other varieties of theistic mysticism. This usually involves a process of introversion in which all images and memories of outer things must be set aside so that the eye of inner vision may be opened and readied for the appearance of God. Introversion leads to ecstasy in which "the mind is ravished into the abyss of divine Light" (Richard of Saint-Victor, *The Four Grades of Violent Love*). Illumination may express itself in actual radiance. Symeon the New Theologian speaks of himself as a young man who saw "a brilliant divine Radiance" filling the room. In the path to union many of the Christian mystics experienced unusual and extraordinary psychic phenomena—visions, locutions, and other altered states of consciousness. The majority of mystics have insisted that such phenomena are secondary to the true essence of mysticism and can even be dangerous. "We must never rely on them or accept them," as John of the Cross said in *The Ascent of Mount Carmel,* 2.11.

**The union with God.** Christian mystics claim that the soul may be lifted into a union with God so close and so complete that it is in some way merged in the being of God and loses the sense of any separate existence. Jan van Ruysbroeck wrote that in the experience of union "we can nevermore find any distinction between ourselves and God" (*The Sparkling Stone,* chapter 10); and Eckhart speaks of the birth of the Son in the soul in which God "makes me his only-begotten Son without any difference" (German *Sermons,* 6). These strong expressions of a unity of indistinction have seemed dangerous to many, but Eckhart and Ruysbroeck insisted that, properly understood, they were quite orthodox. Bernard of Clairvaux, who insisted that in becoming one spirit with God the human "substance remains though under another form" (*On Loving God,* chapter 10), and John of the Cross, who wrote "the soul seems to be God rather than a soul, and is indeed God by participation" (*The Ascent of Mount Carmel* ii, 5:7), express the more traditional view of loving union.

**The readjustment.** The goal of the mystic is not simply a transient ecstasy; it is a permanent state of being in which the person's nature is transformed or deified. This state is frequently spoken of as a spiritual marriage that weds God and the soul. This unitive life has two main aspects. First, while the consciousness of self and the world remains, that consciousness is accompanied by a continuous sense of union with God, as Teresa of Ávila clearly shows in discussing the seventh mansion in *The Interior Castle.* Brother Lawrence wrote that while he was at work in his kitchen he possessed God "in as great tranquillity as if I were upon my knees at the Blessed Sacrament" (*The Practice of the Presence of God,* chapter 4). Second, the spiritual marriage is a theopathic state: the soul is felt to be in all things the organ or instrument of God. In the unitive life Mme Guyon says that the soul "no longer lives or works of herself, but God lives, acts and works in her." In this state the mystic is able to engage in manifold activities without losing the grace of union. In the words of Ignatius of Loyola, the mystic is "contemplative in action."

The spiritual marriage

## FORMS OF CHRISTIAN MYSTICISM

Christian mysticism has expressed itself in many forms during the last two millennia. Three broad types characterize much of Christian mysticism, though these should not be seen as mutually exclusive. Some mystics tend to emphasize one form over the others, while others make use of all three.

**Christ-mysticism.** The earliest form of Christian mysticism was the Christ-mysticism of Paul and John. Although Christian mysticism in its traditional expression has centred on aspiration for union with God, Christ-mysticism has always been present in the church. In the Eastern Church emphasis was placed on the divine Light that appeared to the disciples at the Transfiguration, and mystics sought to identify with this light of Christ in his divine glory. Symeon says of a certain mystic that "he possessed Christ wholly. . . . He was, in fact entirely Christ." In the Catholic West, with reference to the founding figure of Augustine, it is evident that it is in and through the one Christ, the union of Head and body that is the church, that humans come to experience God. For Augustine the mystical life is Christ "transforming us into himself" (*Homily on Psalm,* 32.2.2). In the medieval period some of the most profound expressions of Christ-mysticism are found in the women mystics, such as Catherine of Siena and Julian of Norwich. Luis de León spoke of the theopathic life in terms of Christ-mysticism: "The very Spirit of Christ comes and is united with the soul—nay, is infused throughout its being, as though he were soul of its soul indeed."

With Protestants the attempt to return to primitive Christianity has led to strong affirmations of Christ-mysticism. The early Quaker George Keith wrote that Christ is born spiritually in humanity when "his life and spirit are united unto the soul." The chief representative of Christ-mysticism among the early Protestants was Kaspar Schwenckfeld. For him Christ was from all eternity the God-man, and as such he possessed a body of spiritual flesh in which he lived on Earth and which he now possesses in heaven. In his exalted life Christ unites himself inwardly with human souls and imparts to them his own divinity.

**Trinitarian mysticism.** Pure God-mysticism is rare in Christianity, though not unknown, as Catherine of Genoa shows. Christ as God incarnate is the Word, the second Person of the Trinity, and Christian mysticism has, from an early era, exhibited a strong Trinitarian dimension, though this has been understood in different ways. What ties the diverse forms of Trinitarian mysticism together is the insistence that through Christ the Christian comes to partake of the inner life of the Trinity. The mysticism of Origen, for example, emphasizes the marriage of the Word and the soul within the union of Christ and the church but holds out the promise that through this action souls will be made capable of receiving the Father (*First Principles,* 3.6.9). The mystical thought of Augustine and of such medieval followers as Richard of Saint-Victor, William of Saint-Thierry, and Bonaventure is deeply Trinitarian. Meister Eckhart taught that the soul's indistinction from God meant that it was to be identified with the inner life of the Trinity—that is, with the Father giving birth to the Son, the Son being born, and the Holy Spirit proceeding from both. A similar teaching is found in Ruysbroeck. John of the Cross wrote of mystical union that "it would not be a true and total transformation if the soul were not transformed into the three Persons of the Most Holy Trinity" (*Spiritual Canticle,* stanza 39.3). Such strong Trinitarian emphasis is rarer, but not absent from Protestant mysticism.

**Negative mysticism: God and the Godhead.** The most daring forms of Christian mysticism have emphasized the absolute unknowability of God. They suggest that true contact with the transcendent involves going beyond all that we speak of as God—even the Trinity—to an inner "God beyond God," a divine Darkness or Desert in which all distinction is lost. This form of "mystical atheism" has seemed suspicious to established religion; its adherents have usually tried to calm the suspicions of the orthodox by an insistence on the necessity, though incompleteness, of the affirmative ways to God. The main exponent of this teaching in the early centuries was the Pseudo-Dionysius, who distinguished "the super-essential Godhead" from all positive terms ascribed to God, even the Trinity (*The Divine Names,* chapter 13). In the West this tradition is first found in Erigena and is especially evident in the Rhineland school. According to Eckhart, even being and goodness are "garments" or "veils" under which God is hidden. In inviting his hearers to "break through" to the hidden Godhead, he daringly exclaimed, "let us pray to God that we may be free of 'God,' and that we may apprehend and rejoice in that everlasting truth in which the highest angel and the fly and the soul are equal" (German *Sermons,* 52). The notion of the hidden Godhead was renewed in the teaching of Jakob Böhme, who spoke of it as the *Ungrund*—"the great Mystery," "the Abyss," "the eternal Stillness." He stressed the fact of divine becoming (in a nontemporal sense): God is eternally the dark mystery of which nothing can be said but ever puts on the nature of light, love, and goodness wherein the divine is revealed to human beings.

The unknowability of God

### SIGNIFICANCE OF CHRISTIAN MYSTICISM

The study of Christian mysticism presents both the unity of mysticism as an aspect of religion and the diversity of expression that it has received in the history of Christian faith. The mystic claims contact with an order of reality transcending the world of the senses and the ordinary forms of discursive intellectual knowing. Christian mystics affirm that this contact is with God the Trinity and can take place only through the mediation of Christ and the church, whether explicitly or implicitly at work. The claim is all the more significant in that Eastern Orthodox Christians, Roman Catholics, and Protestants are here in agreement.

Without in any way affirming that all mysticism is everywhere one and the same, it can be said that the Christian mystics take their stand with the mystics of other traditions in pointing to "the Beyond that is within." In an age when the claims of established religion are so widely questioned, the witness of the mystics is of particular appeal; but it should be remembered that most mystics have not been rebels against their respective religious confessions. Another great question that confronts the present age is the relation of Christianity to other world religions. If Christianity is to embark upon truly cooperative relations with other religions, it must be deeply imbued with the insight and experience of the mystics. Even if it is to attempt to plumb the depths of its own history, it cannot neglect its mystical dimensions. (S.Sp./B.J.McG.)

## Christian myth and legend

Myths and legends number among the most creative and abundant contributions of Christianity to the history of human culture. They inspire artists, dramatists, clerics, and others to contemplate the wondrous effects of Christian salvation on the cosmos and its inhabitants. They conjoin diverse cultural horizons, taking those worldviews bounded by Christian revelation and fusing them creatively with the religious histories that exist prior to and alongside the orthodox Christian world. Even for the less pious and the nonbelievers, the distinctive visions of reality presented in Christian legend or myth and the symbolic actions based upon them have helped to form the fundaments of Western civilization. Pilgrimage to the shrines of legendary saints, to mention but one example, touches economic and political life, military history, visual and musical arts, popular devotion, and the exchange of scientific information. Moreover, the content of the legends and myths themselves has contributed directly to theories about religion, society, politics, art, astronomy, economics, music, and history.

### CHARACTERISTICS OF CHRISTIAN MYTH AND LEGEND

An appreciation of the positive role of myth and legend in culture has been long in coming. Christian theology, taking its lead from Greek philosophy, at first denigrated the value of myth. In constructing the Christian canon and in choosing authoritative interpretations of it, the early church suppressed or excluded myth and legend in favour of the genres of philosophy, history, and law. The opinion expressed in the First Letter of Paul to Timothy only echoes the prevalent Hellenistic view of myth: "Have nothing to do with godless and silly myths" (1 Timothy 4:7). In spite of that, a number of important mythical themes remain central to the New Testament—*e.g.*, Christ as the second Adam (Romans 5:12–14), the heavenly spheres (2 Corinthians 12:2–4), and the celestial battle between angels and demons. Still visible, but barely so, in the writings of Paul (Galatians 3:28) is the early Christian theme of the androgyny of Christ and of his spiritually accomplished disciples.

Early exclusion of myth and legend

The Apologists from the 2nd to the 5th centuries used legend and myth. Clement of Alexandria employed them as allegories to make Christian concepts intelligible to Greek converts. But Clement (*e.g.*, in his *Protreptikos* ["Exhortation"]) and other Church Fathers roundly condemned the belief that Greek myths might be autonomous sources of truth. In spite of its ambiguous use of mythic symbols and themes, the history of Christian doctrine testifies to the systematic excision of legendary and mythical elements from Christian orthodoxy. Even folk practices, based on legend, were policed and suppressed. In 692, for example, the Quinisext Council (also known as the Trullan Synod), a precedent-setting episcopal council convoked by the Byzantine emperor Justinian II, prohibited baking bread in the form of the Virgin Mary's placenta, as was the custom on the afterbirth day (that is, the day after Christ's birth). This ambiguous, but ultimately negative, evaluation of Christian myth and legend lingers to this day.

A second cause for the delay in evaluating the positive contributions of myth and legend to religious life is the theories of religion that have flourished since the time of the European Enlightenment. These theories treated myths as infantile projections of the prerational childhood of the human race (projections surpassed by the mature rationalism of the Enlightenment). More intimate knowledge of mythic traditions in Africa, India, Oceania, and the Americas, however, has disclosed the important role myth plays in culture and highlighted the coherence and sophisticated order of myth.

Role of myth in society

Myths narrate the sacred events that unfolded in the first time, the epoch of creative beginnings. In that primordial period supernatural beings brought reality—in part or in whole—into existence. In that sense, myth relates only those things that have really occurred—that is, those realities that have revealed themselves completely. These realities become the foundation of the world, society, and human destiny. Myths manifest the acts and beings that are sacred, that are completely other than the world encountered in day-to-day experience. Myths are always paradoxical because realities that are other than those of this world have nonetheless established it. The intervention of sacred and supernatural beings accounts for the conditions of the world and humanity today. Myth describes the acts and beings whose appearance shaped material existence in all its concrete specificity.

Legends are episodic continuations of mythic narratives, for they describe the effects of primordial events on history—a history revealed through the imagination and one as fabulous as the primordial mysteries that brought that history into being. Legends must describe history in fantastic terms in order to clarify the significance of the powers that underlie it. The repetitiveness and redundancy of legends emphasize the fact that many different legends spring from the same mythic sources—that is, from the same primordial events and creative powers. But variants of legend are reminders that myths and their outcomes are historically conditioned and questioned. Christian legend contends with the question of what the Christian mystery means here and now, in these particular, everyday circumstances. Because of their local frame of reference, legends vary incessantly, and widely different accounts emerge from diverse locales and periods. Favourite legendary themes are the struggles and miraculous adventures of heroes in the faith. Such accounts edify the faith and bolster the courage of the listener.

There is no complete account of Christian legend and myth, nor is there a full outline of the mythic world engendered by the economy of salvation set in motion with the life of Christ and his disciples. Above all, the theologies of rural populations and oral traditions have been slighted in the study of Christian thought. A historical interpretation of the full mythical and legendary expression of Christianity would probably reveal a surprising adherence to tradition even while it uncovered startling reinterpretations of the Christian message over time. Christian myths and legends exuberantly express the truths of Christian existence, viewed as a religious situation in the social and physical world. The mystery of salvation unfolds when the eternal God dramatically enters the created universe in the form of an incarnate, mortal creature. "Like us in all things but sin," God's presence among human beings is mysterious, and the meaning of this mystery risks remaining hidden and undeciphered. Legends and myths spell out the effects of these salvific secrets not only for human individuals but also for all realms of reality—animal, vegetal, astral, material, corporeal, social, and intellectual. By quickening the listeners' religious awareness of the salvation unfold-

ing around them, the symbols of legend and myth often aimed to further the redemptive effects of the mysteries and wonders they describe. (L.E.S.)

## HISTORY OF CHRISTIAN MYTH AND LEGEND

**The early church.** Early Christianity appropriated mythological motifs and genres from the Greek and Middle Eastern cultures that dominated the Hellenistic Age (*c.* 300 BC–*c.* AD 300). Among them was the miraculous birth of a deity; the virgin birth of a god or goddess was a theme common in the mythology of the Hellenistic world. Aphrodite (Venus), the goddess of sexual love, for example, emerged from sea foam. Athena sprang, in full battle array, from the head of Zeus, her father. The legend of the virgin birth of Alexander the Great (4th century BC) from his mother, Olympias, whose reputation was not that of a virgin, demonstrated Alexander's divinity. Mithra, the Iranian god of light and of sacred contracts, is described as a divine child of radiant heavenly beams. Mithra was born from the rock of a cave, the birth witnessed by shepherds on a day (December 25) that was later claimed by Christians as the Nativity of Christ.

Influence of Judaism

Hellenistic Judaism had already reinterpreted many Gentile motifs and set them within a biblical context. From Greek and Jewish sources Christians adopted and adapted some favourite mythical themes: the creation of the world, the end of the paradisal condition and the fall of humankind, the assumption of human form by a god, the saved saviour, the cataclysm at the end of time, and the final judgment. Christians reframed these motifs within their new images of history and their doctrines concerning the nature of God, sin, and redemption. As it spread beyond Palestine and the Hellenistic world over the course of time, Christianity continued to develop mythical themes important to the religious consciousness of converted peoples.

*The ages of the world.* One fascinating mythical theme in the New Testament is that time consists of a series of ages. Each age of the world (or kingdom) is dominated by a powerful force or figure. This motif exists throughout the globe with a range of specific cultural meanings. In the 8th century BC in Greece, the poet Hesiod described the ages of the world as four in number and symbolized by gold, silver, bronze, and iron, each age successively declining in morality. In India the four *yugas* (Sanskrit: "world ages"), symbolized by the four throws of a dice game, also are viewed as descending—though in repetitive cycles—from perfection to moral chaos. Other original schematizations of this theme can be found in the mythologies of Chinese, Polynesian, and American Indian cultures.

By the time the New Testament was written, Jewish apocalyptic writings (symbolic or cryptographic literature portraying God's dramatic intervention in history and catastrophic dramas at the end of a cosmic epoch) had already produced theories of history that reworked Indo-Iranian notions about the ages of the world. Iranian concepts most influenced Christian views of time, history, and ultimate human destiny. The prophet Zoroaster (*c.* 7th century BC) and his followers in Iran taught a doctrine of the four ages of the world in which each age was a different phase in the struggle between two kinds of powers—light and darkness, goodness and evil, spirit and matter, infinity and finitude, health and sickness, time and eternity. The forces of good and evil battled for the allegiance and the souls of human beings. In the last days a promised saviour (Saoshyant) would pronounce final judgment and announce the coming of a new world without end in which truth, immortality, and righteousness would have everlasting reign.

Stages of the Christian apocalypse

Drawing on Jewish apocalyptic literature (exemplified in the Book of Daniel), early Christian apocalypse (exemplified in the Book of Revelation) elaborated the theme of the ages of the world as a series of historical periods in which good struggles against evil: (1) from the creation of the world and of humanity to the Fall into sin and out of Eden; (2) from the Fall to the first coming of Christ; (3) from the first to the second advent of Christ, which includes the 1,000-year reign of Christ and his saints and the Last Judgment; and (4) the creation of a new heaven and a new earth in which those who have chosen the good (*i.e.,* Christ) will live in eternity. Within this framework of the mythical history of the ages of the world, Christian apocalyptic re-envisions a number of themes important to Jewish apocalypticism: the Son of man and the great tribulation prior to the judgment of the world; the battle between Christ and the Antichrist, a false messiah or "great liar" who denies that Jesus is the Christ and who pitches the world into moral confusion and physical chaos; and the ultimate triumph over Satan, who appears as a dragon but who no longer deceives the nations of the world.

The theme of the several ages of the world has a long and fruitful life in Christian thought and undergirds many Western concepts of progress toward a better state of existence or of decline toward extinction. Montanus, a heretical Christian prophet of the early 2nd century, claimed that history progressed from an age of the Father to an age of the Son to an age of the Holy Spirit, of whom Montanus was the manifestation. One could fruitfully explore the degree to which such apocalyptic myths underlie not only the religious theories of a multistage history, as propagated by Martin Luther, the early Jesuits, Christopher Columbus (in his *Book of Prophecies*), and Giambattista Vico, but also the more secular philosophies of history developed by Gotthold Ephraim Lessing, the comte de Saint-Simon, Auguste Comte, Johann Gottlieb Fichte, G.W.F. Hegel, Friedrich Schelling, and Karl Marx.

*Messianic secrets and the mysteries of salvation.* New Testament references to the "mysteries of the kingdom of heaven" (for example, Matthew 13:11; Mark 4:11; Luke 8:10) generated myth and legend. The New Testament emphasis on secrecy and on the mysteries of salvation became fertile ground for the exfoliations of myth and legend. Things hidden from the beginning of the world now blossomed in the signs of the new messianic age. These truths, now come to light, should be proclaimed to the whole world. Through myth and legend Christians transmitted and explored, with the full force of the imagination, the wonders revealed in Christ and the secrets of his salvation.

Esoteric traditions, especially those based on apocalypses and apocrypha (such as the *Apocalypse of Peter, Gospel of Thomas, Secret Gospel of Mark,* and *Gospel of Philip*) preserve some legends and myths descending from the early Christian centres of Edessa, Alexandria, and Asia Minor.

The childhood of Jesus

The *First Gospel of the Infancy of Jesus* (known also as the Arabic Infancy Gospel) recounts that, one day, Jesus and his playmates were playing on a rooftop and one fell down and died. The other playmates ran away, leaving Jesus accused of pushing the dead boy. Jesus, however, went to the dead boy and asked, "Zeinunus, Zeinunus, who threw you down from the housetop?" The dead boy answered that Jesus had not done it and named another (*I Infancy* 19:4–11). This and other such narratives describe the "hidden life" of Jesus in the 30 years before his public ministry began. The *Acts of Paul and Thecla* narrates the story of a friend of Paul who was thrown to the lions—one of which defended her in a manner similar to that of the lion in the story of Androcles, a well-known legend. Other exemplary legends appear in the *Acts of the Martyrs* and other histories. After Christian theologians defined orthodoxies in terms of Greek philosophy or Roman juridical code, these mythic themes appeared clumsy or tasteless and, in retrospect, heterodox or even heretical.

Groups of Gnostics and heretics who based their ideas on alternative mythologies of the economy of Christian salvation furnished exotic Christian myths, legends, and practices. In the 2nd and 3rd centuries these groups often subscribed to theories of dualism: the world of matter created by an evil god (of the Book of Genesis) and the realm of the spirit created by a good god (revealed in the New Testament) were irreconcilably pitted against one another. The many Gnostic sects—among them the Valentinians, Basilidians, Ophites, and Simonians—developed a variety of myths. Valentinus lived in Rome and Alexandria in the mid-2nd century. Valentinian myths describe how the pleroma (spiritual realm) that existed in the beginning was disrupted by a Fall. The Creator God of Genesis, aborted from the primordial world, became a Demiurge and cre-

ated the material universe. He deliberately created two kinds of human being and animated them with his breath: the hylics and the psychics. Unknown to the Demiurge, however, certain remnants of pleromic wisdom contained in his breath lodged as spiritual particles in matter and produced a third group of beings called pneumatics. The God of Genesis now tries to prevent Gnostics from discovering their past origins, present powers, and future destinies. Gnostics (the pneumatics) contain within themselves divine sparks expelled from the pleroma. Christ was sent from the pleroma to teach Gnostics the saving knowledge (gnosis) of their true identities and was crucified when the Demiurge of Genesis discovered that Christ (the male partner of the feminine Holy Spirit) was in Jesus. After Christ returned to the pleroma, the Holy Spirit descended.

Myths of the Ophites

The Ophites (from the Greek word *ophis,* "serpent") reinterpreted the mythological theme of the Fall of Man in Genesis. According to the Ophite view, the serpent of the Garden of Eden wanted Adam and Eve, the first man and woman, to eat from the tree of knowledge (*gnosis*) so that they would know their true identities and "be like God" (Genesis 3:5). The serpent, thus, is interpreted as a messenger of the spiritual god, and the one who wanted to prevent Adam and Eve from eating the fruit of the tree of knowledge is viewed as the Demiurge. In their rejection of the God of the Old Testament, who gave the Ten Commandments, the Ophites flaunted their sexual freedom from the law and conventionality by extreme sexual license, a trait common to other Gnostic groups as well.

The Phibionites in Alexandria were a Gnostic sect described by Epiphanius. They gathered at banquets that became ecstatic orgies. Married couples changed partners for dramatic sexual performances. Sperm and menstrual blood were gathered and offered as a gift to God before being consumed as the Body and Blood of Christ. By such erotic communions they sought to regather the elements of the world-soul (*psyche*) from the material forms into which it had been dispersed through a cosmic tragedy at the beginning of time. The regathering amounted to salvation, for all things would be gathered up into the one glorious body of Christ. (L.F./L.E.S.)

*The Magi and the Child of Wondrous Light.* The legend of the Magi-Kings was embellished in apocryphal books and Christian folklore. The *Protogospel of James* and the *Chronicle of Zuqnin* describe the birth of the Saviour. Like the god Mithra, the divine child is consubstantial with celestial light and was born in a mountain cave on December 25. Such imagery of the Nativity of Christ and the symbolism of the royal visitors may originally have descended from Iranian accounts of the birth of the cosmic saviour, for the accounts seem to owe a great deal to Iranian theologies of light. But the themes have been recast in Christian terms. The *Opus imperfectum in Matthaeum* relates that 12 Magi-Kings lived near the Mountain of Victories, which they climbed every year in the hope of finding the messiah in a cave on the mountaintop. Each year they entered the cave and prayed for three days, waiting for the promised star to appear. Adam had revealed this location and the secret promises to his son Seth. Seth transmitted the mysteries to his sons, who passed the information from generation to generation. Eventually the Magi, sons of kings, entered the cave to find a star of unspeakable brightness, glowing more than many suns together. The star and its bright light led to, or became, the Holy Child, the son of the Light, who redeems the world.

Giruadon/Art Resource, New York City



Angel appearing to the Magi, warning them not to return to King Herod; relief by Gislebertus; cathedral of Saint-Lazare, Autun, Fr., 12th century.

*Relics and saints.* The cult of saints gained momentum from the 4th to the 6th century. The bones of martyrs gave stirring evidence of God's power at work in the world, producing miracles and spectacles of the effectiveness of faith. The martyrs had imitated Christ even unto death, and the remains of their holy bodies served as contacts between earth and heaven. On the model of Christ's Incarnation, the bones of martyred saints embodied God's salvific power and thus became the centre of active cults. Relics were installed in special churches called martyria or in basilicas. The tombs of martyrs, on the margins of cities and towns, attracted pilgrims and processions. Legends described the prodigious virtues of martyrs and saints, as well as the dreams or visions that revealed the resting places of still more powerful relics. Each discovery (*inventio*) promised new and effective signs of divine redemption. Returning from distant places, especially Rome, pilgrims brought relics to their home churches. Thus, during the 8th century, bones and other relics were moved from southern Europe to the north and west. During the Middle Ages especially, deities and cultural heroes became elements of Christian hagiography.

The True Cross and the World Tree

Of all discovered relics the most impressive was the True Cross, found in September 335 (or in 326, according to other accounts). Prompted by a dream, Helena, mother of the emperor Constantine, located the place where the Cross lay buried and had the wood unearthed. The power of the Cross, the history of the wood, and the story of its discovery became legendary. In Christian myth this relic of Christ's death dated back to the mortal origins of humanity. Innumerable cures attested to the authenticity of the Cross.

Through the symbolism of the Cross early Christian imagery perpetuated, and at the same time transformed, the myths of the World Tree. The sacred drama of Christ's birth, death, and Resurrection participates in the rejuvenating rhythms of the fecund cosmos. Early Christians identified the Cross of Christ as the World Tree, which stood at the centre of cosmic space and stretched from earth to heaven. The Cross was fashioned of wood from the Tree of Good and Evil, which grew in the Garden of Eden. Below the tree lies Adam's buried skull, baptized in Christ's blood. The bloodied Cross-Tree gives forth the oil, wheat, grapes, and herbs used to prepare the materials administered in the sacraments that revitalize a fallen world. The Italian Renaissance painter Piero della Francesca later depicted the myth of the True Cross in his frescoes in Arezzo, Italy. They portray the death of Adam, fallen at the foot of the Tree that provides wood for the crucifix on which Jesus is slain. But the wood of the Cross becomes the instrument of salvation and the holiest matter in Christendom. Fabulous accounts and fantastic historical episodes surround the Cross.

Another 4th-century event, the discovery of Christ's tomb, the Holy Sepulchre in Jerusalem, also became a highlight of Christian legend. Like the body of the Saviour, the tomb is a "holy of holies." Its discovery was tantamount to the Resurrection, for its reemergence into the light of day was seen as a restoration of life where before only darkness reigned. The Cross and the tomb were woven together in legend. The desire to regain possession of the True Cross and the Holy Sepulchre eventually fueled the territorial expansion of Christian empires and spurred Christian knights to crusade. (L.E.S.)

**The Middle Ages.** As Christianity expanded from the cultural milieu of the Mediterranean area to the north and east, the various converted tribes and peoples did not, understandably, forget their own religious heritages. Just as attributes of the Roman god of war, Mars, had been transferred to Michael, the archangel who is the leader of the heavenly hosts, in the early centuries of the church,

so also the attributes of the gods of the Germanic, Baltic, Slavic, and other peoples were transferred to angels and saints during the Middle Ages. For example, St. George, who rescued a maiden after slaying a dragon, became the patron saint of England and one of the most popular saints among the Balts (among whom St. George replaced the god Kalvis, the heavenly smith and dragon slayer).

**Prester John and the Holy Grail**

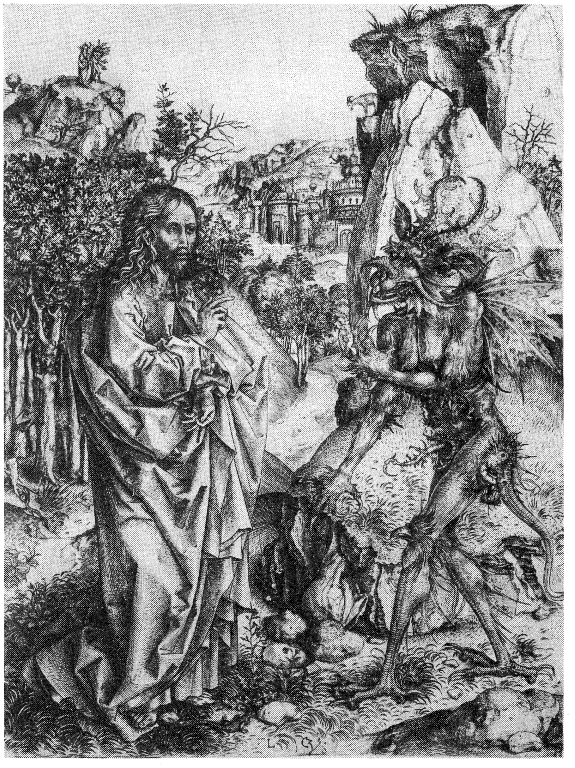
Nor were saints the only legendary figures important to Christendom. Prester (Presbyter) John, a fabled Christian priest-king of the Orient, became so believable a figure in the Middle Ages that Pope Alexander III dispatched a letter to him in 1177. Similarly, the legend of the wandering Jew, who had taunted Jesus on his way to be crucified, was popular in the 13th century and again, from the 17th century on, in the stories about the wanderer Ahasuerus.

For the Christian medieval world the Holy Grail (the chalice used by Jesus at the Last Supper) symbolized the truth and knowledge needed to achieve the experience of salvation. Led in search of the Grail by divine grace, the naive hero Perceval inquired directly about the Grail, a question other knights had failed to ask. His simplistic question, put to the ailing Fisher King, revitalized not only the royal body but the entire drooping cosmos. The human condition is rejuvenated by the graceful quest for the truth of salvation. Perceval was superseded by Galahad as the winner of the Holy Grail in later variations, Galahad being viewed as a descendant of Joseph of Arimathea (the member of the Jerusalem council in whose tomb the body of Jesus was laid), who was believed to have gone to Glastonbury, Eng., with the Holy Grail. (L.F./L.E.S.)

Probably under the influence of Bogomil and Cathar heretical tendencies toward dualism, apocryphal books of Christian legends (such as *The Wood of the Cross, Gospel of Nicodemus, How Christ Became a Priest, Adam and Eve,* and *Interrogatio Iohannis*) circulated in both eastern and western Europe. They usually stressed the role of Satan as co-creator of the world or as a being whose fall is responsible for the evil world that exists. The devil plays a major role in legend, and his activity usually exhausts the creative energies of the good God, who falls into passivity.

A number of Christian myths, legends, and works of art were aimed at awakening religious capacities, turning the viewer or listener against repulsive forms of evil, and hastening the effects of the salvation achieved in Christ. Nowhere is this better illustrated than in the bestiaries, fables, and cosmic dramas sculpted into Romanesque cathedrals. Christ, the glorious King, and his saintly cohorts confront armies of monsters and demons. Together the two sides show forth the full spectrum of the imaginary world of Christian legend and myth of the day.

National Gallery of Art, Washington, D.C., Rosenwald Collection, B-11149



Christ being tempted by a devil, "Temptation of Christ," engraving by Master LCZ (Lorenz Katzheimer), *c.* 1492.

Christian legends and myths were also woven into long-lived literary creations: the late medieval chansons de geste yielded to the epic tales, lyric poetry, and songs that conducted audiences into an enchanted symbolic world that paralleled their mundane one. Such are the enigmatic poems of the 12th-century Court of Love and the literature patronized by Eleanor of Aquitaine and her daughter, Marie, countess of Champagne. Similarly the troubadours of 12th-century Provence creatively refashioned, in Christian terms, the inspirations they received from the Arabic poetry of Spain and the influences of Celtic, Gnostic, and Oriental themes in circulation at the time. These tendencies toward the fantastic in Christian expression reached their literary peak in the works of Dante (1265–1321), whose *Divine Comedy* depicts the terrifying and attractive visions of Paradise, Purgatory, and Hell in such a way as to quicken the ultimate powers of the imagination and thereby draw the reader toward the effective images of the mystery of their own salvation.

**King Arthur and the Round Table**

In the place of Charlemagne, a favourite hero of the old chansons de geste, the legendary cycles of the 12th century spawned a new generation of romantic heroes—King Arthur and the knights of his Round Table. Marie, countess of Champagne, sponsored Chrétien de Troyes, the poet who composed five long romances that became the mythic foundation for chivalry. These cycles interweave Christian, Muslim, and Gnostic elements into a singular cosmic vision. Suffering ordeals during their adventures, the knights of the Arthurian cycle (Arthur, the Fisher King, Perceval, and Lancelot) journey through the Wasteland on their heroic quests for the Holy Grail and for the cure that will revitalize king and cosmos. Wolfram von Eschenbach offers the most coherent mythology of the Grail in his *Parzival,* a refinement of Christian legends that draws on the worlds visited by the crusaders and by Italian merchants—Syria, Persia, India, and China. At the conclusion of many of these cycles, the Holy Grail, often in the image of the chalice of salvation in Christ, is transported to a fabulous mythical location in the Orient.

**Joachim of Fiore's universal history**

The 12th century also witnessed the rise of a new mythology of Christian history. Joachim of Fiore (1130/35–1201/02) was an abbot of the Calabrian monastery of Fiore and was well-known in the Christian world of his day. On the vigil of Easter and on Pentecost Sunday, God infused him with special knowledge, which enabled him to decode history as a series of divine signs. According to Joachim, universal history has three stages, each age (*status*) corresponding to a person of the Holy Trinity. The first age, presided over by God the Father, was ruled by married men and propelled by their labour. Jesus Christ presided over the age of the New Testament, an epoch ruled by the clergy and driven forward by the power of science and discipline. The two testamental periods featured the two kinds of people chosen in each, the Jews and the Gentiles. Joachim fascinated the faithful of his day with a prediction that the second age, the age of the New Testament presided over by Jesus Christ, would end in 1260. Then would dawn a new epoch, the third age, presided over by the Holy Spirit, guided by monks and fueled by their contemplation. It was to be an epoch of total love, joy, and freedom. But three and one-half years of cataclysm ruled by the Antichrist would precede entrance to this bliss.

Joachim promised that God's mysterious saving power would burst fully into history in the immediate future and would change forever the fundamental structures of the cosmos as well as the social and ecclesiastical world. Joachim's new vision of history generated critiques of the 13th-century church and society. His doctrine of the Trinity was condemned at the fourth Lateran Council in 1215. In 1255 Pope Alexander IV suppressed a collection of his written works, and in 1263 the regional Council of Arles condemned many of Joachim's most stirring ideas.

His notions of an impending third epoch, in which history would come to complete fulfillment, lived on.

**Renaissance magic and science.** Christian legend and myth also found fertile ground in the practices of alchemy. Through the perfection of metals the alchemists sought their own perfection and, indeed, the salvation of all matter. Through the mysterious and great work (magnum opus) of alchemy the alchemist dissolved, then fused, his own physical matter and spirit with the prime matter of the universe. These initiatory experiences of reduction into prime matter made possible the re-creation of individual and cosmos as a single, pure element. Even the philosopher's stone or elixir was reinterpreted so that Christ appeared as the perfect matter produced by the alchemical process—that is, Christ was the stone of all wisdom and knowledge. In the alchemist's spiritual forge, the Stone reemerged from the Matrix, the crucible containing the so-called Bath of Mary, whose amniotic fluids dissolved all impurities. This dissolution prepared one for rebirth as a perfect being. All matter was redeemed by immersion in the fluids of the womb where Jesus took flesh. Mystical union with Christ's death and physical regression to that same uterus where God became matter empowered the Christian alchemist to effect a new fusion of redeemed realities, freed of all impure dross. Scientists secretly continued the alchemical tradition. Among them numbered the foremost pioneers of modern physics and chemistry: Robert Fludd, Robert Boyle, and Sir Isaac Newton.

Legends of alchemy

Legends also found their place in the growing science of astronomy. In the Middle Ages it was learned that conjunctions of planets occur every 20 years on a minor scale and every 960 years on a major scale. This theory, described in the *Liber magnarum coniunctionum,* was advocated by Albumazar (787–886), a disciple of al-Kindī (?–*c.* 870), a Muslim philosopher who assimilated Greek philosophy to Islām. Roger Bacon used this theory to work out the chronology of great personalities in history and to map the chronological relationship of true prophets (Alexander the Great, Jesus Christ, Mani, and Muḥammad), one for every 320 years. Based on observations of a supernova in 1604, Johannes Kepler calculated the "true date" of the birth of Jesus. These calculations revitalized an interest in the legendary Magi, who had followed the great star. Kepler believed that the conjunctions were unnatural events brought about by the miraculous acts of God, who had decided to lodge the birth of his son between the significant zodiacal signs of the Fish (Pisces) and the Ram (Aries).

Rosicrucian announcements of the imminent coming of a new world also propagated the theory that great celestial conjunctions appeared at the births of prophets and saviours. The scientific achievements of Kepler became a foundation for the new secret order reputedly founded by Christian Rosenkreuz, for it confirmed their hopes. The editors of Rosicrucian publications dated the death of their founder to 1484 and fixed the time of the discovery of his tomb as 1604 in order to coordinate the events with the last two great conjunctions of stars.

**Christian practice in the modern world.** The 20th century continues to generate important Christian myths and legend-based practices, including pilgrimages made on Marian feast days to holy wells and fairy rings outside the Irish town of Sneem and devotions at the tomb of Christ in Japan, where, according to local legend, Christ ended the long life of missionary travels he began after his mock death in Jerusalem. These acts and the legendary explanations that accompany them detail the impact of Christian salvation on present-day reality. In all the cultures where Christianity has been propagated, myth and legend express the fulfillment of the religious desires and hopes that constituted the religious traditions before contact with Christian revelation. The following examples suggest their variety and vitality.

Healing of the sick

The healing of sickness is, as it was in the time of the New Testament, a sign of the coming of the Kingdom of Christ in its fullness. In Africa, for example, many so-called Independent Churches creatively reinterpret disease and rites of cure along Christian lines. In Douala, Cameroon, during the 1980s, two healing prophets named Mallah and Marie-Lumière divided their disciples, whom they called the "sick ones of the Father," into groups named for the important categories of illness described in the Gospels: the Blind, the Halt, the Lame, the Deaf, the Epileptic, the Dumb, and the Paralyzed. The disciples evidenced none of these physical symptoms, but they were asked to identify deep within themselves with the affliction described in the Gospel, so that salvation might touch them in their inner being. By becoming sick, they could be healed and thus join the elect. In lengthy sermons the healing prophets reimagined traditional African religious imagery and refashioned it in the light of Christian belief. The experience of their peculiar mystical disorders afforded a basis for social regrouping and for rethinking the past and present.

The Christian expression of sacred music and trance is often grounded in legend or myth. In Brazil, for example, Macumba, Candomblé, and other Afro-Brazilian cults have roots sunk deep into the religions of African slaves transplanted to the New World. Afro-Brazilian rites often centre on possession by a supernatural being, called an *orixá.* The innumerable *orixás* are ranked in hierarchies modeled on the pantheons of the Yoruba people of West Africa, among others. In Brazil (and in much of Afro-American religious life of the Americas), each *orixá* is identified with a specific Christian saint. In the Umbanda cult of Brazil, altars hold small plaster images of the Christian saints associated with the *orixás.* Each one of the saints presides over a domain of human activity or over a disease, social group, geographic area, or craft. For example, Omolú, the god of smallpox, is identified with St. Lazarus, whose body, in Christian legend, is pocked with sores and who heals diseases of the skin. Oxossi, the Yoruba god of hunting, is associated with the bellicose St. George or St. Michael, the slayers of dragons and other demonic monsters. Yansan, who ate the "magic" of her husband and now spits up lightning, is associated with St. Barbara, whose father was struck by lightning when he tried to force her to give up her Christian faith. In the worship site each *orixá* has its own stone, which is peculiarly shaped, coloured, or textured; arranged in a distinctive position on the altar; and identified as the Cross of Christ. A single saint may be identified with several *orixás* or vice-versa. Regions vary the saintly identifications and some designations shift over time. Each *orixá* has its own musical rhythms and sounds. When called by drums, dance, and music, the supernatural being may take over the possessed medium, reveal valued information, and carry out effective symbolic acts on behalf of the community.

The saints of Umbanda

European communities continue to be fascinated with the rigorous asceticism of St. Anthony of Egypt, who repulsed wild beasts, reptiles, and other assaults and remained steadfast in the faith. He is considered the patron of domestic animals, and in many parts of Italy, the drama of the feast of St. Anthony, historically associated with the winter solstice, rivals any other feast day of the Christian calendar. To celebrate that festival in Fara Filiorum Petri, a town in the Abruzzi region of Italy, the townspeople ignite enormous bonfires on the night of January 16. Each of the 12 outlying hamlets brings into the main town's square a bundle (*farchia*) of long poles. Set on end, the bundles are lashed together to form a single tall mass, an act that commemorates the historical union of the mountain settlements as one bonded community. Then the bundles of *farchie,* 15 or more feet high, are set ablaze. The fire cleanses the community and holds at bay the evil forces of sickness and death. As the fire dies down, young men jump through the purifying flames. Spectators carry remnants of the blessed fire back to their homes, spreading the ashes in their stalls and on their fields.

The feast of St. Anthony

The birth of Christ is still a focus for traditions of legends and myths that maintain their autonomous existence outside of ecclesiastical institutions. In rural Romania, for instance, on Christmas Eve groups of young carolers (*colindatori*) proceed from house to house in the village, singing and collecting gifts of food. Often these carolers impersonate the saints, especially John, Peter, George, and Nicholas. The words of their songs (*colinde*) describe legendary heroes who carry the sun and wear the moon

on their clothes. They live in paradisal worlds and subdue monstrous animals in order to leave the world free from harm and ready to renew itself in the fertile acts of spring.

The symbolic reenactments of legend often experiment with alternative social orders and criticize or reverse existing divisions of labour and prestige. In Sicilian-American communities of Texas, Louisiana, California, and elsewhere, the female head of the household dedicates and displays an altar to St. Joseph and thus fulfills a promise made in a moment of need. Normally, in Roman Catholicism, a priest who is a celibate male presides at the liturgy and at devotional services. In this case, however, a woman presides, together with other women who assist her. She prepares fruit, hard-boiled eggs, cakes, fig-filled pastries, pies, and special breads and uses them to decorate a series of tiers stretching from floor to ceiling. She also arranges on this festival altar the figurines of saints, the Virgin Mary, and the Sacred Heart of Jesus. The construction of this panorama of fruitfulness takes nine days, a period that constitutes a ritual novena of prayer and devout action. Representatives who act in the accompanying ceremony play the roles of the Holy Family and other saints important to the altar display. Re-creating the Holy Family's search for room in a Bethlehem inn on the night of the Nativity, the ritual drama builds toward the moment when the altar-giver opens her home to Joseph and Mary. As Mother Mary prepares to give birth to Jesus, the hostess readies her home, heart, and community so that they may become fit dwelling places for the sacred being. The presiding women play the roles of Magi-Kings bearing gifts of food and hospitality to the Holy Family and their entourage, which includes most of the neighbouring community. A single family can host from 500 to 1,000 people in the feast that terminates the celebration.

Sometimes the new Christian mythologies function as counter-theologies or theologies of resistance to the impositions of Christian culture. They criticize the Christian missionary enterprise even while they embrace aspects of the new religion. For instance, biblical and Christian themes now occupy a large part of the mythology of the Makiritare Indians in the upper Orinoco River region of Venezuela. For them, Wanadi is the Supreme Being of great light and, although one being, he exists in three distinct persons (*damodede,* "spirit-doubles"). Over the course of creation and human history, Wanadi has sent his three incarnations to earth in order to create human beings and redeem them from the darkness into which they have fallen. In the end, Wanadi, the god incarnate who comes to save humankind, is crucified by mythical monsters called Fañuru (from the Spanish *españoles:* "Spaniards"), at the instigation of an evil being called Fadre (from the Spanish *padre:* "father" or "priest"). To all appearances, Wanadi was slain by the Fañurus, but, in fact, he cut his own insides out and allowed his inner spirit (*akato*) to dance free of his dead, cast-off body. Before Wanadi's spirit ascends into heaven, he gathers his 12 disciples about him and promises that he will return in a new and glorious body to destroy the evil world and create a new earth.

Unlike the orthodox canon of Christian scripture, which was inscribed and closed in the first centuries, authentic Christian myth and legend have arisen anew in all the centuries of the Christian Era. The course of Christian myth and legend can be traced through the whole of Christian history. It offers a record of the spread of Christianity—through the Mediterranean, eastern and western Europe, Asia, Africa, Oceania, and the Americas—and highlights the diversity of cultures brought into contact with the Christian message of salvation. The diverse religious hopes, heroes, and rites of these cultures continue to shape reinterpretations of the life of Christ and his saintly followers.

Legend and myth constitute a record of critical reflection on Christian reality in all its dimensions—social, political, economic, doctrinal, and scriptural. No social class or geographic region can lay exclusive claim to Christian myth and legend; they fill the stanzas of royally sponsored poets, the visions of utopian philosophers, and the folklore of rural populations. Indeed, many ideas widely held about the workings of salvation (especially regarding the saints, angels, the devil, and the powers of nature) find their origin in legendary episodes rather than biblical text. Through myth and legend, diverse local communities across the globe have creatively absorbed into their rich religious histories the message of Christian salvation and, through the same fabulous means, they have evaluated the impact of Christian temporal power on their world.

(L.E.S.)

# THE CHRISTIAN COMMUNITY AND THE WORLD

## The relationships of Christianity

From the perspectives of history and sociology, the Christian community has been related to the world in diverse and even paradoxical ways. This is reflected not only in changes in this relationship over time but also in simultaneously expressed alternatives ranging from withdrawal from and rejection of the world to theocratic triumphalism. For example, early Christians so consistently rejected imperial deities that they were known as radical atheists, while later Christians so embraced European monarchies that they were known as reactionary theists. Radical medieval Franciscans proclaimed that true Christians should divest themselves of money at the same time that the papacy expended great sums to manipulate the political landscape of Europe. Another classic example of this paradoxical relationship is the early monastic withdrawal from the world that at the same time preserved and transmitted classical culture and learning to medieval Europe. In the modern period some Christian communities regard secularization as a fall from true Christianity; others view it as a legitimate consequence of a desacralization of the world initiated by Christ.

The Christian community is always part of the world in which it exists. Thus, the church has served the typical religious function of legitimating social systems and values and of creating structures of meaning, plausibility, and compensation for society as it faces loss and death. The Christian community has sometimes exercised this religious function in collusion with tribalistic nationalisms (*e.g.,* the "German Christians" and Nazism) by disregarding traditional church tenets. When the Christian community has held to its teachings, however, it has opposed such social systems and values (*e.g.,* the stance of the Confessing Church of Germany against Nazism). Given the inherent fragility of human culture and society, religion in general and the Christian community in particular frequently are conservative forces.

However, the Christian community is not always a conservative force. Its ability to criticize the world was bitterly acknowledged by those Romans who attributed the fall of their empire to Christian undermining of their "civil religion." Contemporary black theology and Latin-American liberation theology share the conviction that God takes the side of the oppressed against the world's injustices. From the perspective of theology or faith, the criticism of the world of which the Christian community itself is a part is the exercise of its commitment to Jesus Christ. For the Christian community, the death and Resurrection of Jesus call into question all structures, systems, and values of the world that claim ultimacy.

The relationship of the Christian community to the world may be seen differently depending upon one's historical, sociological, and theological perspectives because the Christian community is both a creation in the world and an influence upon it. This complexity led the American theologian H. Richard Niebuhr to comment in *Christ and Culture* (1956) that "the many-sided debate about the relations of Christianity and civilization . . . is as confused as it is many-sided."

An influential effort to reduce this confusion to manageable and meaningful patterns was articulated by the German scholar Ernst Troeltsch. He organized the complex relationships of the Christian community to the world into three ideal types of religious social organization: church, sect, and mystical movement. The church is described as a conservative institution that affirms the world and mediates salvation through clergy and sacraments. It is also characterized by inclusivity and continuity, signified by its adherence to infant baptism and historical creeds, doctrines, liturgies, and forms of organization. The objective-institutional character of the church increases as it relinquishes its commitment to eschatological perfection in order to create the *corpus Christianum,* the Christian commonwealth or society. This development stimulates opposition from those who understand the Gospel in terms of personal commitment and detachment from the world. The opposition develops into sects, which are comparatively small groups that strive for subjective, unmediated salvation and that are related indifferently or antagonistically to the world. The exclusivity and historical discontinuity of the sect is signified by its adherence to believers' baptism and efforts to imitate what it believes is the New Testament community. Mystical movements are the expression of a radical religious individualism that strives to interiorize and live out the personal example of Jesus. They are not interested in creating a community but strive toward universal tolerance, a fellowship of spiritual religion beyond creeds and dogmas. The Methodist Church exemplifies the dynamic of these types. The Methodist movement began as a sectarian protest against the worldliness of the Church of England; its success stimulated it to become a church, which in turn spawned various sectarian protests, including charismatic communities.

Church, sect, and mystical movement

Niebuhr further developed Troeltsch's efforts by distinguishing five repetitive types of the Christian community's relations to the world. Niebuhr's types are: Christ against culture, Christ of culture, Christ above culture, Christ and culture in paradox, and Christ the transformer of culture. The first two are expressions of opposition to and endorsement of the world, while the last three share a concern to mediate in distinctive ways the opposition between the first two.

Relationships of Christ and culture

Opposition to the world is exemplified by Tertullian's question, "What has Athens to do with Jerusalem?" This sharp opposition to the world was expressed in the biblical disjunction between the children of God and the children of the world and between "the light" and "the darkness" (1 John 2:15, 4:4–5; Revelation); and it has continued to find personal exponents, such as Leo Tolstoy, and communal expressions, such as the Hutterites.

Endorsement of the world emerged in the 4th century with the imperial legal recognition of Christianity by the Roman emperor Constantine. Although frequently associated with the medieval efforts to construct a Christian commonwealth, this type is present wherever national, social, political, and economic programs are "baptized" as Christian. Thus, its historical expressions may be as diverse as the Jeffersonian United States and Hitlerian Germany.

The other three types that Niebuhr proposed are variations on the theme of mediation between rejection and uncritical endorsement of the world. The "Christ above culture" type sees a continuity between the world and faith. This was probably best expressed by Thomas Aquinas' conviction that grace or the supernatural does not destroy nature but completes it. The "Christ and culture in paradox" type views the Christian community's relationship to the world in terms of a permanent and dynamic tension in which the Kingdom of God is not of this world and yet is to be proclaimed in it. A well-known expression of this position is Martin Luther's law–gospel dialectic, distinguishing how the Christian community is to live in the world as both sinful and righteous at the same time. The conviction that the world may be transformed and regenerated by Christianity ("Christ the transformer of culture") has been attributed to expressions that have theocratic tendencies, such as those of Augustine and John Calvin.

Efforts by scholars such as Troeltsch and Niebuhr to provide typical patterns of Christian relations to the world enable appreciation of the multiformity of these relationships without being overwhelmed by historical data. These models relieve the illusion that the Christian community has ever been monolithic, homogeneous, or static. This "many-sidedness" may be seen in the Christian community's relationships to the state, society, education, the arts, social welfare, and family and personal life. (C.H.L.)

## CHURCH AND STATE

The relationship of Christians and Christian institutions to forms of the political order has shown an extraordinary diversity in the course of church history; there have been, for example, theocratically founded monarchies, democracies, and communist community orders. In various periods, however, political revolution, based on theological foundations, to eliminate older "Christian" state forms has also belonged to this diversity.

In certain eras of church history the aspiration for the Kingdom of God stimulated political and social strivings for its realization that included elements of power and dominion. The political power of the Christian proclamation of the coming sovereignty of God resided in its promise of both the establishment of a kingdom of peace and the execution of judgment.

The church, like the state, has been exposed to the temptation of power. The attempt to establish a kingdom of peace resulted in the transformation of the church into an ecclesiastical state. This took place in the development of the Roman Papal States, but it also occurred to a lesser degree in several theocratic churches and was attempted in Calvin's ecclesiastical state in Geneva in the 16th century. In these cases the state declared itself a Christian state and the executor of the spiritual, political, and social commission of the church; it understood itself to be the representative of the Kingdom of God. This development took place in both the Byzantine and the Carolingian empires as well as in the medieval Holy Roman Empire.

The church as an ecclesiastical state

The struggle between the church, understanding itself as state, and the state, understanding itself as representative of the church, not only dominated the Middle Ages but also continued into the Reformation period. The wars of religion in the era of the Reformation and Counter-Reformation discredited in the eyes of many the theological and metaphysical rationales for a Christian state. In the period of the Enlightenment, this led to the idea of the relationship of church and state as grounded upon ideas of natural law and, with Friedrich Schleiermacher among others, to the advocacy of legal separation of church and state.

**The history of church and state.** *The church and the Roman Empire.* In the early church the attitude of the Christian toward the political order was determined by the imminent expectation of the Kingdom of God, whose miraculous power was already beginning to be visibly realized in the figure of Jesus Christ. The importance of the existing political order was, thus, negligible, as expressed in the saying of Jesus, "My kingship is not of this world." Orientation toward the coming kingdom of peace placed Christians in tension with the state, which made demands upon them that were in direct conflict with their faith.

This contrast was developed most pointedly in the rejection of the emperor cult and of certain state offices—above all, that of judge—to which the power over life and death was professionally entrusted. Although opposition to fundamental orderings of the ruling state was not based upon any conscious revolutionary program, contemporaries blamed the expansion of the Christian Church in the Roman Empire for an internal weakening of the empire on the basis of this conscious avoidance of many aspects of public life, including military service.

The role of the church in weakening the Roman Empire

Despite the early Christian longing for the coming Kingdom of God, even the Christians of the early generations acknowledged the pagan state as the bearer of order in the old eon, which for the time being continued to exist. Two contrary views thus faced one another within the Christian communities. On the one hand, under the influence of Pauline missions, was the idea that the "ruling body"—*i.e.,* the existing political order of the Roman Empire—was "from God . . . for your good" (Romans 13:1–4) and that Christians should be "subject to the govern-

ing authorities." Another similar idea held by Paul (in 2 Thessalonians) was that the Roman state, through its legal order, "restrains" the downfall of the world that the Antichrist is attempting to bring about. On the other hand, and existing at the same time, was the apocalyptic identification of the imperial city of Rome with the great whore of Babylon (Revelation 17:3–7). The first attitude, formulated by Paul, was decisive in the development of a Christian political consciousness. The second was noticeable especially in the history of radical Christianity and in radical Christian pacifism, which rejects cooperation as much in military service as in public judgeship.

*The church and the Byzantine, or Eastern, Empire.* In the Byzantine Empire the emperor Constantine granted himself, as "bishop of foreign affairs," certain rights to church leadership. These concerned not only the "outward" activity of the church but also encroached upon the inner life of the church—as was shown by the role of the emperor in summoning and leading imperial councils to formulate fundamental Christian doctrine and to ratify their decisions.

In the Byzantine era there evolved the concept of what has been called caesaropapism, a system in which the harmony between church and state shifted more and more in favour (in terms of power) of the emperor. His ecclesiastical authority was endowed with the idea of the divine right of kings, which was symbolically expressed in the ceremony of crowning and anointing the emperor. This tradition was later also continued in the Russian realms, where the tsardom claimed a growing authority for itself even in the area of the church.

*The church and Western states.* Conversely, the theocratic claim to dominion by the church freely developed in the sphere of the Roman Catholic Church after the state and administrative organization of the Roman Empire in the West collapsed in the chaos following the barbarian ethnic migrations. In the political vacuum that arose in the West because of the invasion by the German tribes, the Roman Church was the single institution that still preserved in its episcopal dioceses the Roman provincial arrangement. In its administration of justice the church largely depended upon the old imperial law and—in a period of legal and administrative chaos—was viewed as the only guarantor of order. The Roman popes used this power, which was in fact allotted to them by circumstances, to develop a specific ecclesiastical state and to base this state upon a new theocratic ideology—the idea that the pope was the representative of Christ and the successor of Peter. From this perspective the Roman popes detached themselves from the power of the Byzantine emperor, to whom they were indeed subordinate according to prevailing imperial law.

The break between Rome and Byzantium

The Roman bishops beginning with Gregory I the Great (reigned 590–604) turned to missionizing the peoples of the West. Under Gregory the church in Spain, Gaul, and northern Italy was strengthened, and England was converted to Roman Christianity. Succeeding popes convinced the rulers of the Frankish (Germanic) kingdom in the 8th century of their leadership role; they also succeeded in winning them as protectors of the papal dominion. These rulers were the first of the German kings to join themselves to the Roman Church. The relationship created a new area of tension. Whereas rulers considered the pope as a member of the Christian state and therefore under its protection and laws, the popes saw rulers as members of the church and therefore subject to the rule of God through St. Peter's successors. Moreover, the emperor Charlemagne claimed for himself the right to appoint the bishops of his empire, who were more and more involved in political affairs. These conflicting perspectives were the cause of interminable struggles between popes and rulers throughout the Middle Ages.

The feudalization of the church in the West

In the course of this development, the process of the feudalization of the church—unique in church history—occurred. Ruling political leaders in this system occupied significant positions in the church; by virtue of patronage this development encompassed the whole imperial church. At the conclusion of this development, bishops in the empire were simultaneously the reigning princes of their dioceses; they often were much more interested in the political tasks of their dominion than in the spiritual.

In the great church-renewal movement, which extended from its beginnings at the monastery at Cluny (France) in the 10th century and lasted until the reign of Pope Gregory VII in the 11th century, the papal church rejected both the sacred position of the king and the temporal position of bishops, who were awarded their rights and privileges by the king. This renewal movement proclaimed the freedom of the church from state authority as well as its preeminence over worldly powers. This struggle, now remembered as the Investiture Controversy, was fought out as a dramatic altercation between the papacy and the empire. The church did not, however, gain a complete victory in terms of papal claims of full authority over the worldly as well as the spiritual realms.

With the weakening of the Holy Roman Empire, the European nation-states arose as opponents of the church. The papal ideology had developed with respect to controlling emperors and was not suited to deal effectively with kings of nation-states. This was first clearly evident with the humiliation of Pope Boniface VIII by King Philip of France and the subsequent Babylonian Captivity of the church, when the papacy was forced to reside in Avignon (1309–77).

Contributing to the strengthening of the nation rulers' right of ecclesiastical supervision was the problem of papal schism, initiated upon the return of the papacy to Rome by the deposition of one pope and the election of another, with both claiming legitimacy. Popes and counterpopes reigning simultaneously mutually excommunicated one another, thus demeaning the esteem of the papacy. The schisms spread great uncertainty among the believers of the empire about the validity of the consecration of bishops and the sacraments as administered by the priests they ordained. The schism also fueled desires for a parliamentary form of church government and contributed to the rise of the 15th-century conciliar movement, which posited the supreme authority of ecumenical councils in the church.

The 16th-century Reformation forced the church to face its purely spiritual tasks and placed Reformation law as well as the legal powers of church leadership in the hands of the princes. Under King Henry VIII a revolutionary dissociation of the English Church from papal supremacy took place. In the German territories the reigning princes became, in effect, the legal guardians of the Protestant episcopate—a movement already in the process of consolidation in the late Middle Ages. The development in the Catholic nation-states, such as Spain, Portugal, and France, occurred in a similar way.

Church leadership under the control of the state

The democratic ideas of the freedom and equality of Christians and their representation in a communion of saints by virtue of voluntary membership had been disseminated in various medieval sects (*e.g.*, Cathari, Waldenses, Hussites, and the Bohemian Brethren) and were reinforced during the Reformation by groups such as the Hutterites, Mennonites, and Schwenckfelders and the followers of Thomas Müntzer. Under the old ideal of an uncompromising realization of the Sermon on the Mount, there arose anew in these groups a renunciation of certain regulations of the state, such as military service and the acceptance of state offices (judgeship), a radical pacifism, and the attempt to structure their own form of common life in Christian, communist communities. Many of their political ideas—at first bloodily suppressed by the Reformation and Counter-Reformation states and churches—were later prominent in the Dutch wars of independence and in the English Revolution, which led to a new relationship between church and state.

In the Thirty Years' War (1618–48), confessional antitheses were settled in devastating religious wars, and the credibility of the feuding ecclesiastical parties was thereby called into question. Subsequently, from the 17th century on, the tendency toward a new, natural-law conception of the relationship between state and church was begun and continued. Henceforth, in the Protestant countries, state sovereignty was increasingly emphasized vis-à-vis the churches. The state established the right to regulate ed-

ucational and marriage concerns as well as all foreign affairs of the church. A similar development also occurred in Roman Catholic areas. In the second half of the 18th century Febronianism demanded a replacement of papal centralism with a national church episcopal system; in the German *Reich* an enlightened state-church concept was established under Josephinism (a view advocated by Joseph II [reigned 1765–90]) through the dismantling of numerous ecclesiastical privileges. The Eastern Orthodox Church also was drawn into this development under Peter the Great.

The separation of church and state

The separation of church and state as proclaimed during the French Revolution in the latter part of the 18th century was the result of Reformational strivings toward a guarantee for the freedom of the church and the natural-law ideas of the Enlightenment; it was aggravated by the social revolutionary criticism against the wealthy ecclesiastical hierarchy. The separation of church and state was also achieved during and after the American Revolution as a result of ideas arising from the struggle of the Puritans against the English episcopal system and the English throne. After the state in France had undertaken the task of creating its own political, revolutionary substitute religion in the form of a "cult of reason," which was foreshadowed by Rousseau's discourse on "la religion civile," a type of separation of church and state was achieved. The French state took over education and other hitherto churchly functions of a civic nature.

From the late 18th century on, two fundamental attitudes developed in matters related to the separation of church and state. The first, as implied in the Constitution of the United States, was supported by a tendency to leave to the church, set free from state supervision, a maximum freedom in the realization of its spiritual, moral, and educational tasks. In the United States, for example, a comprehensive church school and educational system has been created by the churches on the basis of this freedom, and numerous universities have been founded by churches. The separation of church and state by the French Revolution and later in the Soviet Union and the countries under the Soviet Union's sphere of influence was based upon an opposite tendency. The attempt was to totally exterminate the church and to replace it with nationalism.

In contrast to this, National Socialism in Germany under Hitler showed paradoxical contradictions. On the one hand, Nazi propaganda pursued a consciously anti-Christian polemic against the church; it proceeded to arrest those clergy opposed to the Nazi worldview and policies.

Hitler's concordat with the Vatican

On the other hand, Hitler placed the greatest value upon concluding with the Vatican in 1934 a concordat that granted the Roman Catholic Church more special rights in the German *Reich* than had ever been granted it in any earlier concordat. The concordat with the Vatican represented the first recognition of the Hitler regime by a European government and was viewed by Hitler as a method of entrance into the circle of internationally recognized political powers.

In Germany the old state-church traditions had already been eliminated in the revolution of 1918, which, with the abolition of the monarchical system of government, also deprived the territorial churches of their supreme Protestant episcopal heads. In the German Weimar Constitution the revolution had earlier sanctioned the separation of church and state. State-church traditions were maintained in various forms in Germany, not only during the Weimar Republic but also during the Hitler regime and afterward in the Federal Republic of Germany. Thus, through state agreements, definite special rights, primarily in the areas of taxes and education, were granted to both the Roman Catholic Church and the Evangelical (Lutheran-Reformed) churches of the individual states.

Even in the United States, however, the old state-church system, overcome during the American Revolution, still produces aftereffects in the form of tax privileges of the church (exemption from most taxation), the exemption of the clergy from military service, and the financial furtherance of confessional school and educational systems through the state. These privileges have been questioned and even attacked by certain segments of the American public.

**Church and state in Eastern and Western theology.** The two main forms of the relationship between church and state that have been predominant and decisive through the centuries and in which the structural difference between the Roman Catholic Church and Eastern Orthodoxy becomes most evident can best be explained by comparing the views of two great theologians: Eusebius of Caesarea and Augustine.

*The views of Eusebius of Caesarea.* Eusebius of Caesarea (*c.* 260–*c.* 340) was the court theologian of Emperor Constantine the Great, who formed the Orthodox understanding of the mutual relationship of church and state. He saw the empire and the imperial church as sharing a close bond with one another; in the centre of the Christian empire stood the figure of the Christian emperor rather than that of the spiritual head of the church.

The mutual relationship of church and state

Eusebius made this idea the basis of his political theology, in which the Christian emperor appears as God's representative on Earth in whom God himself "lets shine forth the image of his absolute power." He is the "Godloved, three times blessed" servant of the highest ruler, who, "armed with divine armor cleans the world from the horde of the godless, the strong-voiced heralds of undeceiving fear of God," the rays of which "penetrate the world." Through the possession of these characteristics the Christian emperor is the archetype not only of justice but also of the love of humankind. When it is said about Constantine, "God himself has chosen him to be the lord and leader so that no man can praise himself to have raised him up," the rule of the Orthodox emperor has been based on the immediate grace of God.

The Orthodox interpretation of the Christian emperor

This religious interpretation of the Christian emperor reinterpreted in the Christian sense the ancient Roman institution of the god-emperor. Some of Eusebius' remarks echo the cult of the Unconquered Sun, the Sol Invictus, who was represented by the emperor according to pagan understanding. The emperor—in this respect he also resembled the pagan god-emperor who played the role of the *pontifex maximus* (high priest) in the state cult—took the central position within the church as well. He summoned the synods of bishops, "as though he had been appointed bishop by God," presided over the synods, and granted judicial power for the empire to their decisions. He was the protector of the church who stood up for the preservation of unity and truth of the Christian faith and who fought not only as a warrior but also as an intercessor, as a second Moses during the battle against God's enemies, "holy and purely praying to God, sending his prayers up to him." The Christian emperor entered not only the political but also the sacred succession of the Roman god-emperor. Next to such a figure, an independent leadership of the church could hardly develop.

Orthodox theologians have understood the coexistence of the Christian emperor and the head of the Christian church as *symphōnia,* or "harmony." The church recognized the powers of the emperor as protector of the church and preserver of the unity of faith and limited its own authority to the purely spiritual domain of preserving the Orthodox truth and order in the church. The emperor, on the other hand, was subject to the spiritual leadership of the church as far as he was a son of the church.

The relationship between patriarch and emperor

The special position of the imperial ruler and the function of the Byzantine patriarch as the spiritual head of the church have been defined in the *Epanagoge,* the judicial ruling establishing this relationship of church and state. The church-judicial affirmation of this relationship in the 6th and 7th centuries made the development of a judicial independence of the Byzantine patriarch in the style of the Roman papacy impossible from the beginning.

The *Epanagoge,* however, did not completely subject the patriarch to the supervision of the emperor but rather directed him expressly "to support the truth and to undertake the defense of the holy teachings without fear of the emperor." Therefore, the tension between the imperial reign that misused its absolutism against the spiritual freedom of the church and a church that claimed its spiritual freedom against an absolutist emperor or tsar was

characteristic for the Byzantine and Slavic political history but not the same as the political tension between the imperial power and the politicized papacy that occurred in the West.

The answers of the *City of God*

*The views of Augustine.* Augustine's *City of God* attempted to answer the most painful event of his century: the fall of Rome. Augustine responded to the existential shock and dismay his contemporaries experienced with the collapse of their world by a literary demolition of their nostalgic paganism. From Augustine's perspective the "splendid vices" of the pagans had led inexorably to the fall of an idolatrous world. In sharp contrast to this "earthly city," epitomized by Rome but everywhere energized by the same human desires for praise and glory, Augustine projected the "most glorious city" of praise and thanks to God, the heavenly Jerusalem, a historical image of which was the new Rome of the Catholic Church. However, Augustine did not simply identify the state with the earthly city and the church with the city of God. He perceived that the state existed not simply in opposition to God but as a divine instrument for the welfare of humankind. The *civitas dei* and the *civitas terrena* finally correspond neither to church and state nor to heaven and earth. They are rather two opposed societies with antagonistic orders of value that intersect both state and church and in each case show the radical incompatibility of the love of God with the values of worldly society.

*Later developments.* Based upon Augustine's views, the historical development of the church in the Latin West took a different course, one away from the Byzantine imperial church. In the West a new power was formed—the Roman Church, the church of the bishop of Rome. This church understood itself as the successor of the extinct Roman Empire. In the political vacuum of the West that was created by the invasion of the Germans and the destruction of the Roman state and administrative apparatus, the church became great and powerful as the heir to the Roman Empire. Only within this vacuum could the idea of the papacy develop in which the great popes, as bishops of Rome, stepped into the position of the vanished emperors.

The Donation of Constantine

It was in this context that the judicial pretense of the "Gift of the emperor Constantine"—the Donation of Constantine—became possible, to which the later development of the papacy was connected. The Donation attempted to reconstruct the history of the Roman papacy in retrospect in order to make legitimate the newly gained ecclesiastical and political position of the popes after the extinction of the Western Roman imperial reign. This fabrication entered papal ideology in written form through the mid-9th-century resource for canon law known as the Pseudo-Isadorian Decretals. The exposure of the Donation as a forgery did not occur until the 15th century. The Donation is the account of Constantine's purported conferring upon Pope Sylvester I (reigned 314–335) of the primacy of the West, including the imperial symbols of rulership. The Pope returned the crown to Constantine, who in gratitude moved the capital to Byzantium (Constantinople). The Donation thereby explained and legitimated a number of important political developments and papal claims, including the transfer of the capital to Byzantium, the displacement of old Rome by the new Rome of the church, papal secular authority, and the papal right to create an emperor by crowning him. The latter would be used to great effect when Pope Leo III crowned Charlemagne king of the Romans in 800. The force of this action was of great significance throughout the Middle Ages as popes exerted authority over the emperors of the Holy Roman Empire, and it explains the symbolic significance of Napoleon's taking the crown from Pope Pius VII's hands to crown himself.

This was the point of separation from which the developments in the East and in the West led in two different directions. The growing independence of the West was markedly illustrated by the Donation of Pepin (Pepin, father of Charlemagne, was anointed king of the Franks by Pope Stephen III in 754), which laid the foundation of the Papal States as independent of any temporal power and gave the pope the Byzantine exarchate of Ravenna. At this time the development of two different types of a Christian idea of the state and of the church began, and it subsequently ended in the schism between Rome and Byzantium in 1054.

The idea of the church as a state existed not only in the Roman theocracy and in the papal idea of the church, but it also appeared in a new democratic form and in strict contrast to its absolutist Roman model in some Reformation church and sect developments and in Free churches of the post-Reformation period. The sects of the Reformation period renewed the old idea of the Christian congregation as God's people, wandering on this Earth—a people connected with God, like Israel, through a special covenant. This idea of God's people and the special covenant of God with a certain chosen group caused the influx of theocratic ideas, which were expressed in forms of theocratic communities similar to states and led to formations similar to an ecclesiastical state. Such tendencies were exhibited among radical Reformation groups (*e.g.,* the Münster prophets), Puritans in Massachusetts, and various groups of the American Western frontier. One of the rare exceptions to early modern theocratic theology was Luther's sharp distinction of political and ecclesial responsibilities by his dialectic of law and gospel. He commented that it is not necessary that an emperor be a Christian to rule, only that he possess reason.

The Mormon attempt to establish a theocratic state

The latest attempt to form a church-state by a sect that understood itself as the chosen people distinguished by God through a special new revelation was undertaken by the Mormons, the "Latter-day Saints." Based on the prophetic direction of their leaders, they attempted to found the state Deseret, after their entrance into the desert around the Great Salt Lake in Utah. The borders of the state were expected to include the largest part of the area of the present states of Utah, California, Arizona, Nevada, and Colorado. The Mormons, however, eventually had to recognize the fact that the comparatively small centre state, Utah, of the originally intended larger Mormon territory, could not exist as a theocracy (though structured as other secular models) under a government of Mormon Church leaders. Reports (some apparently spurious) by federal agents hostile to the church and widespread revulsion toward the Mormon practice of polygyny mitigated against federal sanction of the church leadership as the governmental heads of the proposed state. Utah eventually became a federal state of the United States.

The enlightened absolutist state of the 18th century basically took over the secularized form of the old Christian government that consciously took into account the equality of Christian denominations.

## CHURCH AND SOCIETY

Preserving or improving the status quo

Ever since the Reformation, the development of Christianity's influence on the character of society has been twofold. In the realm of state churches and territorial churches, its influence has been a strong element in preserving the status quo of society. Thus, in England, the Anglican Church remained an ally of the throne, as did the Protestant churches of the German states. In Russia the Orthodox Church continued to support the feudal society founded upon the monarchy, and even the monarch carried out a leading function within the church as protector.

Though the impulses for transformation of the social order according to the spirit of the Christian ethic came more strongly from the radical Free churches and sects, churches within the established system of state and territorial churches made positive contributions in improving the status quo. In 17th- and 18th-century Germany, Lutheran clergy, such as August Francke (1663–1727), were active in establishing poorhouses, orphanages, schools, and hospitals. In England, Anglican clergymen, such as Frederick Denison Maurice and Charles Kingsley in the 19th century, began a Christian social movement in the throes of the Industrial Revolution. Their movement brought a Christian influence to the conditions of life and work in industry. Johann Hinrich Wichern proclaimed, "There is a Christian Socialism," at the Kirchentag Church Convention in Wittenberg, Ger., in 1848, the year of the publication of the *Communist Manifesto,* and created the

"Inner Mission" in order to address "works of saving love" to all suffering spiritual and physical distress. The diaconal movements of the Inner Mission were concerned with social issues, prison reform, and care of the mentally ill. Only in tsarist Russia did the church fail in matters concerning social problems and the Industrial Revolution.

The Anglo-Saxon Free churches made great efforts to bring the social atmosphere and living conditions into line with a Christian understanding of human life. Methodists and Baptists addressed their message mainly to those segments of society that were neglected by the established church. They recognized that the distress of the newly formed working class, a consequence of industrialization, could not be removed by the traditional charitable means used by the state churches. The fact that in Germany, in particular, the spiritual leaders of the so-called revival movement, such as Friedrich Wilhelm Krummacher (1796–1868) and others, denied the right of self-organization to the workers by claiming that all earthly social injustices would receive compensation in heaven caused Karl Marx and Friedrich Engels to separate themselves completely from the church and its purely charitable attempts at a settlement of social conflicts and to declare religion with its promise of a better beyond as the "opiate of the people." This reproach, however, was as little in keeping with the social-ethical activities of the Inner Mission and of Methodists and Baptists as it was with the selfless courage of the Quakers, who fought against social demoralization, against the catastrophic situation in the prisons, and, most of all, against slavery.

The consequences of industrialization

**The problem of slavery and persecution.** The fight against slavery has passed through many controversial phases in the history of Christianity. Paul recommended to Philemon that he accept back his runaway slave Onesimus, "no longer as a slave but more than a slave, as a beloved brother . . . both in the flesh and in the Lord" (verse 16). Although the biblical writings made no direct attack upon the ancient world's institution of slavery, its proleptic abolition in community with Christ—"There is neither Jew nor Greek, there is neither slave nor free, there is neither male nor female; for you are all one in Christ Jesus" (Galatians 3:28)—has been a judgment upon the world's and the Christian community's failure to overcome slavery and all forms of oppression. Medieval society made only slow progress in the abolition of slavery. One of the special tasks of the orders of knighthood was the liberation of Christian slaves who had fallen captive to the Muslims; and special knightly orders were even founded for the ransom of Christian slaves.

With the discovery of the New World, the institution of slavery grew to proportions greater than had been previously conceived. The widespread conviction of the Spanish conquerors of the New World that its inhabitants were not really human in the full sense of the word and therefore could be made slaves in good conscience added to the problem. The attempt of missionaries, such as Bartolomé de Las Casas in 16th-century Peru, to counter the inhuman system of slavery in the colonial economic systems finally introduced the great basic debate concerning the question of human rights. A decisive part in the elaboration of the general principles of human rights was taken by the Spanish and Portuguese theologians of the 16th and 17th centuries, especially Francisco de Vitoria. Even modern natural law, however, could still be interpreted in a conservative sense that did not make slavery contrary to its provisions. Puritanism, however, fought against slavery as an institution. In German Pietism, Nikolaus Ludwig, Graf von Zinzendorf, who became acquainted with slavery on the island of Saint Croix in the Virgin Islands, used his influence on the King of Denmark for the human rights of the slaves. The Methodist and Baptist churches advocated abolition of slavery in the United States in the decisive years preceding the foundation of the New England Anti-Slavery Society in Boston in 1832 by William Lloyd Garrison. In regard to the fight against slavery in England and in The Netherlands, which was directed mainly against the participation of Christian trade and shipping companies in the profitable slave trade, the Free churches were very active. The overcoming of the institution of slavery did not end racial discrimination. Martin Luther King, Jr., Baptist pastor and Nobel laureate, led the struggle for civil rights in the United States until his assassination in 1968. In South Africa in the 1980s, Desmond Tutu, Anglican archbishop and Nobel laureate, exemplified a continuing Christian struggle for human rights.

The beginning of the abolition of slavery

The fight against slavery is only a model case in the active fight of the Christian churches and fellowships against numerous other attempts at desecration of a Christian understanding of the nature of humanity, which sees in every human being a neighbour created in God's image and redeemed by Christ. Similar struggles arose against the persecution of the Jews and the elimination of members of society characterized by political or racist ideology as "inferior." In Germany the members of the Confessing Church fought against the practices of National Socialism, which called for the elimination of the mentally ill and the inmates of mental and nursing institutions, who were considered "unfit to live."

**Theological and humanitarian motivations.** Decisive impulses for achieving changes in the social realm in the sense of a Christian ethic have been and are initiated by men and women in the grasp of a deep personal Christian experience of faith, for whom the message of the coming Kingdom of God forms the foundation for faithful affirmation of social responsibility in the present world. Revival movements have viewed the Christian message of the Kingdom of God mainly as an impulse for reorganization of the secular conditions of society in the sense of a Kingdom of God ethic. Under the leadership of an American Baptist theologian, Walter Rauschenbusch (1861–1918), the so-called Social Gospel movement spread in the Anglo-Saxon countries. A corresponding movement was started with the Christian social conferences by German Protestant theologians, such as Paul Martin Rade (1857–1940) of Marburg. The basic idea of the Social Gospel—*i.e.,* the emphasis on the social-ethical tasks of the church—gained widespread influence within the ecumenical movement and especially affected Christian world missions. In many respects modern economic and other forms of aid to developing countries—including significant ecumenical contributions from the World Council of Churches, the World Alliance of Reformed Churches, the Lutheran World Federation, and the Roman Catholic Church—have now succeeded the Social Gospel.

Decisive impulses for achieving social change

There is concern on the part of some Christians that these developments reduce the Christian message to a purely secular social program that is absorbed by political programs. Others in the Christian community believe that faithful responsibility in and to the world requires political, economic, and social assistance to oppressed peoples with the goal of their liberation to a full human life.

Changes and problems in the Christian social message

## CHURCH AND EDUCATION

**Intellectualism versus anti-intellectualism.** In contrast to Tertullian's anti-intellectual attitude, an exactly opposite attitude toward intellectual activities has also made itself heard from the beginning of the Christian Church (*e.g.,* by Clement of Alexandria). It also has its basis in the nature of Christian faith. In the 11th century Anselm of Canterbury expressed it in the formula *fides quaerens intellectum* ("faith seeking understanding"), a formula that has become the rallying point for scholastics of all times. Because people have been endowed with reason, they have an urge to express their experience of faith intellectually, to translate the contents of faith into concepts, and to formulate beliefs in a systematic understanding of the correlation between God, humankind, and creation. Christians of the 1st century came from the upper levels of society and were acquainted with the philosophy and natural science of their time. Justin Martyr, a professional philosopher, saw Christian revelation as the fulfillment, not the elimination, of philosophical understanding. The Logos term of the opening chapter of the Gospel According to John is the point of departure for the intellectual history of salvation. The light of the Logos (a Greek word meaning "word" or "reason," with the sense of divine or universal reason permeating the intelligible world) had made itself manifest in a number of sparks and seeds in

human history even before its incarnation in the person of Jesus Christ.

These two contrasting opinions have stood in permanent tension with one another. In medieval Scholasticism the elevation of Christian belief to the status of scientific universal knowledge was dominant. Theology became the instructor of the different sciences, organized according to the traditional classification of trivium (grammar, rhetoric, and dialectic) and quadrivium (music, arithmetic, geometry, and astronomy) and incorporated into the system of education as "servants of theology." This system of education became part of the structure of the universities that were founded in the 13th century. The different sciences only gradually gained a certain independence.

Theology as the instructor of the various sciences

With the Reformation there was widespread concern for education because the Reformers desired everyone to be able to read the Bible. Their concern was the beginning of universal, public education. Luther also argued that it was necessary for society that its youth be educated. He held that it was the duty of civil authorities to compel their subjects to keep their children in school so "that there will always be preachers, jurists, pastors, writers, physicians, schoolmasters, and the like, for we cannot do without them."

Hostility toward intellectual endeavours

Open conflict between science and theology occurred only when the traditional biblical view of the world was seriously questioned, as in the case of the Italian astronomer Galileo (1633). The principles of Galileo's scientific research, however, were themselves the result of a Christian idea of science and truth. The biblical faith in God as Creator and incarnate Redeemer is an explicit affirmation of the goodness, reality, and contingency of the created world—assumptions underlying scientific work. Thus, in the 20th century, William Temple, archbishop of Canterbury, could assert that Christianity is an avowedly materialistic religion. Positive tendencies concerning education and science have always been dominant in the history of Christianity, even though the opposite attitude arose occasionally during certain periods. Thus the German astronomer Johannes Kepler (1571–1630) spoke of celebrating God in science.

The attitude that had been hostile toward intellectual endeavours was less frequently heard after the Christian Church had become the church of the Roman Empire. But the relationship between science and theology was also attacked when the understanding of truth that had been developed within theology was turned critically against the dogma of the church itself. This occurred, for instance, after the natural sciences and theology had turned away from total dependence upon tradition and directed their attention toward experience—observation and experiment. A number of fundamental dogmatic principles and understandings were thus questioned and eventually abandoned. The struggle concerning the theory of evolution (*e.g.*, the Scopes Trial in Tennessee in 1925) has been a conspicuous modern symptom of this trend.

The estrangement between theology and science

The estrangement of theology and natural science in the modern period was a complex development related to confessional controversies and wars in the 16th and 17th centuries and philosophical perspectives in the 18th and 19th centuries. The epistemological foundation of faith was radically called into question by the Scottish philosopher David Hume. Building upon Hume's work, the German philosopher Immanuel Kant advocated freedom from any heteronomous authority, such as the church and dogmas, that could not be established by reason alone. Scholars withdrew from the decisions of church authorities and were willing only to subject themselves to critical reason and experience. The rationalism of the Enlightenment appeared to be the answer of science to the claim of true faith that had been made by the churches, which had become untrustworthy through the religious wars and the influence of philosophy.

**Forms of Christian education.** The Christian Church created the bases of the Western system of education. From its beginning the Christian community faced external and internal challenges to its faith, which it met by developing and utilizing intellectual and educational resources. The response to the external challenge of rival religions and philosophical perspectives is termed apologetics—*i.e.*, the intellectual defense of the faith. Apologetic theologians from Justin Martyr in the 2nd century to Paul Tillich in the 20th have promoted critical dialogue between the Christian community, the educated world, and other religions. The internal challenges to the Christian community were met not only by formulating the faith in creeds and dogmas but also by passing this faith on to the next generations through education.

In the early Middle Ages a system of schools was formed at the seats of bishops to educate clergy and to teach the civil servants of the government and administrative offices. The school at the court of Charlemagne (which was conducted by clergy), the medieval schools of the religious orders, cathedrals, monasteries, convents, and churches, the flourishing schools of the Brethren of the Common Life, and the Roman Catholic school systems that came into existence during the Counter-Reformation under the leadership of the Jesuits and other new teaching orders contributed much to the civilization of the West. Equally important were the schools and educational reforms started by the German Reformers Luther, Philipp Melanchthon, Johann Bugenhagen, John and August Hermann Francke, and by the Moravian reformers John Amos Comenius and the Graf von Zinzendorf. The church was responsible for the system of schools even after the Reformation. Only in the 18th century did the school system start to separate itself from its Christian roots and fall more and more under state control.

Effect of the separation of church and state

With the separation of church and state, both institutions have entered into tensely manifold relationships. In some countries the state has taken over the school system completely and does not allow private church schools except in a few special cases in which constant control is maintained regarding religious instruction as a part of the state's educational task. Other countries (*e.g.*, France) maintain school systems basically free of religion and leave the religious instruction to the private undertakings of the different churches. In the American Revolution the concept of the separation of state and church was a lofty goal that was supposed to free the church from all patronization by the state and to make possible a maximum of free activity, particularly in the area of education. On the other hand, the Soviet Union has used its schools particularly for an anti-religious education based upon the state philosophy of dialectical materialism, practicing the constitutionally guaranteed freedom of anti-religious propaganda in schools, though the churches are forbidden to give any education outside their worship services.

A second issue that results from the separation of church and state is the question of state subsidies to private church schools. These are claimed in those countries in which the church schools in many places take over part of the functions of the state schools (*e.g.*, in the United States). After the ideological Positivism and the Materialism of the 19th century faded away in many areas, it was realized that religious life had had an important role in the cultural development of the West and the New World and that the practiced exclusion of religious instruction from the curricula of the schools indicated a lack of balance in education. Based on new insights, it has therefore been maintained in the 20th century that religion should be adopted as a subject among the humanities. State universities in the United States, Canada, and Australia, which did not have theological faculties because of the separation of church and state, founded departments of religion of an interdenominational nature and included non-Christians as academic teachers of religion.

The founding of medieval universities

The Christian system of education led to the early founding of universities. The university was a creation of medieval Europe and spread from there to other continents after the 16th century. The universities that had been formed in the beginning through the unification of schools for monks and schools for regular clergy succeeded in gaining their relative independence by agreements with church and state. The universities represented the unity of education that was apparent in the common use of the Latin language, the teaching methods of lecture and disputation, the extended communal living in colleges, the

periodically changing leadership of an elected dean, the inner structure according to faculties or "nations," and the European recognition of the academic degrees.

The advent of humanism and the Reformation created a new situation for all systems of education, especially the universities. Humanists demanded plans to provide designated places for free research in academies that were princely or private institutions and, as such, not controlled by the church. On the other hand, the Protestant states of the Reformation created their own new state universities, such as Marburg in 1527, Königsberg in 1544, and Jena in 1558. As a counteraction, the Jesuits took over the leadership in the older universities that had remained Roman Catholic or else founded new ones in Europe and overseas.

Education in mission areas

In overseas areas, Christian education has had a twofold task. First, its function was to lay an educational foundation for evangelization of non-Christian peoples by forming a system of education for all levels from grammar school to university. Second, its function was to take care of the education of European settlers. To a large extent the European colonial powers had left the formation of an educational system in their colonies or dominions to the churches. In the Spanish colonial regions in America, Roman Catholic universities were founded very early (*e.g.,* Santo Domingo in 1538, Mexico and Lima in 1551, Guatemala in 1562, and Bogotá in 1573). In China, Jesuit missionaries acted mainly as agents of European education and culture (*e.g.,* astronomy, mathematics, and technology) in their positions as civil servants of the court.

Since the 18th century, the activities of competing Christian denominations in mission areas has led to an intensification of the Christian system of education in Asia and Africa. Even where the African and Asian states have their own system of schools and universities, Christian educational institutions have performed a significant function (St. Xavier University in Bombay and Sophia University in Tokyo are Jesuit foundations, Dōshisha University in Kyōto is a Japanese Presbyterian foundation).

Denominational colleges in the United States

In North America Christian education took a different course. From the beginning the churches took over the creation of general educational institutions. The various denominations did pioneer work in the field of education; a state school system was established only after the situation had consolidated itself. In the English colonies, later the United States, the denominations founded theological colleges for the purpose of educating their ministers and established universities dealing with all major disciplines, including theology, often emphasizing a denominational slant. Harvard University was founded in 1636 and Yale University in 1701 as Congregational establishments, and the College of William and Mary was established in 1693 as an Anglican institution. They were followed during the 19th century by other Protestant universities (*e.g.,* Southern Methodist University, Dallas, Texas) and colleges (*e.g.,* Augustana College, Rock Island, Ill.). In addition, many private universities were based upon a Christian idea of education according to the wishes of their founders.

The purpose of confirmation instruction

Christian education has been undertaken in a variety of forms. The system of Sunday schools is nearly universal in all denominations. Confirmation instruction is more specialized, serving different tasks, such as preparation of the children for confirmation, their conscious acknowledgment of the Christian ethic, of the Christian confessions, of the meaning of the sacraments, and of the special forms of congregational life. (E.W.B./C.H.L.)

## CHURCH AND THE ARTS

The Christian community has fostered the development of a Christian culture in all areas of life. In this development the arts have played major roles in expressing, communicating, and deepening the faith of the community. At times some art has been rejected because either the content or the form was perceived as incompatible with the faith of the community. For example, the antipathy of the early church, especially in Orthodox realms, to theatre was related to the perception that the Greek theatre was preoccupied with ancient myths and deities and that comedy excused immorality. Instrumental music also was rejected in Orthodoxy because of both its accustomed role in pagan cults and the belief that God should not be praised with "dead" materials. The Orthodox community did not completely reject dramatic and musical art, however, but transformed them into the service of the church through rich liturgical dramas and extraordinary choir music.

The didactic use of art

The didactic use of art, succinctly expressed by Pope Gregory I the Great in the phrase "images are the books of the laity," was present in the architecture and art of the churches constructed up to the modern period. The symbolic significance of the church building in the shape of a cross, its deep entrances leading into sacred space, the carefully engineered proportions of the building, the use of light and shadow in relation to statues and stained glass, the smells of candles and incense, not to mention the liturgy itself—all these forms, colours, sights, sounds, and smells worked to communicate a sense of the sacred. The sculptures and stained-glass windows were graphic presentations of biblical stories and moral instructions to which the preacher could point and to which the largely illiterate congregation could return for reflection and edification. The very location of the church in the centre of medieval towns and villages gave physical expression to the community's faith in God's presence in the world.

Until the Renaissance the arts were patronized by and in service to the Christian community. Since then there has been a growing independence of artists from the church. In the modern period the older "Christ against culture" concerns of the early church have once again been raised. The Christian community has by and large not found the criterion of religious subject matter satisfactory for its decisions about art. While the music of Richard Wagner may be regarded as inappropriate for the church because of his preoccupation with Teutonic gods and goddesses, a painting such as Picasso's "Guernica" depicting human sin and evil may be appropriate. At the same time a musical, literary, plastic, or graphic representation of Jesus may either vitiate and trivialize the faith of the community, as in modern sentimentalized "portraits" of Jesus, or profoundly express that faith, as in Matthias Grünewald's "Isenheim Altarpiece" portraying the Crucifixion.

The Christian community, sensitized by its Jewish roots and the Hebrew prohibition of idolatry, has been aware that the beauty of the holy may be twisted into the holiness of beauty. Thus, at various times in history there have been iconoclastic reactions to art. But for the most part the Christian community has appreciated and contributed to the didactic, expressive, and symbolic representation of faith and the human condition. (C.H.L.)

## CHURCH AND SOCIAL WELFARE

**Curing and caring for the sick.** *Healing the sick.* The Christian Church has administered its concern for the sick in a twofold manner: both by healing the sick and by expressing concern and caring for them. The practice of healing has retreated into the background in modern times, but healing played a decisive role in the success of the early church and was important in missionary apologetics. In the Gospels, Jesus appears as a healer of body and soul. The title "Christ the Physician" was the most popular name for the Lord in missionary preaching of the first centuries. Even the Apostles are characterized as healers. The apologetics of the church of the 2nd to 4th century used numerous miraculous healings as arguments for the visible presence of the Holy Spirit in the church. The Fathers of the first centuries interpreted the entire sphere of charismatic life from the basic concepts that Christ is the physician, the church the hospital, the sacraments the medication, and orthodox theology the medicine chest against heresy. Ignatius of Antioch called the Eucharist the "medication that produces immortality."

The history of charismatic healing has hardly been explored. Miracles of healing remain a characteristic attribute of the great Christian charismatics of the Roman Catholic Church as well as of the Eastern Orthodox. Healing within the church began to retreat only in connection with the transformation of the church into a state church under Constantine and with the replacement of free charismatics by ecclesiastical officials.

The early basis for healing was generally a demonological

interpretation of sickness: healing was often carried out as an exorcism—that is, a ceremonial liturgical adjuration of the demon that was supposed to cause the illness and its expulsion from the sick person. The development of exorcism is characteristic in that the office of the exorcist eventually became one of the lower levels of ordination, which led to the priesthood. Traditionally, exorcisms were connected not with the rite of baptism alone; the *Rituale Romanum* (*Roman Ritual*) contains many liturgical formulas for cases of demoniacal possession. Only the Enlightenment in the 18th century repressed the practice of exorcisms within the Roman Catholic Church.

The development of exorcism

In the churches of the Reformation, exorcism never completely vanished; in Pietistic circles several exorcists have appeared; *e.g.,* Johann Christoph Blumhardt the Elder (1805–80). With the motto "Jesus is Conquerer," Blumhardt transformed his healing centre at Bad Boll, Ger., into an influential resource for international missionary work. His son, Christoph Friedrich Blumhardt (1842–1919), continued his father's work and in sympathy with working-class needs included political action as a member of the Württemberg Diet. Since the latter part of the 19th century, different groups of the Pentecostal and charismatic movements have re-accepted the use of exorcistic rituals with great emphasis and—pointing to the power of the Holy Spirit—they claim the charisma of healing as one of the spiritual gifts granted the believing Christian. After the basic connection between healing of the body and healing of the soul and the psychogenic origin of many illnesses was acknowledged theologically and medically, different older churches, such as the Protestant Episcopal Church and even the Roman Catholic Church in the United States, have reinstituted healing services.

In terms of spiritual healing, one church has stood out in this respect in North America. Mary Baker Eddy (1821–1910), the founder of Christian Science, referred particularly to healing through the Spirit as her special mission. Based on her experience of a successful healing from a serious illness by Phineas Quimby, a pupil of the German hypnotist Franz Mesmer, she wrote her work *Science and Health with Key to the Scriptures* and founded the Church of Christ, Scientist. According to the instructions of its founder, Christian Science today carries out a practice of "spiritual healing" throughout the world.

*Care for the sick.* From the beginning another concern besides healing was care for the sick, an element of the earliest commandments of Christian ethics. At the Last Judgment, Christ the Judge will say to the chosen ones on his right hand: "I was sick and you visited me," and to the condemned on his left hand: "I was sick and you did not visit me." To the condemneds' surprised questions as to when they saw Christ sick and did not visit him, they will receive the answer: "As you did it not to one of the least of these, you did it not to me."

The concern for the Last Judgment

The first office created by the church in Jerusalem was the diaconate; it spread rapidly throughout the entire church. The care of the sick was carried out by the deacons and widows under the leadership of the bishop. This service was not limited to members of the Christian congregation but was directed toward the larger community, particularly in times of pestilence and plague. Eusebius included in his *Ecclesiastical History* the report that while the heathen fled the plague at Alexandria, "most of our brother-Christians showed unbounded love and loyalty" in caring for and frequently dying with the victims.

During the Middle Ages the monasteries took over the care of the sick and created a new institution, the hospital. The growing number of pilgrims to the Holy Land and the necessity of care of their numerous sick, who had fallen victim to the unfamiliar conditions of climate and life, led to knightly hospital orders, the most important of which was the Order of the Hospital of St. John of Jerusalem (later called the Knights of Malta). The service for the sick, which was carried out by the knights besides their military service for the protection of the pilgrims, was not elaborate.

In connection with the orders of mendicant friars, especially the Franciscans, civil hospital orders were formed. Even the hospital in Marburg that was founded by St. Elizabeth of Hungary (1207–31) on the territory of the Knights of the Teutonic Order was influenced by the spirit of St. Francis. Other hospitals were founded as autonomous institutions under the leadership or supervision of a bishop. The meaningful centralization of the different existing institutions became necessary with the growth of cities and was most frequently undertaken by city councils. The laity began to take over, but the spiritual and pastoral care of the patients remained a major concern.

The development of nursing orders and deaconess orders

In the realm of the Lutheran Reformation, the medieval nursing institutions were adapted to new conditions. The church constitutions in the different territories of Reformed churches stressed the duty of caring for the sick and gave suggestions for its adequate realization. The office of the deacon was supplemented by that of the deaconess; and these offices of service were considered part of the polity of the church of the New Testament. The Counter-Reformation brought a new impulse for caring for the sick in the Roman Catholic Church, insofar as special orders for nursing service were founded—*e.g.,* the Daughters of Charity, a non-enclosed congregation of women devoted to the care of the sick and the poor, by Vincent de Paul, who was a notable charismatic healer. A great number of new orders came into existence and spread the spirit and institutions of ecclesiastical nursing care throughout the world as part of Roman Catholic world missions.

In the realm of Protestantism, the Free churches led in the care of the sick. Methodists, Baptists, and Quakers all had a great share in this development, founding numerous hospitals throughout the world and supplying them with willing male and female helpers. German Lutheranism was influenced by these developments. In 1823 Amalie Sieveking developed a sisterhood analogous to the Daughters of Charity and was active in caring for the cholera victims of the great Hamburg epidemic of 1831. She was an inspiration to Theodor Fliedner, who founded the first Protestant hospital in Kaiserswerth in 1836 and created at the same time the female diaconate, an order of nurses that soon found worldwide membership and recognition. Florence Nightingale received training at Kaiserswerth, which was an important model for modern nursing schools.

Church hospitals and ecclesiastical nursing care still maintain a leading and exemplary role in the 20th century, although along with the general political and social development of the 19th century the city or communal hospital was founded and overtook the church hospital.

The founding of the Red Cross

The most impressive example of the universal spread of care for the sick was the founding of the Red Cross by the Swiss humanitarian Henri Dunant. The religious influence of Dunant's pious parental home in Geneva and the shocking impression he received on the battlefield of Solferino in June 1859 led him to work out suggestions that—after difficult negotiations with representatives of numerous states—led to the conclusion of the "Geneva convention regarding the care and treatment in wartime of the wounded military personnel." In the 20th century the activity of the Red Cross has embraced not only the victims of military actions but also peace activity, which includes aid for the sick, for the handicapped, for the elderly and children, and for the victims of all types of disasters everywhere in the world.

**Care for widows and orphans.** From the beginning the Christian congregation cared for the poor, the sick, widows, and orphans. The Letter of James says: "Religion that is pure and undefiled before God is this: to visit orphans and widows in their affliction." Widows formed a special group in the congregations and were asked to help with nursing care and other diaconic (from *diakonia,* or faith active in love and service to all) congregational tasks as long as they did not need help and care themselves.

The founding of orphanages and other institutions of mercy

The church had founded orphanages during the 4th century, and the monasteries took over this task during the Middle Ages. They also fought against the practice of abandoning unwanted children and established foundling hospitals. In this area, as in others, a secularization of church institutions took place in connection with the spreading autonomy of the cities. In the Reformed churches the establishment of orphanages was furthered systematically. In Holland almost every congregation had

its own orphanage, which was sustained through the gifts of the members.

Following the great wars of the 17th century the orphanages were reorganized pedagogically, notably by August Hermann Francke, who connected the orphanage in Glaucha, Ger., which he had founded, with a modern system of secondary schools. Francke's orphanage became a model that was frequently imitated in England and also in North America. An exemplary proponent of comprehensive Christian caring and curing for the whole person and community was the Alsatian Lutheran pastor Johann Friedrich Oberlin (1740–1826). Responsible for a remote and barren area in the Vosges Mountains, Oberlin transformed the impoverished villages into prosperous communities. He led in establishing schools, roads, bridges, banks, stores, agricultural societies (with the introduction of potato cultivation), and industries. His nursery schools were imitated in many areas through "Oberlin Societies." These efforts provided a significant contribution to the development of modern welfare, which in the 20th century is mainly the responsibility of state, communal, or humanitarian organizations but is still characterized strongly by its Christian roots. (E.W.B./C.H.L.)

**Property, poverty, and the poor.** The Christian community's relation to the questions of property, poverty, and the poor may be sketched in terms of four major perspectives, which have historically overlapped and sometimes coexisted in mutuality or contradiction. The first perspective, both chronologically and in continuing popularity, is personal charity. This was the predominant form of the church's relationship to the poor from the 1st to the 16th century. The second perspective supplements the remedial work of charity by efforts for preventive welfare through structural changes in society. This concern to remove causes of poverty was clearly expressed in the Reformation but was soon submerged in the profound sociopolitical and economic changes of the time. The third perspective is a retreat into the charity models of the earlier Christian community. Because of the overwhelming effects of the process of secularization and the human misery caused by industrialization, the key to social welfare was expressed in the Pietist maxim that social change depended upon the conversion of individuals. The fourth perspective, present in churches of the modern period, envisions systemic social change to facilitate redistribution of the world's wealth. Personal charity is not neglected, but the major goal is to change the unjust structures of world society.

The relationship of rich and poor

The early Christian community's teachings on property and poverty were marked by the tension between its expectation of support from the wealthy and its biblically rooted criticism of wealth. The solution was to place rich and poor in a symbiotic relationship oriented toward salvation. The rich supply the needs of the poor, who in turn provide the rich with the opportunity for good works and prayers for their salvation.

Augustine's doctrine of charity became the heart of Christian thought and practice. Augustine portrayed the Christian pilgrimage toward the heavenly city by analogy to a traveler's journey home. The city of God, humankind's true home, is characterized by the love of God even to the contempt of self, whereas the earthly city is characterized by the love of self even to the contempt of God. It is the goal—not the journey—that is ultimately important. The world and its goods must be used for the journey, but if they are enjoyed they direct the traveler away from God to the earth. This imagery incorporates into the heart of Christian theology the great medieval themes of pilgrimage, renunciation, alienation, and asceticism; and the biblical and early Christian suspicion of riches receives systematic theological articulation. Pride and covetousness are the major vices; humility and almsgiving are the major virtues; and poverty is endorsed as the favoured status for the Christian life.

This view did not, however, lead to a rejection of property and its importance for society. Against both Marcionite denigration of the world and Gnostic communism, respect for private property was maintained as integral to a comprehensive ethic. It was clear that without property Christians could not care for the needy. And although Gregory of Nazianzus (*c.* 330–*c.* 389) linked private property to the Fall, he understood that the abolition of private property would not cure sin. Property and wealth should be shared, not relinquished. Yet the paradox of 2 Corinthians 6:10 remained: How could a Christian be poor yet make many rich, have nothing yet possess everything? The answers given were communal property, charity to the needy, avoidance of avarice, and concentration upon heavenly treasure. In this way the early Christian community achieved an aristocratic attitude to riches. The solutions of institutionalizing poverty in priesthood and monasticism, while rationalizing poverty as poverty of the spirit and material wealth as God's provision for ministry, formed the basis for medieval care of the poor.

The medieval Christian community promoted almsgiving within a theological framework oriented to the future salvation of the individual. Although this framework was a stimulus to insightful and humane laws and actions, it did not result in the formulation of policies to deal with the major social and economic changes that accompanied the late medieval shift from rural-agricultural society to urban-commercial society.

The most influential medieval thinker on the problem of property was Thomas Aquinas. Thomas saw community of goods as rooted in natural law because it makes no distinction of possessions. The natural law of common use protects every person's access to earthly goods and requires responsibility by everyone to provide for the needs of others. Private property, on the other hand, is rooted in positive law through human reason. In history, reason leads to the conclusion that the common good is served if everyone has disposition of his own property because there is more incentive to work, goods are more carefully used, and peace is better preserved when all are satisfied with what they have. Private property exists to serve the common good; thus, superfluous property is to be distributed as alms to the needy.

The other major effort to deal with property and poverty at this time was through rational direction and administration. As cities developed into political corporations, a new element entered welfare work: an organizing citizenry. Through their town councils, citizens began to claim the authority to administer the ecclesiastical welfare work of hospitals and poor relief. The process was accelerated by the Reformers, whose theology undercut the medieval idealization of poverty. According to the Reformers, righteousness before God was by faith alone apart from human works, and salvation was perceived as the foundation of life rather than its goal. Thus, the Reformation community found it difficult to rationalize the plight of the poor as a peculiar form of blessedness, and no salvific value in either being poor or in giving alms could be identified. When the Reformers turned to poor relief and social welfare, their new theological perspectives led them to raise questions of social justice and social structures. This was institutionalized in the "common chest" sections of Protestant church legislation, which spread throughout Europe from its origin in Wittenberg. The common chest—funded by church endowments, offerings, and taxes—was the community's financial resource for providing support to the poor, orphans, aged, unemployed, and underemployed through subsidies, low-interest loans, and gifts. The attempt to resolve social problems in the cities was a constitutive part of the early Reformation.

Reformation views of poverty

In the following centuries the heirs of Luther and Calvin, although producing noteworthy examples of compassion and charity for the poor, nevertheless lost their "fathers' " vision of a social ethic that was preventive as well as therapeutic. Like their Roman Catholic counterparts, the Protestants made noteworthy efforts to serve the poor but ignored the root causes of poverty.

In the 18th and 19th centuries the social institutions of Pietism, the Inner Mission, and European revival movements inspired social concern for the masses of people pauperized and proletarianized by industrialism. The Methodists in England undertook adult education, schooling, reform of prisons, abolition of slavery, and aid to alcoholics. Famous missions arose in Basel, London, and Paris. The Young Men's Christian Association (YMCA;

1844), Young Women's Christian Association (YWCA; 1855), and the Salvation Army (1865) were only some of the numerous charitable institutions and organizations created to alleviate modern ills. In 1848, Johann Wichern, founder of the Inner Mission, proclaimed that "love no less than faith is the church's indispensable mark."

Yet this Christian social concern hardly was aware of and rarely attempted to expose the origins of the social ills it strove to remedy. Wichern himself was aware that poverty is social, not natural, but his orientation, like that of others, was toward renewing society through evangelization. This attitude—that society is changed by changing the hearts of individuals—is still prevalent.

In recent years, however, the Christian community, especially in its ecumenical organizations, has begun to analyze the social problems of property and poverty from the standpoint of justice and the perspectives of the poor and oppressed. In 1970 the World Council of Churches (WCC) established the Commission for the Churches' Participation in Development (CCPD). Initially involved in development programs and the provision of technical services, the CCPD focus has shifted to the psychological and political character of the symbiosis of development and underdevelopment. This focus was endorsed at the 1975 WCC Assembly at Nairobi, Kenya, as "a liberating process aimed at justice, self-reliance and economic growth." Other church bodies, such as the Lutheran World Federation and the World Alliance of Reformed Churches, share this perspective. Emergency relief and development projects—the modern equivalents of charity—have not ceased, but there is growing realization, due to the increasing participation of so-called Third World churches, that the biblical themes of justice and liberation entail the creation of social structures to enhance human life, economic structures for just distribution of goods, and political structures to promote participation and minimize dependence. The present WCC paradigm for this mission is the church in solidarity with the poor.

**Pastoral care.** Pastoral care has always been of special importance in the Christian community. The biographies of the great charismatic ministers, beginning with the Fathers of the Eastern Church and the Western Church, testify to surprising variations of this pastoral care. The principal interest of pastoral care—whether exercised by clergy or laity—is the personal welfare of persons who are hurt, troubled, alienated, or confused within the context of ultimate concerns and meanings. The historical expressions of pastoral care have focused on the predominant—but not exclusive—expressions of ultimate concern characteristic of the period in question. According to Paul Tillich, in *The Courage to Be* (1952), these concerns may be described in terms of the anxieties of death (early church), guilt (Middle Ages), and meaninglessness (modern period). Thus, Ignatius addressed the terror of death when he termed the sacrament "the medicine of immortality"; Luther responded to the conscience tortured by guilt and uncertainty by proclaiming the free forgiveness of sin by grace alone, apart from human accomplishment; and the modern Christian community has utilized the insights of psychology and psychiatry in developing pastoral counseling and therapy responsive to modern anxieties. Fundamentally, however, pastoral care has always attempted to respond to the totality of human needs in every age in consonance with the words of Jesus Christ: "I was hungry and you gave me food, I was thirsty and you gave me drink, I was a stranger and you welcomed me, I was naked and you clothed me, I was sick and you visited me, I was in prison and you came to me" (Matthew 25:35–36).

The purpose of pastoral care

The first influential contribution to pastoral care after the New Testament was by Pope Gregory I the Great. His *Pastoral Care,* written after he became bishop of Rome in 590, was so influential that it became customary to present it to new bishops upon their ordination. This textbook of the medieval episcopate emphasized the role of the pastor as shepherd of souls.

The medieval institutionalization of pastoral care in the sacrament of penance led to certain deficits in practice: the exclusion of the laity by emphasis upon the central role of the priest and the distortion of its original spiritual purposes of prayer, repentance, and forgiveness of sins by the introduction of paid indulgences. The indulgence abuse sparked the Reformation critique of the sacrament of penance. This in turn led to the Reformers' emphasis upon lay as well as clerical responsibility for pastoral care as expressed in their teaching of "the priesthood of all believers." The Reformation insistence upon justification by grace alone shifted the burden of proof for salvation from human accomplishment to divine promise. By "letting God be God," the Reformers claimed that persons were free to be human. This shift of theological focus, from an otherworldly achievement to a this-worldly trust in God, facilitated a renewed holistic awareness of human needs and pointed the way for the Christian community's appreciation of the benefits available in modern medicine and therapy. (C.H.L.)

## CHURCH AND MINORITIES

The tendency to develop an identifiable Christian culture is apparent even where Christian minorities live in a non-Christian environment—*i.e.,* in an environment the life of which has been shaped and is characterized by a non-Christian religion. This is the case with most Christian churches in Asia and Africa.

In some countries Christian minorities have had to struggle for their existence and recognition, and there are cases of persecutions of Christians. On the other hand, in some cases the situation of Christian minorities is ideally suited to demonstrate to outsiders the peculiar style of life of a Christian culture. This is particularly advantageous for the church within a caste state, in which the church itself has developed into a caste, with special extrinsic characteristics in clothing and customs. An example of this phenomenon is the Mar Thoma Church of South India.

The problem of the coexistence of racially different Christian cultures

A special problem presents itself through the coexistence of racially different Christian cultures in racially mixed states. The influence of the Christian black churches, especially of Baptist denominations, has been thoroughly imprinted upon the culture of North American blacks. The churches themselves were founded through the missionary work of white Baptist churches but became independent of their mother churches or were established as autonomous churches within the framework of the Baptist denomination. A similar situation exists in South Africa, where white congregations and separate black congregations have been established within the white mission churches; independent messianic black churches have appeared outside the older organized congregations. In the 20th century much tension exists in this area.

On the one hand, the Christian Church has from the beginning urged the overcoming of racism. In the early church, racism was unknown; the Jewish synagogues allowed black proselytes. The first Jewish proselyte mentioned in the Acts of the Apostles was a governmental administrator from Ethiopia, who was baptized by the Apostle Philip. The early congregations in Alexandria included many Ethiopians and blacks. Among the evangelizing churches, the Portuguese Catholic mission in principle did not recognize differences between races—whoever was baptized became a "human being" and became a member not only of the Christian congregation but also of the Christian society and was allowed to marry another Christian of any race. In contrast to this practice, the Catholic mission of the Spaniards introduced the separation of races under the term *casticismo* (purity of the Castilian heritage) in the American mission regions and sometimes restricted marriage between Castilian Spanish immigrants and native Christians. Like the Portuguese in Africa and Brazil, the French Catholic mission in Canada and in the regions around the Great Lakes in North America did not prohibit marriage of whites with Indians but tolerated and even encouraged it during the 17th and 18th centuries.

Problems of racial integration and segregation

Consequently, the Christian churches led in endeavours for racial integration, with the exception of those churches that maintained racial segregation from the beginning, in deference to theological arguments deduced from the "order of the creation" and "predestination." The latter was the case in some Reformed churches of the United States and of South Africa. On the other hand, the ideologically

and politically founded racial theory has been introduced into black churches in recent times. The demand for a black theology with a black Christ in its centre has been made and, just as much as a theologically and ideologically founded racial theory on the part of whites, aggravates the specifically Christian task of racial integration within the church.

The promise of recent liberation theologies such as black theology, Latin-American theology, and feminist theology is that of expanding awareness of the history and praxis of Christianity beyond the history of doctrines, the ideas of the elite, and the institutions that convey these ideas. Such reflection—which arises out of lived situations—reveals roles of the poor, the oppressed, and women that have too often been ignored and suppressed. These new orientations serve the church and the world not only by recalling hitherto unnoticed aspects of the past but also by strengthening peoples' awareness of their own causes.

## CHURCH AND FAMILY

The Christian understanding of sexuality, marriage, and family has been strongly influenced by the Old Testament view of marriage as an institution primarily concerned with the establishment of a family, rather than sustaining the individual happiness of the marriage partners. Until the Reformation the patriarchal family structure not only had been preserved but also had been defended from all attacks by sectarian groups. In spite of this, a transformation occurred from the early days of Christianity.

This transformation is evident in the New Testament departure from the Hellenistic understanding of love. The classical understanding of love, expressed in the Platonic concept of eros, was opposed in the Christian community by the biblical understanding of love, agape. Although erotic love has frequently been understood primarily as sexual desire and passion, its classical religious and philosophical meaning was the idealistic desire to acquire the highest spiritual and intellectual good. The early Christian perception of eros as the most sublime form of egocentricity and self-assertion, the drive to acquire the divine itself, is reflected in the fact that the Greek New Testament does not use the word *erōs* but rather the relatively rare word *agapē*. *Agapē* was translated into Latin as *cāritās* and thus appears in English as "charity" and, later, "love." The Christian concept of love understood human mutuality and reciprocity within the context of God's self-giving love, which creates value in the person loved. "We love, because he first loved us. If any one says, 'I love God,' and hates his brother, he is a liar; for he who does not love his brother whom he has seen, cannot love God whom he has not seen. And this commandment we have from him, that he who loves God should love his brother also" (1 John 4:19–21). Love is presented as the greatest of the virtues (1 Corinthians 13:13) as well as a commandment. The Christian community understood faith active in love primarily in terms of voluntary obedience rather than emotion and applied this understanding to every aspect of life, including sexuality, marriage, and family.

**The tendency to spiritualize and individualize marriage.** Christianity has contributed to a spiritualization of marriage and family life, to a personal deepening of the relations between marriage partners and between parents and children, as well as between heads of households and domestic servants in large families. Marriage can be called the most intimate form in which the fellowship of believers in Christ is realized. In the early church, children were included in this fellowship. They were baptized when their parents were baptized, took part in the worship life of the congregation, and received Holy Communion with their parents. The Eastern Orthodox Church still practices as part of the eucharistic rite Jesus' teaching, "Let the children come to me, and do not hinder them." During the first decades of the church, congregational meetings took place in the homes of Christian families. The family became the archetype of the church. Paul called the members of his congregation in Ephesus "members of the household of God" (Ephesians 2:19).

Deepening of the relations between marriage partners

In the early church the Christian foundation of marriage—in the participation of Christians in the body of Christ—postulated a generous interpretation of the fellowship between a Christian and a pagan marriage partner: the pagan one is saved with the Christian one "for the unbelieving husband is consecrated through his wife, and the unbelieving wife is consecrated through her husband"; even the children from such a marriage in which at least one partner belongs to the body of Christ "are holy" (1 Corinthians 7:14). If the pagan partner, however, does not want to sustain the marriage relationship with a Christian partner under any circumstances, the Christian partner should grant him a divorce.

New Testament views on marriage

Jesus himself based his parables of the Kingdom of God on the idea of love between a bride and groom and frequently used parables of a wedding that describe the messianic meal as a wedding feast. In Revelation the glorious finale of salvation history is depicted as the wedding of the Lamb with the bride, as the beginning of the meal of the chosen ones with the Messiah–Son of man (Revelation 19:9: "Blessed are those who are invited to the marriage supper of the Lamb"). The wedding character of the eucharistic meal is also expressed in the liturgy of the early church. It is deepened through the specifically Christian belief that understands the word of the creation story in Genesis "and they become one flesh" as indicative of the oneness of Christ, the head, with the congregation as his body. With this in mind the Christian demand of monogamy becomes understandable.

In the so-called ethical lists in the Letter of Paul to the Colossians and in 1 Peter, Christian marriage is distinguished from the marriage practices of its pagan environment by its stricter ethical demands. The rules concern the mutual relationship of the marriage partners, fidelity, as well as attitudes toward children and slaves of the house.

The position of women

Christianity did not bring a revolutionary social change to the position of women, but it made possible a new position in the family and congregation. In the world of the early church, women were held in very low esteem, and this was the basis for divorce practices that put women practically at men's complete disposal. With the prohibition of divorce, Jesus himself did away with this low estimation of women. The decisive turning point came in connection with the understanding of Christ and of the Holy Spirit. Even the Jewish view of the patriarchal position of man was substituted by Paul with a new spiritual interpretation of marriage. "There is neither male nor female; for you are all one in Christ Jesus" (Galatians 3:28). In fulfillment of the prophecy in Joel 3:1, the Holy Spirit was poured out over the female disciples of Jesus, as well.

This created a complete change in the position of women in the congregation: in the synagogue the women were inactive participants in the worship service and sat veiled on the women's side, usually separated from the rest by an opaque lattice. In the Christian congregation, however, women appeared as members with full rights, who used their charismatic gifts within the congregation. In the letters of Paul, women are mentioned as Christians of full value. Paul addresses Prisca (Priscilla) in Romans 16:3 as his fellow worker. The four daughters of Philip were active as prophets in the congregation. Peter, in a sermon on Pentecost, spoke about men and women as recipients of the gifts of the Holy Spirit: "Your sons and your daughters shall prophesy" (Acts 2:17). Pagan critics of the church, such as Porphyry (*c.* 234–*c.* 305), maintained that the church was ruled by women. During the periods of Christian persecution, women as well as men showed great courage in their suffering. The fact that they were spontaneously honoured as martyrs demonstrates their well-known active roles in the congregations. In this, representatives of patriarchal, rabbinic, and synagogic traditions within the Christian Church saw a danger to congregational constitutions. Paul, on the one hand, included women in his instruction, "Do not quench the Spirit" (1 Thessalonians 5:19), but, on the other hand, carried over the rule of the synagogue into the Christian congregation that "women should keep silence in the churches" (1 Corinthians 14:34). In the 20th century the Roman Catholic Church still refuses to ordain women as priests.

**The tendency toward asceticism.** The proponents of an ascetic theology demanded exclusiveness of devotion

The demand for celibacy

by faithful Christians to Christ and deduced from it the demand of celibacy. This is found in arguments for the monastic life and in the Roman Catholic view of the priesthood. The radical-ascetic interpretation stands in constant tension with the positive understanding of Christian marriage. This tension has led to seemingly unsolvable conflicts and to numerous compromises in the history of Christianity. Without doubt, from the beginning a strong ascetic tendency dominant in Christianity was emphatically directed against the oversexualization of the Hellenistic culture, against the decay of marital life in the Hellenistic world, against the spreading of pederasty and its social recognition and open institutionalization, against cultic and non-cultic prostitution, and against the more or less tolerated sodomy that was excused with pagan mythology.

In the light of the beginning Kingdom of God, marriage was understood as an order of the old passing eon, which would not exist in the approaching new age. The risen ones will "neither marry nor are given in marriage, but are like angels in heaven" (Mark 12:25). Similarly, Paul understood marriage in the light of the coming Kingdom of God: "The appointed time has grown very short; from now on, let those who have wives live as though they had none . . . for the form of this world is passing away" (1 Corinthians 7:29–31). In view of the proximity of the Kingdom of God, it was considered not worthwhile to marry; and marriage was seen to involve unnecessary troubles: "I want you to be free from anxieties" (1 Corinthians 7:32). Therefore, the unmarried, the widowers, and widows "do better" if they do not marry, if they remain single. But according to this point of view marriage was recommended to those who "cannot exercise self-control . . . for it is better to marry than to be aflame with passion" (1 Corinthians 7:9). With the waning of the eschatological expectation that formed the original context for the Pauline views on marriage, his writings were interpreted ascetically. While these texts have been used alone in the course of church history, however, they do not stand alone in the New Testament, which also portrays marriage feasts as joyous occasions and sexual intercourse between spouses as good and holy (Ephesians 5:25–33).

The demonization of sex

A demonization of sex in general occurred in dualistic Gnostic movements. This was particularly apparent in the ascetic branches of Gnosticism and especially in Manichaeism (an Iranian dualistic religion). The conscious renunciation by Christians of the customs of their oversexualized pagan environment supported these tendencies. Their motives are apparent in the biographies and letters of the great ascetics, such as Anthony and Jerome. Within the Roman Catholic Church the tension between the Christian high esteem and the ascetic devaluation of marriage led to a constantly challenged compromise: celibacy was demanded not only of ascetics and monks but also more and more of members of the clergy as a duty of their office.

The Reformation rejected clerical celibacy because it removed men and women from service to the neighbour, contravened the divine order of marriage and the family, and denied the goodness of sexuality. Luther viewed marriage as not merely the legitimation of sexual fulfillment but as above all the context for creating a new awareness of human community through the mutuality and companionship of spouses and family. The demand that priests and monks observe celibacy was not fully accepted in the East. The early church, and following it the Eastern Orthodox Church, decided on a compromise at the Council of Nicaea (325): the lower clergy, including the archimandrite, would be allowed to enter matrimony before receiving the higher degrees of ordination; of the higher clergy—*i.e.,* bishops—celibacy would be demanded. This solution has saved the Eastern Orthodox from a permanent fight for the demand of celibacy for all clergymen, but it has resulted in a grave separation of the clergy into a white (celibate) and a black (married) clergy, which led to severe disagreements in times of crisis within Orthodoxy.

Problem of birth control

The early Christian community's attitude to birth control was formed partly in reaction against secular attitudes of indifference to sexual exploitation and infanticide and partly against the Gnostic denigration of the material world and consequent hostility to procreation. In upholding its faith in the goodness of creation, sexuality, marriage, and family, the early church was also influenced by the prevalent Stoic philosophy, which emphasized procreation as the rational purpose in marriage.

The question of birth control entered a new phase through the invention and mass distribution of technical contraceptive devices on the one hand and through the appearance of a new attitude toward sexual questions on the other. In this situation an obvious differentiation of interpretation developed within Christianity: with a few exceptions—*e.g.,* the Mormons—the Protestant churches accepted birth control in terms of a Christian social ethic. In contrast, the Roman Catholic Church, in the encyclical of Pius XI *Casti Connubii* (1930) and in the encyclical of Paul VI *Humanae Vitae* (1968), completely rejected any kind of contraception. Modern economic and population concerns in connection with improved medical care and social and technological progress have once again confronted the Christian community with the issue of contraception.

## CHURCH AND THE INDIVIDUAL

Christianity received the main commandment of its ethic from the Old Testament: "You shall love your neighbour as yourself" (Leviticus 19:18), but Jesus filled this commandment with a new, twofold meaning. First, he closely connected the commandment "love your neighbour" with the commandment to love God. In the dispute with the scribes described in Matthew, chapter 22, he quoted the commandment of Deuteronomy 6:5, "You shall love the Lord your God with all your heart, and with all your soul, and with all your might." He spoke of the commandment of love for neighbour, however, as being equal to it. With that he lifted it to the same level as the highest and greatest commandment, the commandment to love God. In the Gospel According to Luke, both commandments have grown together into one single pronouncement with the addition: "Do this, and you will live." Second, the commandment received a new content in view of God and in view of the neighbour through the relationship of the believer with Christ. Love of God and love of the neighbour is possible because the Son proclaims the Gospel of the Father and brings to it reality and credibility through his life, death, and Resurrection. Based on this connection of the Christian commandment of love with the understanding of Christ's person and work, the demand of love for the neighbour appears as a new commandment: "A new commandment I give to you, that you love one another; even as I have loved you, that you also love one another" (John 13:34). The love for each other is supposed to characterize the disciples: "By this all men will know that you are my disciples, if you have love for one another" (John 13:35).

Love as a new commandment

This is an ethic that does not base its norms on social, biologic, psychological, physiological, intellectual, or educational differences and levels but on an understanding and treatment of human beings as created in the image of God. Furthermore, the ethic does not deal with humanity in an abstract sense but with the actual neighbour. The Christian ethic understands the individual always as a neighbour in Christ.

The new element of the Christian ethic is the founding of the individual ethic in a corporate ethic, in the understanding of the fellowship of Christians as the body of Christ. The individual believer is not understood as a separate individual who has found a new spiritual and moral relationship with God but as a "living stone" (1 Peter 2:4), as a living cell in the body of Christ in which the powers of the Kingdom of God are already working.

The realization of Christian love leads to the peculiar exchange of gifts and suffering, of exaltation and humiliations, of defeat and victory; the individual is able through personal sacrifice and suffering to contribute to the development of the whole. In this basic idea of the fellowship of believers as the body of Christ, all forms of ecclesiastical, political, and social communities of Christianity are founded. It also has influenced numerous secularized

forms of Christian society, even among those that have forgotten or denied their Christian origins.

The question as to who is one's neighbour

From the beginning, the commandment contains a certain tension concerning the answer to the question: Does it refer only to "the disciples," that is, fellow Christians, or to "all"? The practice of love of neighbour within the inner circle of the disciples was a conspicuous characteristic of the young church. Pagans said: "Look, how they love each other" (Justin). Christian congregations and, above all, small fellowships and sects have stood out throughout the centuries because of the fact that within their communities love of the neighbour was highly developed in the form of personal pastoral care, social welfare, and help in all situations of life.

The Christian commandment of love, however, has never been limited to fellow Christians. On the contrary, the new factor in the Christian ethic was that it crossed all social and religious barriers and saw a neighbour in every suffering human being. Characteristically, Jesus himself explicated his understanding of the practical implications of the commandment of love in the parable of the Good Samaritan, a non-Jew who followed the commandment of love and helped a person in need whom the believing wanderers—a priest and a Levite—had chosen to ignore (Luke 10:29–37). A demand in the Letter of James, that the "royal law" of neighbourly love has to be fulfilled without "partiality" (James 2:9), points to its universal validity.

Love of one's enemies

The universalism of the Christian command to love is most strongly expressed in its demand to love one's enemies. Jesus himself emphasized this with these words: "Love your enemies and pray for those who persecute you, so that you may be sons of your Father who is in heaven; for he makes his sun rise on the evil and on the good, and sends rain on the just and on the unjust" (Matthew 5:44–45). According to this understanding, love of the enemy is the immediate emission of God's love, which includes God's friends and God's enemies.

The Reformation revitalized a personal sense of Christian responsibility by anchoring it in the free forgiveness of sins. Luther summarized this in "The Freedom of a Christian Man" (1520): "A Christian is a perfectly free lord of all, subject to none. A Christian is a perfectly dutiful servant of all, subject to all." The second sentence expressed the theme of Christian vocation developed by Luther and Calvin. While the medieval church understood vocation in terms of the specific religious calling of priesthood and monastery, the Reformers expanded the concept of vocation to all Christians and to everyday responsibility for the neighbour and for the world. The Reformers emphasized that Christian service is not limited to a narrow religious sphere of life but finds means to help others in the everyday relationships of family, marriage, work, and politics.

Personal and social ethics

Later Protestantism under the influence of Pietism and Romanticism restricted the social and communal orientation of the Reformers to a more individualistic orientation. This met, however, with an energetic counterattack from the circles of the Free churches (*e.g.,* Baptists and Methodists) who supported the social task of Christian ethic (mainly through the Social Gospel of the American theologian Walter Rauschenbusch, who attempted to change social institutions and bring about a Kingdom of God), which spread through the whole church, penetrating the area of Christian mission. Love rooted in faith has continued to play an important role in the 20th century in the struggle between Christianity and all ideologies, such as Fascism, Communism, and jingoistic nationalisms.

(E.W.B./C.H.L.)

## Christian missions

In the late 20th century about one-third of the world's people claimed the Christian faith. Christians thus constituted the world's largest religious community and embraced remarkable diversity, with churches in every nation. Christianity's demographic and dynamic centre had shifted from its Western base to Latin America, Africa, Asia, and the Pacific region, where more than half the world's Christians lived. This trend steadily accelerated as the church declined in Europe. The tangibly real universal church represented a new phenomenon in the history of religions. This was the fruit of mission.

### BIBLICAL FOUNDATIONS

The word mission (Latin: *missiō*), as a translation of the Greek *apostolē,* "a sending," appears only once in the English New Testament (Galatians 2:8). An apostle (*apostolos*) is one commissioned and sent to fulfill a special purpose. The roots of mission, Christians have believed, lie in God's active outreach to humanity in history—as a call to those able to fulfill the divine purpose, among them Abraham, Moses, Jonah, and Paul. The New Testament designated Jesus as God's apostle (Hebrews 3:1). Jesus' prayer in the Gospel According to John includes the words "As thou didst send me into the world, so I have sent them into the world. . . . [I pray also] for those who believe in me through their word, that they may all be one . . . so that the world may believe that thou hast sent me" (John 17:18, 20–21).

The ground for mission appeared early in the Old Testament in God's concern for all nations (Genesis, chapters 10 [the "Table of Nations"] and 11) and the calling of Abraham—implicitly of Israel (Genesis 12:1–3). The Jews acknowledged God's sovereignty over all the world's peoples but believed God had chosen Israel to be the sign to all nations of the divine will and purpose (compare the "Covenant on Sinai," Exodus 19:5–6). Echoing throughout the Old Testament, this theme found its clearest voice in Isaiah: "I have given you as a covenant to the people, a light to the nations" (42:6); and in God's universal task for Israel as servant to the nations: "I will give you as a light to the nations, that my salvation may reach to the end of the earth" (49:6).

The Maccabean wars filled the Jews with fervent hope for a powerful messiah to bring political triumph to a suffering Israel. Jesus Christ, professed by his followers as Messiah in the role of suffering servant presented in Isaiah (52:13–53:12), found widespread rejection among his people. Yet those from Israel who confessed faith in him as Messiah and Lord saw in Christ's Incarnation, death, and Resurrection God's decisive entry into history—an act in continuity with God's incursions in Israel's past. In that reality they and their successors viewed the Old and New Testaments (or "Covenants") as inseparably united and mutually interdependent. The church was born and grew as the covenanted instrument of and witness to God's mission (*missio Dei*), the human agency of God's outreach to all the peoples of the world.

The Great Commission

The "Great Commission" of Jesus declares: "Go therefore and make disciples of all nations, baptizing them in the name of the Father and of the Son and of the Holy Spirit, teaching them to observe all that I have commanded you; and lo, I am with you always, to the close of the age" (Matthew 28:19–20; compare Mark 16:15, Luke 24:47, John 20:21–22, and Acts 1:8). Not an isolated command, it re-expressed, in Christian perspective, the obedience of a servant in universal witness to the mission of God as declared in the Old and New Covenants.

### THE HISTORY OF CHRISTIAN MISSIONS

The Christian mission, the church, and Christianity—each distinguishable, but inseparably related—have experienced across 20 centuries of world history four major transitions.

**First transition, to AD 500.** Born on Jewish soil but quickly emerging from Palestine to cover the rim of the Mediterranean world, the new missionary faith made its first major transition. The Apostle Paul became the missionary to the Gentile world. With help from Barnabas and a local network of coworkers, many of them women, he evangelized Asia Minor and southern Greece and eventually reached Rome. When Rome destroyed Jerusalem in AD 70, Antioch became Christianity's centre in the Eastern Empire, and mission became one of Gentiles to Gentiles. Thus began the transition.

Dominated politically by the Roman Empire, the new religion benefited from one unifying factor in the Greco-Roman world: common, or koine, Greek provided its lingua franca. Alexandrian Jews had already translated (250 BCE) the Hebrew Bible into koine Greek for dis-

persed Greek-speaking Jews. The New Testament writers also wrote in koine Greek. In that largely literate empire early Christians used and widely distributed the Hebrew Scriptures.

Channels of growth

Several factors brought growth to the faith. From the beginning laypeople—women and men—conducted the largest part of mission. Congregations grew in homes used as churches. Although it was owned by the husband, inside the house the wife was its mistress, and thousands of women opened their homes to newly forming churches. Most evangelization occurred in the daily routine as men and women shared their faith with others. Christianity's monotheism, morality, assurance of eternal life with God, and ancient Scriptures attracted many to the faith.

From the empire-wide occasions of emperor worship, Rome had exempted only the Jews. Christians also refused to engage in emperor worship. Rome declared their faith an illegal religion, and persecutions ensued. In the persecutions so many Christians had borne powerful witness (Greek: *martyria*) that the word martyr quickly evolved into its current meaning. Christian faith—not least that of young women such as Blandina, Perpetua, and Felicity—had made an impact, and many who beheld that witness became Christian. In 313 when the new emperor, Constantine, declared the persecutions ended, Christians probably constituted 10 percent of the empire's population.

Christians daily encountered members of other religions—the mysteries, Gnosticism, and philosophical cults. In the 2nd and 3rd centuries external and internal pressures drove the young church to strengthen itself through creating a structured ministry, formulating beliefs in creeds, and producing a canon of Scripture. That process transformed a movement into a young religious institution. The major thrust of the pre-Constantinian church-mission sprang from the conviction that Christians and congregations were fulfilling a mission and ministry begun in Jesus Christ. Baptism provided induction into the vibrant company of "God's own people" (1 Peter 2:9–10).

By 315 many who saw advantage in belonging to Constantine's new imperial faith poured into the churches. The result was striking: small congregations of convinced Christians serving God's outreach in the world became large churches with many nominal members whose instruction and needs had to be met within the new churches. As multitudes entered the churches, the need for outreach to others was much reduced, and most churches shifted from an outward thrust to an inward focus upon themselves. Mission and service became the province of priests, deacons, and, increasingly, monks. This Constantinian inversion helped shape the churches of Christendom.

Mission outside the empire

At the same time, mission beyond the frontiers of the empire continued. Ulfilas (*c.* 311–*c.* 382), Arian apostle to the Goths, translated the Bible into their tongue. Martin of Tours (*c.* 316–397) served in Gaul, and Patrick (*c.* 389–*c.* 461) laboured in Ireland. In Malabar, South India, a church of ancient tradition, demonstrably present since the 3rd century, held the Apostle Thomas to be its founder. Frumentius (d. *c.* 380) from Tyre evangelized in Ethiopia and became the first patriarch of its church. In the 5th century Nestorians pushed into Central Asia and began a mission that eventually reached the capital of China.

In 410 Rome fell to the barbarians, and by 476 the entire Western Roman Empire had collapsed. In the eyes of many members of the still-pagan Roman nobility, the rise of the Christian faith in the Mediterranean world had caused the empire's downfall. Yet where Constantine had built his capital, Constantinople—the "second Rome"—the Eastern Empire continued.

In its first 500 years Christianity achieved remarkable missionary and theological acculturation. Through the first four ecumenical councils (325–451), and in the Nicene Creed (on the Trinity) and Definition of Chalcedon (on Christology), the church had stated its faith with meaning for the Greek and Latin worlds.

By the close of the period Jerome's Latin translation of the Bible, the Vulgate, had appeared. Church and state already were locked in uneasy embrace. The first great transition of the Christian mission—from Judaic Palestine to the Mediterranean world—had ended.

**Second transition, to AD 1500.** Rome's urban and literate world quickly disappeared under the barbarians' westward onslaught. These rough conquerors filled Europe's rural lands; however, they recognized in missionary monks the bearers of a new faith and preservers of a higher civilization. The monks instructed them in the faith and in statecraft. The mission thrust of these monks contrasted sharply with that of the tiny persecuted church in the first three centuries. Then, except for the conversions of the city-state of Edessa, in AD 200, and Armenia, declared a Christian nation in AD 300, people joined the new faith individually. In this second transition whole peoples followed their sovereigns into the new faith.

Christianity expanded in the Byzantine Empire, most notably in Russia, but it experienced a widening breach, and a split of the Eastern and Western churches occurred in 1054. Yet the major result of this 1,000-year mission was the creation of European civilization. Its emergence marked the second great transition of the faith.

*Western mission.* The medieval mission began in 496 with the baptism of Clovis, king of the Franks, and his soldiers. Baptized by a Catholic bishop rather than an Arian one (through the influence of Clovis' Catholic wife), they helped to turn the tide against the Arians.

Celtic missions

Irish Celtic Christianity differed from that on the Continent. It was organized into communalized groups under an abbot and nurtured intense missionary conviction and outreach. It did not recognize Rome's authority. The abbot Columba (*c.* 521–597) built a monastery on Iona, off Scotland's western coast, as a base for mission to Scotland and northern England. From it Aidan (d. 651) traveled to Lindisfarne, off England's northern coast, where he and a successor, Cuthbert (634/635–687), led in evangelizing in Northumbria. Moving southward, the Celtic monks might have evangelized all of Britain, but midway they met Roman missionaries. Other Celtic *peregrini,* or "wanderers," evangelized on the Continent.

*Papal mission.* Pope Gregory I the Great (reigned 590–604), who possessed the mind of both a statesman and a theologian, operated as a Roman emperor and greatly magnified papal power and temporal involvement. In 596 he launched, through Augustine of Canterbury, a mission to England based on gradualism and accommodation—the first papally sponsored mission. For the next 1,000 years Roman missions operated with the pope's direction, the king's support, and the monks' services.

Augustine's missionaries reached England's southern coast in 597. King Aethelberht of Kent and his wife, Bertha, a Christian, enabled them to make their base at Canterbury. Within the year the King and 10,000 subjects had received baptism. Roman missionaries moving northward met the Celts, and at the Synod of Whitby in 664 the Celts accepted Roman jurisdiction and patterns.

Inspired by Irish missionary enthusiasm, the English Christians began a 500-year mission across northern Europe and finally into Scandinavia. Outstanding in this effort were Willibrord (658?–739), "Apostle to the Frisians" (Friesland, Holland, and Belgium), and Wynfrid, renamed Boniface (*c.* 675–754), one of the greatest of all Roman missionaries. In central and southern Germany Boniface established Benedictine monasteries for evangelization. With full papal trust and Carolingian support he strengthened and reformed the Frankish church.

Boniface also saw the need for women in mission. From England he recruited Lioba (d. 782) and entrusted her with developing Benedictine monasteries for women. Despite her outstanding and unique achievements, with her death that movement ended, and Roman Catholic women reentered mission service only in the 19th century. But the Christian wives of pagan kings, who led their husbands into the faith and through them hastened the Christianizing of whole peoples, also contributed to its spread.

In Rome on Christmas Day, AD 800, Pope Leo III crowned Charlemagne (d. 814) Holy Roman emperor. Leo thus demonstrated the primacy of papal power over temporal rulers and symbolized the growing gulf between the Eastern and Western churches. Charlemagne's missionary zeal and political goals fused. Saxons in his territories faced a choice: become Christian or die.

Mission to eastern Europe and Scandinavia

From the Holy Roman Empire, Catholic outreach into Bohemia took root under King Wenceslas I (*c.* 907–929), with evangelization complete by about AD 1000. In Poland, Mieszko I, under the influence of his wife, accepted baptism in 966 or 967. His reign saw the beginning of the evangelization of the country, which continued under his able son, Boleslav. In 955 the Holy Roman emperor Otto I defeated the Magyars and brought them to Christian faith. Later, the country's first king, Stephen (reigned 1000–38), made Hungary a Christian land.

Early attempts at evangelization in Denmark and Sweden were made by a German monk, Ansgar (801–865). Canute (d. 1035), Danish king of England, of Denmark, and of Norway, was probably raised as a Christian and determined that Denmark should become a Christian country. The archbishop of Canterbury consecrated bishops for him, and he saw his goal realized before he died. Olaf I Tryggvason (reigned 995–*c.* 1000) was baptized by a Christian hermit, returned to Norway and was accepted as king, and sought to make his realm Christian—a task completed by King Olaf II Haraldsson (reigned 1016–30), later St. Olaf. Olaf I also presented Christianity to a receptive Iceland. Leif Eriksson took the faith to Greenland's Viking settlers, who quickly accepted it. After several efforts Sweden became Christian during the reign of Sverker (*c.* 1130–56). Sweden's Eric IX controlled Finland and in 1155 required the Finns to be baptized, but only in 1291, with the appointment of Magnus, the first Finnish bishop, was evangelization completed.

*Eastern and Nestorian missions.* Removal of the empire's capital from Rome to Constantinople, the "second Rome," in 330 greatly strengthened the temporal power of the bishop of Rome. In the Byzantine Empire the patriarch of Constantinople remained under the political control of the Christian emperor. Cultural, political, philosophical, and theological differences strained relations between the two cities. Rome demanded Latin as the one ecclesiastical language, but Constantinople encouraged national languages for the liturgy and emphasized translation of the Scriptures. In 1054 leaders of the two bodies excommunicated each other.

One reflection of growing difficulties lay in counterclaims to pursue mission in and hold the allegiance of border areas between the two jurisdictions. Rostislav of Great Moravia sought help from the Emperor, who (presumably through the Patriarch) in about 862 sent two brothers, Constantine (later called Cyril; *c.* 827–869) and Methodius (*c.* 825–884), from Constantinople to Moravia. They provided Scriptures and liturgy in the mother tongue of each people evangelized. They also trained others in their methods—a major factor in winning Bulgaria.

Constantinople's greatest mission outreach was to areas that later became Russia. In the 10th century the Scandinavian Rus controlled the areas around Kiev. Undoubtedly influenced by his Christian grandmother Olga and by a proposed marriage alliance with the Byzantine imperial family, Vladimir I (*c.* 956–1015) of Kiev, from among several options, chose the Byzantine rite. Baptized in 988, he led the Kievans to Christianity. His son Yaroslav encouraged translations and built monasteries.

From 1240, and continuing for 200 years, the Mongol Golden Horde was suzerain over Russia but generally allowed freedom to the church. For Russians the church proved to be the one means through which they could express national unity. They moved the metropolitanate from Kiev to Moscow, and their church became and remained the largest of the Orthodox bodies, protector and leader for the others. In 1453 Constantinople fell to the Ottoman Turks. Moscow became "the third Rome" and accepted for itself the mystique, dynamism, and messianic destiny of the first Rome—a reality essential to understanding Russian Orthodoxy and nationalism.

Early missions to China

East of the Euphrates River, Nestorians and Jacobites maintained headquarters in Persia for eastern outreach. The more numerous Nestorians developed a far-flung mission network throughout Central Asia. The Persian bishop A-lo-pen reached China's capital, Ch'ang-an (modern Sian), in 635 and founded monasteries to spread the Christian faith. By the end of the T'ang dynasty (618–907), however, the Nestorian community had disappeared.

In 1289 the Pope—responding to a request made 20 years earlier by Kublai Khan for 100 Christian scholars to be brought by the Polo brothers—sent one Franciscan, Giovanni da Montecorvino (1247–1328). He reached Khanbaliq (Peking) in 1294 and launched a small but successful mission. In 1342 Giovanni dei Marignolli arrived with 32 other missionaries, but their work flourished for less than 25 years. The succeeding Ming dynasty excluded foreigners. Twice Christianity had entered and disappeared from China.

*The rise of Islām.* Between Muḥammad's death in 632 and the defeat of Muslim forces at Poitiers by Charles Martel's Franks in 732, Arab Muslims had taken the Middle East and Egypt, then swept across North Africa, turned northward through Spain, and ventured briefly into southwestern France. Within a century Islām had eliminated more than half of Christendom.

Encouraged by the papacy, the Iberian reconquest (742–1492) became a crusade against Islām and fused an Iberian Catholicism that Spain and Portugal later transplanted around the globe. In the late 20th century its members represented more than half the world's Roman Catholics. The Crusades (1095–1396) produced among many Christians an adversarial approach to those of other faiths. Ramon Llull (*c.* 1235–1316) pursued a different way. He studied Arabic and sought through dialogue and reason the conversion of Muslims and Jews.

As a result of the second great transition the faith of the Mediterranean world had become that of all Europe and had largely created its civilization. Christendom had lost half its members to Islām, but Europe had become the new centre of the Christian faith.

**Third transition, to AD 1950.** By 1500 Europe was bursting with new energy and achievement, and from it Christianity spread worldwide. Iberian monks in the 16th century spanned the globe, and 300 years later Protestant missionaries did the same.

*Roman Catholic mission, 1500–1950.* With Europe cut off from Asia by the Muslims, Portugal's Prince Henry the Navigator (1394–1460) launched exploratory voyages along the western coast of Africa. In 1498 Vasco da Gama reached India; others pushed to Asia's eastern limits. Papal grants in 1454 and 1456 gave Henry all lands, power over the missionary bishops therein, and trading rights south of the Tropic of Cancer. An early Portuguese mission to the Congo produced an African bishop, but the church quickly disappeared. Other efforts on both African coasts also were unsuccessful.

Spain sought a route to India through Columbus' westward voyages. In 1494 a papal grant gave Spain everything west of 47° W longitude (eastern Brazil). Under royal patronage (*patronato real,* or *padroado*), monarchs of both nations accepted responsibility for evangelizing the newly found peoples. Franciscans, Dominicans, Augustinians, and, from 1542, Jesuits staffed the resulting missions.

By 1600 France was becoming the third great imperial Roman Catholic power and was also deeply involved in mission. At the same time England, Holland, and Denmark—all Protestant—began an imperial thrust that challenged the Roman Catholic powers in their own territories.

When the Europeans arrived in the Americas, the Indian population south of the Rio Grande numbered some 35,000,000, but in North America there were at most 1,200,000 Indians—a marked difference. The great majority of European males entering Latin America were unmarried and quickly produced a mestizo, or mixed, population. European settlers, who expected to instruct the Indians in the faith and protect them, gained their labour. The Indians were used widely as slaves and often were treated cruelly. Bartolomé de Las Casas (1474–1566) championed their cause but, ironically, favoured increasing the already growing number of African slaves.

Despite its weaknesses, the Roman Catholic mission gained vast numbers. Although later modified, a 1555 decree which held that Indians, mestizos, and mulattos could not be ordained proved ruinous. Never inspired to produce their own clergy, the new Christians became dependent upon European clergy.

Franciscans and Dominicans traveled widely and built mission churches. The most notable development—used widely but most fully developed by Jesuits in Paraguay—was the appearance of the *reducciones.* In these mission-operated villages Indians were instructed in the faith, taught to develop trades, and protected from the Europeans. Despite good intentions this environment did not produce strong Christians. The movement dissolved when the Jesuits were disbanded in 1773.

Much of the evangelization appeared to be an integral part of military conquest. Yet in whatever way Indians and mestizos intermingled past beliefs and practices with their Christian faith, the majority thought of themselves as Roman Catholic.

Evangelization in French North America followed a somewhat different course. In 1534 Jacques Cartier claimed New France (Canada) for his homeland. A century later French missionaries began to enter the territory. In their work these missionaries sought to reshape Indian life as little as possible.

Efforts of St. Francis Xavier

In Asia, chiefly through the Jesuits, some of the most productive missions appeared. Under a papal commission the Jesuit missionary Francis Xavier (1506–52) reached Goa in 1542. He established Christian communities in India, built a college in Goa for training priests, began a prospering mission in Japan, and died off the coast of China while hoping to enter that land.

By 1600 there were about 300,000 Christians in Japan. Christianity was proscribed, thousands were martyred, and the Japanese sealed themselves off from the West.

China also was closed to foreigners, but the Italian Jesuit missionary Matteo Ricci (1552–1610) arrived in 1582 and eventually reached the capital. His efforts brought success, and other Jesuits followed. An edict of toleration was proclaimed in 1692. Ricci's conviction that the honouring of ancestors and Confucius was a social rite that could be accommodated within the church produced the Chinese Rites Controversy (1634–1742). It brought bitter opposition from Dominicans and Franciscans. Attempts at papal intervention at the beginning of the 18th century angered the Emperor. The Chinese forced missionaries to leave the country and persecuted Christians. Yet by 1800 some 250,000 remained, and since the 16th century the church has been continuously present in China.

In India Jesuits were welcomed to the court during the reign of Mughal emperor Akbar (1556–1605). The noted Jesuit Roberto de Nobili (1577–1656) sought points of agreement between Hinduism and Christianity as a means of evangelization, but this caused difficulty with the church. The missionaries also worked among India's existing Christian communities. In 1599 the Roman Catholic Church brought the South Indian Christians (Nestorians) into its fold, but in 1653 about 40 percent of the Syrian, or Thomas, Christians revolted and linked themselves with the Jacobites. Nevertheless, the Roman Catholics retained a solid base of Christians on which to build.

To provide knowledgeable oversight and to coordinate policy, in 1622 Pope Gregory XV established the Sacred Congregation for the Propagation of the Faith (Propaganda Fide), or the Propaganda. It provided a library for research and a school for training priests and missionaries, assigned territories, and directed ecclesiastical matters overseas. The Foreign Missionary Society of Paris (1663), directed exclusively toward outreach to non-Christian peoples, sought to produce rapidly an indigenous secular clergy (*i.e.,* one not bound to a religious order), and focused its efforts on Vietnam, Cambodia, Laos, and Thailand.

With the suppression of the Jesuits (1773–1814) and the decline of Spanish and Portuguese influence, Roman Catholic missions found themselves at low ebb, but French and other European missionaries steadily took up the slack. Between 1800 and 1950 new vigour paralleled that seen in Protestantism and brought new orders—such as the Society of the Divine Word (1875) and the Catholic Foreign Missionary Society of North America (1911) of Maryknoll fathers and sisters—and voluntary societies to promote and support missions. The missionary force remained overwhelmingly European.

*Protestant missions, 1500–1950.* Protestant missions emerged some 275 years after Martin Luther launched the Reformation in 1517. Reasons for the delay included Protestantism's thorough rejection of the theocratic and universal claims of the papacy and its rationale for papal mission. It also vigorously rejected monasticism and lacked the structure for mission that monasticism had supplied. Some Protestants—especially the Anabaptist but also other prophetic voices, including Adrian Saravia (1531–1613) and Justinian von Welz (1621–68)—called for mission but were scarcely heard.

Protestant expansion through colonization

Protestants began to expand overseas through migration, notably to North America. To minister to the colonists' needs, individual Anglicans formed the Society for Promoting Christian Knowledge (SPCK; 1698) and the Society for the Propagation of the Gospel in Foreign Parts (SPG; 1701), whose chaplains were also to spread the Gospel among non-Christians. The Dutch East India Company trained ministers in Leiden to serve their employees in Indonesia and Ceylon (Sri Lanka), but they were also encouraged to catechize and baptize local people.

European colonization of North America aroused interest in the American Indians, and the Virginia and Massachusetts charters enjoined their conversion. The mission of John Eliot (1604–90) to the Pequot Iroquois and that of the Thomas Mayhew family encouraged formation of supporting societies in Britain.

The German Lutheran Pietists were the first Protestant group to launch church-supported continuing missions from the Continent. Philipp Jakob Spener (1635–1705) and August Hermann Francke (1663–1727) at the Pietists' University of Halle trained Bartholomäus Ziegenbalg (1683–1719) and Heinrich Plütschau (1678–1747). From 1706 they served the Danish mission of King Frederick IV at Tranquebar, in South India. Also trained at Halle, Nikolaus Ludwig, Count von Zinzendorf (1700–60), received Moravian refugees at his Herrnhut estate and in 1732 molded them into a missionary church. Their small, self-supporting communities spread from Greenland to South Africa.

William Carey's *Enquiry into the Obligations of Christians, to Use Means for the Conversion of the Heathens* (1792) became the "charter" for Protestant missions and produced the Baptist Missionary Society. In 1793 Carey went to India. His first letter to an England stirred by the Evangelical Revival resulted in the formation of the London Missionary Society (1795). The Scottish Missionary Society (1796) and the Netherlands Missionary Society (1797) soon appeared, Anglican evangelicals organized the Church Missionary Society (1799), and many others followed. Like the SPCK and SPG, they were founded not by churches but as autonomous societies supported chiefly by denominational constituencies. Similarly, in Europe these organizations were usually created geographically—such as the Basel (1815), Berlin (1824), and Leipzig (1836) societies.

With separation of church and state in the United States, American churches made plain that mission was the responsibility of each Christian. Most denominations developed their own boards or societies. The American Board of Commissioners for Foreign Missions (1810) was the first, and the pattern of denominational societies spread.

Until 1890 American Protestant missions centred on the new immigrants and those following the westward-moving frontier, but from 1890 they turned their attention to areas abroad. In 20th-century "overseas" missions, English-speaking participants have represented from 80 to 89 percent, and North Americans about 67 percent, of all Protestant missionaries.

Women in Protestant missions

Women have not only provided the major support for mission in the modern era but also early recognized the need to found their own societies and send their own missionaries. In much of the world, because of local customs, women missionaries could perform services for other women and for children, especially in medicine and education, that men could not undertake. Their greatest impact was in the production of vast corps of able and educated women, especially in Asia, who played major roles in the professions and in church leadership.

Nondenominational faith missions viewed J. Hudson

Taylor's China Inland Mission (1865; after 1965 called the Overseas Missionary Fellowship) as the great prototype. Missions such as these often sought to work in areas unoccupied by other missionaries, guaranteed no salaries, and left financial support in God's hands; but most bodies made their financial needs known to a wide constituency. Their chief aim has been to proclaim the Gospel and eschew the provision of social services. These societies joined together in the Interdenominational Foreign Mission Association (IFMA; 1917). Since the 1960s they have cooperated with the Evangelical Foreign Missions Association (EFMA; 1945), the missionary arm of the National Association of Evangelicals (1943), and, at the international level, with the World Evangelical Fellowship (1952).

In the early 19th century in India, William Carey, Joshua Marshman, and William Ward—the Serampore trio—worked just north of Calcutta. Their fundamental approach included translating the Scriptures, establishing a college to educate an Indian ministry, printing Christian literature, promoting social reform, and recruiting missionaries for new areas as soon as translations into that area's language were ready.

Alexander Duff (1806–78) gave India the pattern for an entire educational system, including colleges. By the 1860s education for women had advanced and nurses' training had begun; the education of women physicians began at the turn of the century. The Vellore Medical College is a monument to the missionary physician Ida Scudder (1870–1959). The vast majority of Indian nurses also have been Christian.

In Indonesia, Dutch chaplains established churches in the 17th century. In the mid-19th century, the German Rhenish Missionary Society enabled the Batak Church of Sumatra to grow in size and commitment and to provide leadership for the nation. Other strong churches developed in various parts of predominantly Muslim Indonesia.

Following the Opium War treaties of 1842–44 and 1858–60, China was opened to Westerners. Although Roman Catholics remained from the efforts of the 16th-century Jesuit mission, the Chinese viewed Christianity as entering their homeland at gunpoint. The Boxer Rebellion of 1899–1900 brought death to thousands of Chinese Christians and several hundred missionaries. Yet what Protestant schools, colleges, and hospitals represented attracted Chinese youth to the Christian faith. With the fall of the Ch'ing, or Manchu, dynasty in 1911, Sun Yat-sen, a Christian favouring parliamentary government, became the provisional president. The Christian influence in China, particularly in education, was significant. In 1949, when the People's Republic of China was formed, Christians represented only 1 percent of the Chinese population, but they exercised an influence out of all proportion to their size.

The Chinese government expelled all missionaries in 1950–51, confiscated churches, and brought pressure on Christians. During the Cultural Revolution (1966–76) no churches or other religious bodies could operate. Christians continued to exist in China, but they suffered grievously. From 1976, as the government allowed some churches to open, Christians reemerged throughout the country. Roman Catholic and Protestant churches were filled, and in varied ways "silent" house-churches testified that the underground church had been dynamically growing. The state of the church in China, despite persecution, is considerably larger and stronger than it had been in 1949.

Koreans baptized as Roman Catholics in China returned in 1784 but remained underground when their faith was soon proscribed. A handful of American Presbyterians and Methodists entered Korea in 1884, and the faith they planted flourished through the 20th century, despite Korea's long wartime devastation. Evangelistic and self-supporting Korean churches were known throughout Asia for their effective promotion of Bible study. Helen Kim, a Korean graduate of Ewha College, built it into the world's largest women's university.

Christianity in Korea

Unlike other Asian countries, Korea did not experience Christianity's arrival with Western imperialism but rather saw that faith as reinforcing Korean nationalism against Japanese imperialism from 1910 to 1945. Korean evangelization enabled the church to grow in less than a century to about one-third of the population in South Korea. In the late 20th century, strong annual compounding growth continued—a situation unique among the Asian nations.

The vast Pacific Ocean, with tiny, scattered island kingdoms among the Polynesian, Micronesian, and Melanesian peoples, early attracted missionaries. Most of them were laypeople of deep Christian faith. It was the effort of the Christian islanders, however, that achieved virtually total evangelization of the Pacific.

In the Middle East, Protestants emphasized schools, colleges, and hospitals and witnessed to Muslims, though few Muslims became Christians. Humanitarian assistance by the Near East Relief, begun by American Protestant missions, helped those suffering during and after World War I, but this organization made its greatest efforts to aid the Armenian victims of genocide and forced deportation by the Turks. Later mission work was undertaken by the Near East Christian Council for Missionary Cooperation in Beirut (1924, 1929), which became the broadly ecumenical Near East Council of Churches in 1964.

Three major religions appear in sub-Saharan Africa: African traditional religion, Islām, and Christianity. Protestant missionaries were working in most of the West and Central African colonial nations in the 19th century, but in some parts of East Africa mission began only in the 20th century. After Ghana gained freedom in 1957, many former colonies were granted independence. Cataclysmic change appeared everywhere: in building new nations; rapid shifts from a rural to an urban population; coping with the massive problem, especially in cities, of some 2,500 languages; and developing literacy. Amid all this, Christianity grew with increasing rapidity. By 1980 more than half of the sub-Saharan African population was Christian. African independent, or indigenous nonwhite, churches proliferated, and several of the largest ones joined the World Council of Churches. These churches remain an important factor in African Christianity.

In the 19th century Evangelical churches were begun in Latin America by Protestant missionaries who were largely from the United States but also in some instances from Britain and Germany. Most of these churches have remained small. The exception was the explosion of Pentecostalism throughout the region, with heaviest concentration in Brazil, Chile, and Mexico. Evangelicals also have gained members in Central America.

Cooperation and unity

Protestants quickly discovered the need for cooperation and unity. As tiny minorities in lands of other religions, new Christians and missionaries together saw that denominational separatism hindered evangelization. Four streams led to the cooperation and unity reflected in the World Missionary Conference (WMC) held in Edinburgh in 1910.

First, missionary "field" conferences affirmed comity (separation of spheres of work), cooperation in Bible translation and missionary councils, and shared sponsorship in major enterprises such as hospitals and colleges. A second stream involved missionary conferences in England and the United States from 1854 to 1900. A third force flowed through the missionary concern of the international student Christian and missionary movements. The fourth stream arose in the West from continuing interdenominational conferences of mission leaders to face common concerns and forge common policies. Among others, these included the Continental European Missions Conference (1866) and the Foreign Missions Conference of North America (1893).

The Edinburgh conference was unique—a landmark and watershed for all that was to follow. Largely Western in membership, but with 17 Asian delegates, it created a Continuation Committee that in 1921 became the International Missionary Council (IMC). The IMC consisted of a worldwide network of Christian councils and the Western cooperative agencies. WMC continuation conferences in Asia (1912–13) reproduced the earlier conference in settings that incorporated national leaders of the Asian churches.

From the WMC and IMC also flowed the Faith and Order Movement (concerned with doctrine and ministry), Life and Work Movement (on the churches' moral respon-

sibility in society), and the World Council of Churches (WCC; 1948). The IMC's member bodies became national councils of churches. The IMC and the WCC, officially "in association with" each other, worked closely together. In 1961 the IMC became the Division of World Mission and Evangelism of the WCC.

*Orthodox missions.* Virtually the entire outreach of the Russian Orthodox mission extended to the peoples of the vast Russian Empire across Asia. Its outstanding missionaries included the linguist and translator Nicholas Ilminsky (d. 1891) and Ivan Veniaminov (1797–1879), who in 1823 went as its first missionary to the Aleutian Islands. Veniaminov eventually became Metropolitan Innocent of Moscow, and in 1870 he founded the Russian Orthodox Missionary Society. The Russian Orthodox Church opened a mission to Japan in 1854 and in 1941 turned over all church property to its members.

For some decades the church appointed missionaries to its highest posts. Tikhon (1865–1923), who in 1917 became the first patriarch in two centuries, and Sergius (Stragorodsky; 1867–1944), who followed him in that post, had both served missions abroad. Following the 1917 Revolution, Russian missions became impossible.

The African Orthodox Church, founded in Uganda in the 1920s by Reuben M. Spartas, spread to Kenya and Tanzania.

**Fourth transition, from 1950.** During the third transition, Christianity had spread worldwide from a base in Europe. The fourth transition brought the reality that more Christians lived in Asia, the Pacific Islands, Africa, and Latin America than in the old Christendom, part of a long-term, continuing shift in Christianity's numerical and vivifying centre. The growing churches brought new life and dynamism to the faith, along with new theologies and concerns.

ihift in
Christian-
ty's centre

The growth of the world Christian community kept pace with the 20th-century population explosion, and in the fastest-growing areas the growth rate in numbers of Christians was almost three times greater than the general population increase. The majority of the world's Christians lived in non-Western nations; a universal church had come into being.

In this transition two issues were especially prominent. First, the church found itself engaged with those of traditional or new religions and those for whom ideologies had become religions. In that setting the Roman Catholic Church and the Orthodox and Protestants in the World Council of Churches affirmed evangelization to be essential but also advanced dialogue for clarity, understanding, and basic engagement with other religions. This effort brought dissent and tension from many.

Second, "Third World theologies" often brought angry debate. The underlying questions concerned the identification of what was essentially Christian in Western Christianity and theology and whether Western church structures and theologies were universally normative. But the most basic question asked how Christians of all races could manifest the unity and obedience for which their Lord prayed.

Another force was the worldwide growth in the number of Pentecostals and charismatics. They formed new churches, appeared in traditional churches, and found outlet in many nonwhite indigenous bodies. Pentecostals and charismatics were most heavily concentrated in Latin America and Africa but also had grown in Asia and in the West. They forced theological reflection—perhaps best developed by Roman Catholics—on the doctrine of the Holy Spirit and on authority within the church.

'atican
I and
nission

The second Vatican Council (1962–65) stood as the most important ecclesiological and missiological event for Roman Catholics since the 16th century. Theologically it set itself within the dynamics of the faith's fourth transition. The council's Decree on the Church's Missionary Activity (*Ad Gentes*) built theologically on the council's foundational document, the "Dogmatic Constitution on the Church" (*Lumen Gentium;* "Light of the Nations").

The "Dogmatic Constitution on the Church" rooted the church and mission in the triune Godhead, insisted upon evangelization but presented a larger understanding of God's grace for those outside the church, and urged missionaries to pursue dialogue.

In 1975 Pope Paul VI, responding to the ensuing debate, declared in *Evangelii Nuntiandi* ("Evangelization in the Modern World") that God can achieve salvation in anyone through God's own ways, but that witnessing to and preaching the Gospel is the regular pattern given to Christians. The Pope also presented a theology of liberation. In many respects his statement refined and replaced "The Church's Missionary Activity" (1965).

In 1968 the Latin American Episcopal Conference (CELAM) at Medellín, Colom., worked to apply the insights and intent of Vatican II to Latin America: first, to identify with the aspirations of the masses, and, second, to seek "re-evangelization" and "reconversion" in Latin America. At CELAM III (1979) in Puebla, Mex., the final document, "God's Saving Plan for Latin America," set forth a structure built upon *Evangelii Nuntiandi.*

**Scripture translations.** The translation of the Holy Scriptures has constituted a basic part of mission. During the Middle Ages few could read the Latin Bible. Within 80 years of the invention of printing in the West, however, Reformation leaders such as Luther and Calvin focused on the Word of God. Their cardinal principle remained that each should be able to read the Bible in his own tongue. The result was the development of education and literacy. The printing press greatly aided Protestantism, and widespread literacy again became the hallmark of a civilized society.

In the 20th century most of the world's people speak one of about 75 primary languages. A small minority speak one of 450 secondary languages, and more than 4,400 other languages are in use. Through Christian world mission, printed Scriptures have become available in the mother tongues of almost 99 percent of the world's people. That unprecedented accomplishment marks the greatest achievement in the history of written communications. Bibles are available in more than 300 languages, complete New Testaments in nearly 700 languages, and some portion of the Scriptures is available in 1,000 other languages. The translation effort, most of which has occurred during the past 200 years, has in many cases reduced a language to writing for the first time. The effort involved the production of grammars and dictionaries of these languages as well as scriptural translations, and an additional benefit has been the written preservation of the cultural heritage by native speakers of the language.

Bible societies, including the United Bible Societies (1946), have coordinated and aided the translation work of missionaries in this task for almost 200 years. Wycliffe Bible Translators (1936) concentrated its work among the language groups having the smallest numbers of speakers. From 1968 Roman Catholics and the United Bible Societies have coordinated their efforts and cooperated in translation and production wherever possible.

Christianity, unlike some of the other world religions, is a translating faith. In that area of God's mission the chief work in recent centuries has come from the Protestant community and has been offered as a gift to the church universal. This constitutes one of the great contributions of Christian mission to the world. (W.R.H.)

## Ecumenism

The word ecumenism comes from a family of classical Greek words: *oikos,* meaning a "house," "family," "people," or "nation"; *oikoumenē,* "the whole inhabited world"; and *oikoumenikos,* "open to or participating in the whole world." Like many biblical words, these were invested with Christian meaning. The *oikoumenē* describes the place of God's reconciling mission (Matthew 24:14); the unity of the Roman Empire (Luke 2:1) and of the kingdoms of the earth (Luke 4:5); and the world destined to be redeemed by Christ (Hebrews 2:5). In the biblical community the vision of one church serving the purposes of God in the world came to reflect a central teaching of the early Christian faith, the essence of the church.

In later centuries the word ecumenical was used to denote those councils (*e.g.,* Nicaea, Chalcedon) of bishops whose

decisions represented the universal church, in contrast to other church councils that enjoyed only regional or limited reception. The honorary title of ecumenical patriarch was given to the Greek Orthodox patriarch of Constantinople because his see was located in the capital of the *oikoumenē* and his leadership was accepted as primus inter pares (first among equals) in the faith and mission of the whole church. The Apostles', the Nicene, and the Athanasian creeds are called ecumenical because they witness to the universal faith of all Christians. In the 19th and 20th centuries ecumenism denoted the movement of the renewal, unity, and mission of Christians and churches of different traditions "so that the world may believe."

Ecumenism is a vision, a movement, a theology, and a mode of action. It represents the universality of the people of God, which affects the way Christians think about their faith, the church, and the world. Ecumenism, which is a long process, includes Bible study, dialogue, prayer, eucharistic worship, common witness, diaconal service, and ecclesial unity that draws Christians together, uniting their life and mission and bringing the Body of Christ and the human community closer to the fulfillment of God's purposes. To be involved in ecumenism means to participate in those ideas, activities, and institutions that express a spiritual reality of shared love in the church and the human community. It involves the work of officially organized ecumenical bodies, the confessing and witnessing of Christians in local places, and the spirituality and actions of those who live together in love and prophetic proclamation. Far more than a program or an organization, ecumenism is, according to the British ecumenist Oliver S. Tomkins, "something that happens to the soul of Christians."

Ecumenical participation

Any unity worthy of this vision cannot be identified with political or spiritual coercion, strategies of dominance or superiority, calls for "a return to the mother church," or expectations of monolithic uniformity or a super-church. When serving the cause of faith, the weapons of faith are not those of force or intolerance; neither can divisions be overcome nor authentic unity manifested by syncretism, a least-common-denominator theology, or a casual friendliness. Ecumenism accepts the diversity of God's people, given in creation and redemption, and strives to bring these confessional, cultural, national, and racial differences into one fully committed fellowship.

Ultimately the purpose of ecumenism is to glorify the triune God and to help the one missionary church to witness effectively and faithfully among all peoples and nations. In the last half of the 20th century Christians have learned and confessed new dimensions of this vocation, especially in relation to what divides the churches. Progress has been made on historical theological issues that have alienated Christians through the centuries—baptism, the Eucharist, and ministry. But equally divisive among Christians are the divisions of the human family: racism, poverty, sexism, war, injustice, and differing ideologies. These issues are part of the agenda of ecumenism and bring a particular context, dynamic spirit, and urgency to the pursuit of Christian unity as well as of justice and peace. The church's unity becomes essential for the renewal and unity of the human family. Through its unity the church becomes a sign, the firstfruits of the promised unity and peace among God's peoples and the nations.

## THE BIBLICAL PERSPECTIVE

The unity of the church and of all creation is a dominant motif in the Bible. This witness begins in the Old Testament, or Hebrew Scriptures, not the New Testament. God established a covenant with the Hebrew people and gathered the disparate tribes into one religious nation, Israel, taking steps to overcome the alienation between God and humans and to reconcile God's people. The tradition of ancient Judaism, therefore, was based on the reality of the one people of God. Their unity was an expression of their monotheistic faith, the oneness of God (Yahweh). As Genesis records, God created the world as one cosmos, an ordered unity determined by one single will in which all creatures are responsive to the purposes of the Creator. Yahweh chose Israel from all the nations of the world and entered into covenant with its people. Whenever men and women sinned and alienated themselves from God and from one another, God acted to bring about their reconciliation. Israel's mission was to preserve the faithfulness and unity of all God's people and to prepare them for the realization of the Kingdom of God.

The vision of unity is central to the Gospel of Jesus Christ and the teachings of his Apostles. Those who confess Jesus as Lord and Saviour are brought together in a new community: the church. All New Testament writers assume that to be "in Christ" is to belong to one fellowship (Greek: *koinōnia*). Jesus clearly gave the mandate when at the Last Supper he offered his high-priestly intercession, praying that the disciples and all those who believe in him "may all be one; even as thou, Father, art in me, and I in thee . . . so that the world may believe that thou hast sent me" (John 17:21). This unity was evidenced in the miracle of Pentecost (Acts 2) and other actions that constituted the primitive church—*e.g.*, the epoch-making Council of Jerusalem (Acts 15), which negotiated conflicts between Jewish and Gentile Christians.

Christ's mandate for unity

The early church nevertheless had many tensions and conflicts that called for ecumenical proclamations and pleas from the writers and Apostles of the New Testament. Tensions arose between Jewish Christian churches and Gentile Christian churches, between Paul and the enthusiasts, between John and early Catholicism. Peter and Paul disagreed strongly over whether Gentiles had to fulfill Jewish requirements in order to be welcome at the Lord's Supper (Eucharist). That theological aberrations challenged the young church is shown in the New Testament: Colossians refutes Gnosticism; the Johannine Epistles warn against Docetism; 2 Peter and Revelation attack false prophets.

None of this diversity created schism nor allowed a break in fellowship. There were no denominations or divided communities, as were to develop later in the church's history. Division among Christians is a denial of Christ, an unthinkable distortion of the reality of the church. Amid their diversity and conflicts the early Christians remained of "one accord," visibly sharing the one Eucharist, accepting the ministries of the whole church, reaching out beyond their local situation in faith and witness with a sense of the universal community that held all Christians together. As Paul taught the Ephesians, God's ultimate will and plan is "to unite all things in him [Christ], things in heaven and things on earth" (chapter 1, verse 10).

## THE HISTORY OF ECUMENISM

While unity is given in Christ, church history chronicles two diametric forces in the church's life. One is the tendency toward sectarianism and division; the other is the conviction toward catholicity and unity. Ecumenism represents the struggle between them. Some of the schisms were theological conflicts foreshadowed in the apostolic church; others were internal quarrels related to liturgical differences, power politics between different patriarchates or church centres, problems of discipline and piety, or social and cultural conflicts. Nevertheless, according to the American historian John T. McNeill, "the history of the Christian Church from the first century to the 20th might be written in terms of its struggle to realize ecumenical unity."

**Early controversies.** A long and continuing trail of broken relations among Christians began in the 2nd century. Early in the 2nd century the Gnostics presented a serious doctrinal error and broke fellowship. Quartodecimanism, a dispute over the date of Easter, pitted Christians from Asia Minor against those from Rome. Montanism—which taught a radical enthusiasm, the imminent Second Coming of Christ, and a severe perfection, including abstinence from marriage—split the church. The Novatians broke fellowship with those Christians who, under pressure, offered sacrifices to pagan gods during the persecutions of the Roman emperor Decius in AD 250. In the early 4th century the Donatists, Christians in North Africa who prided themselves as the church of the martyrs, refused to share communion with those who had lapsed (*i.e.*, who had denied the faith under threat of death). The church

in Rome received the lapsed back into fellowship after services of repentance. This schism—like many since—reflected regional, national, cultural, and economic differences between the poor, rural North African Christians and the sophisticated, urban Romans.

In each century leaders and churches sought to reconcile these divisions and to manifest the visible unity of Christ's church. But in the 5th century a severe break in the unity of the church took place. The public issues were doctrinal consensus and heresy, yet in the midst of doctrinal controversy alienation was prompted by political, cultural, philosophical, and linguistic differences. Tensions increased as the church began to define—amid critical distortions by some—the relationship between God the Father and God the Son and later the relation between the divine and human elements in the nature and person of Jesus Christ. The first four ecumenical councils—at Nicaea (AD 325), Constantinople (381), Ephesus (431), and Chalcedon (451)—defined the consensus to be taught and believed, articulating this faith in the Nicene Creed and the Chalcedonian Definition, which stated that Jesus is the only begotten Son of God, true man, and true God, one person in "two natures without confusion, without change, without division, without separation." Two groups deviated doctrinally from the consensus developed in the councils. The Nestorians taught that there are two distinct persons in the incarnate Christ and two natures conjoined as one; Monophysites taught that there is one single nature, primarily divine. Several churches refused to accept the doctrinal and disciplinary decisions of Ephesus and Chalcedon and felt that they had no alternative but schism. These churches, called pre-Chalcedonian or Oriental Orthodox, became great missionary churches and spread to Armenia, Egypt, Ethiopia, Syria, Persia, and the Malabar coast of India in isolation from other churches.

Disputes regarding the nature of Christ

**The schism of 1054.** The greatest schism in church history occurred between the church of Constantinople and the church of Rome. While 1054 is the symbolic date of the separation, the agonizing division was six centuries in the making. The friction was ignited by several issues. The Eastern Church sharply disagreed when the Western Church introduced into the Nicene Creed the doctrine that the Holy Spirit proceeds not from the Father alone—as earlier Church Fathers taught—but from the Father and the Son (Latin: *filioque*). When the Roman Empire was divided into two zones, Latin-speaking Rome began to claim superiority over Greek-speaking Constantinople; disputes arose over church boundaries and control (for example, in Illyricum and Bulgaria). Rivalry developed in Slavic regions between Latin missionaries from the West and Byzantine missionaries from the East, who considered this territory to be Orthodox. Lesser matters related to worship and church discipline—for example, married clergy (Orthodox) versus celibacy (Roman Catholic) and rules of fasting—strained ecclesial relations. The tensions became a schism in 1054, when the uncompromising patriarch of Constantinople, Michael Cerularius, and the envoys of the uncompromising Pope Leo IX excommunicated each other. No act of separation was at this time considered final by either side. Total alienation came a half century later, as a result of the Crusades, when nominally Roman Catholic Christian soldiers made military campaigns to save Jerusalem and the Holy Land from the Muslims. In 1204 the Fourth Crusade was diverted to attack and capture Constantinople brutally. Thousands of Orthodox Christians were murdered; churches and icons were desecrated. As a consequence undying hostility developed between East and West.

Even so, certain leaders and theologians on both sides tried to heal the breach and reunite East and West. In 1274 the second Council of Lyon sought reunion. Agreements among the negotiators were achieved, including Orthodox acceptance of papal primacy and the acceptance of the Nicene Creed with the *filioque* clause. But the agreements were only a rushed action conditioned by political intrigue. As a result, reunion on these terms was fiercely rejected by the clergy and laity in Constantinople and other Orthodox provinces. A second attempt at reunion came at a council that met in Italy at Ferrara in 1438 and Florence in 1439. A formula of union was approved by both delegations, but later it was rejected by rank-and-file Orthodox Christians.

**The Reformation.** The 16th century experienced the next dramatic church division in the Reformation in the West. Like other schisms this one does not yield to simple analysis or exegesis. The Reformation was a mixture of theology, ecclesiology, politics, and nationalism, all of which led to breaks in fellowship and created institutional alienation between Christians in Germany, France, Switzerland, Scotland, England, and elsewhere. In one sense it was a separation, especially a reaction against the rigid juridical structures of medieval Roman Catholicism and its claim to universal truth and jurisdiction. In another sense, however, the Reformation was an evangelical and ecumenical renewal of the church as the Body of Christ, an attempt to return to the apostolic and patristic sources in order, according to Calvin, "to recover the face of the ancient Catholic Church." All the continental Reformers sought to preserve and reclaim the unity of the church.

Once the separation between the Roman Catholic Church and the emerging Protestant churches was conclusive, irenic persons on both sides tried to restore unity. Roman Catholics such as Georg Witzel and George Cassander developed proposals for unity, which all parties rejected. Martin Bucer, celebrated promoter of church unity among the 16th-century leaders, brought Luther and his colleague Philipp Melanchthon into dialogue with the Swiss Reformer Zwingli at Marburg, Ger., in 1529. In 1541 Calvin (who never ceased to view the church in its catholicity), Bucer, and Melanchthon met with Cardinal Gasparo Contarini and other Roman Catholics at Ratisbon (now Regensburg, W.Ger.) to reconcile their differences on justification by faith, the Lord's Supper, and the papacy. Another attempt was made in 1559, when Melanchthon and Patriarch Joasaph II of Constantinople corresponded, with the intention of using the Augsburg Confession as the basis of dialogue between Lutherans and Orthodox. On the eve of the French religious wars (1561) Roman Catholics and Protestants conferred without success in the Colloquy of Poissy. It would seem that the ecumenical projects of theologians and princes in 16th-century Europe failed unequivocally, but they kept alive the vision and the hope.

Ecumenical efforts of Bucer and Melanchthon

**Ecumenism in the 17th and 18th centuries.** During the 17th and 18th centuries storms of contention and division continued to plague the churches throughout Europe. The Church of England, severed from the Roman Catholic Church in the 16th century by Henry VIII and his English theologians, experienced its first internal split when it was unable or unwilling to embrace John Wesley and the Methodist Church's insights for spiritual renewal.

During these two centuries there was an eclipse of official, church-to-church attempts at unity. Instead, ecumenical witness was made by individuals who courageously spoke and acted against all odds to propose Christian unity. John Amos Comenius, a Czech Brethren educator and advocate of union, produced a plan of union for Protestants based upon the adoption of a scriptural basis for all doctrine and polity and the integration of all human culture.

Ecumenical witness by individuals

In England, John Dury, a Scots Presbyterian and (later) an Anglican minister, "a peacemaker without partiality," traveled more extensively than any other ecumenist before the 19th century, negotiating for church unity in his own country and in Sweden, Holland, France, Switzerland, and Germany. Richard Baxter, a Presbyterian Puritan, developed proposals for union, including his Worcestershire Association, a local ecumenical venture uniting Presbyterians, Congregationalists, and Anglicans.

Efforts were undertaken in Germany as well. The German Lutheran George Calixtus called for a united church between Lutherans and Reformed based on the "simplified dogmas," such as the Apostles' Creed and the agreements of the church in the first five centuries. Count Nicholas von Zinzendorf applied his Moravian piety to the practical ways that unity might come to Christians of all persuasions. The philosopher Gottfried Wilhelm Leibniz worked tirelessly for union between Protestants and Roman Catholics, writing an apologia interpreting Roman Catholic doctrines for Protestants.

Orthodox Christians also participated in the search for union. Metropolitan Philaret of Moscow and the Russian Orthodox theologian Aleksey S. Khomyakov expressed enthusiasm for ecumenism. Cyrillus Lukaris, Orthodox patriarch of Alexandria and later of Constantinople, took initiatives to reconcile a divided Christendom. People throughout Europe held tenaciously to the dream of ecumenism, although no attempt at union reached fruition.

**19th-century efforts.** In the 19th century a worldwide movement of evangelical fervour and renewal, noted for its emphasis on personal conversion and missionary expansion, stirred new impulses for Christian unity. The rise of missionary societies and volunteer movements in Germany, Great Britain, The Netherlands, and the United States expressed a zeal that fed the need for church unity. As missionaries in different countries began to experience the harmful results of Christian divisions, cooperation among Protestant missionaries began to take place in India, Japan, China, Africa, Latin America, and the United States.

Growth of ecumenical societies

In 1804 the British and Foreign Bible Society came into existence to bring Protestants and Anglicans together in the translation and distribution of the Scriptures. This was followed, 40 years later, by the founding of two important Christian organizations in England: the Young Men's Christian Association (1844) and the Young Women's Christian Association (1855). Their international bodies, the World Alliance of YMCAs and the World YWCA, were established in 1855 and 1894, respectively. The Evangelical Alliance, possibly the most significant agent of Christian unity in the 19th century, held a unique place among the volunteer associations of individuals for common service and mission. Founded in London in 1846 (with an American section in 1867), the alliance sought to draw individual Christians into fellowship and cooperation in prayer for unity, Christian education, the struggle for human rights, and mission.

Also pivotal in the 19th century were advocates for the visible unity of the church. In the United States, where the most articulate 19th-century unity movements were heard, the witness to the unity and union was led by three traditions. Among Lutherans, Samuel Simon Schmucker and Philip Schaff pleaded for "catholic union on apostolic principles." Among Episcopalians, the visionaries for unity included Thomas Hubbard Vail, William Augustus Muhlenberg, and William Reed Huntington, who proposed the historic "Quadrilateral" of the Scriptures, the creeds, the sacraments of baptism and the Lord's Supper, and episcopacy as the keystone of unity. Among the Disciples of Christ the biblical vision of unity was dramatically offered by Thomas Campbell and his son, Alexander, and Barton Warren Stone—all of whom taught that "the Church of Christ on earth is essentially, intentionally and constitutionally one." Ecumenism was enflamed in the hearts of 19th-century Christians and in the next century began to shape the churches as never before.

**Ecumenism in the 20th century.** The 20th century experienced a flowering of ecumenism. Four different strands—the international Christian movement, cooperation in world mission, Life and Work, and Faith and Order—developed in the early decades and, though distinctive in their emphases, later converged to form one ecumenical movement.

The modern ecumenical era began with a worldwide movement of Christian students, who formed national movements in Great Britain, the United States, Germany, Scandinavia, and Asia. In 1895 the World Student Christian Federation, the vision of American Methodist John R. Mott, was established "to lead students to accept the Christian faith" and to pioneer in Christian unity. The

The 1910 Edinburgh conference

World Missionary Conference at Edinburgh (1910) inaugurated another aspect of ecumenism by dramatizing the necessity of unity and international cooperation in fulfilling the world mission of the church. In 1921 the International Missionary Council (IMC) emerged, bringing together missionary agencies of the West and of the new Christian councils in Asia, Africa, and Latin America for joint consultation, planning, and theological reflection. The Life and Work movement was pledged to practical Christianity and common action by focusing the Christian conscience on international relations and social, industrial, and economic problems. Nathan Söderblom, Lutheran archbishop of Uppsala, inspired world conferences on Life and Work at Stockholm (1925) and Oxford (1937). The Faith and Order movement, which originated in the United States, confronted the doctrinal divisions and sought to overcome them. Charles H. Brent, an Episcopal missionary bishop in the Philippines, was chiefly responsible for this movement, although Peter Ainslie, of the Disciples of Christ, shared the same vision and gave significant leadership. World conferences on Faith and Order at Lausanne (1925), Edinburgh (1937), Lund (1952), and Montreal (1963) guided the process of theological consensus-building between Protestants, Orthodox, and Roman Catholics, which led to approval of the historic convergence text on *Baptism, Eucharist, and Ministry* (1982).

The World Council of Churches

The World Council of Churches (WCC) is a privileged instrument of the ecumenical movement. Constituted at Amsterdam in 1948, the conciliar body includes more than 300 churches—Protestant, Anglican, and Orthodox—which "confess the Lord Jesus Christ as God and Saviour according to the Scriptures and therefore seek to fulfill together their common calling to the glory of the one God, Father, Son, and Holy Spirit." Its general secretaries have been among the architects of 20th-century ecumenism: W.A. Visser 't Hooft, (The Netherlands), Eugene Carson Blake (United States), Philip Potter (Dominica), and Emilio Castro (Uruguay). The witness and programs of the WCC include faith and order, mission and evangelism, refugee and relief work, interfaith dialogue, justice and peace, theological education, and solidarity with women and the poor. What distinguishes the WCC constituency is the forceful involvement of Orthodox churches and churches from the Third World. Through their active presence the WCC, and the wider ecumenical movement, has become a genuinely international community.

Roman Catholic ecumenism received definitions and momentum at the second Vatican Council (1962–64), under the ministries of popes John XXIII and Paul VI, and through the ecumenical diplomacy of Cardinal Augustin Bea, the first president of the Secretariat for Promoting Christian Unity. The church gave the ecumenical movement new hope and language in the "Decree on Ecumenism" (1964), one of the classic ecumenical teaching documents. Another result of Vatican II was the establishment of a wide variety of international theological dialogues, commonly known as bilateral conversations. These include Roman Catholic bilaterals with Lutherans (1965), Orthodox (1967), Anglicans (1967), Methodists (1967), Reformed (1970), and the Disciples of Christ (1977). Topics identified for reconciling discussions include baptism, the Eucharist, episcopacy and papacy, authority in the church, and mixed marriage.

Critical to 20th-century ecumenism is the birth of united churches, which have reconciled formerly divided churches in a given place. In Asia and Africa the first united churches were organized in China (1927), Thailand (1934), Japan (1941), and the Philippines (1944). The most heralded examples of this ecumenism are the United Church of Canada (1925), the Church of South India (1947), and the Church of North India (1970). Statistics of other united churches are revealing. Between 1948 and 1965, 23 churches were formed. In the period from 1965 to 1970 unions involving two or more churches occurred in the West Indies in Jamaica and Grand Cayman, Ecuador, Zambia, Zaire, Pakistan, Madagascar, Papua New Guinea, the Solomon Islands, and Belgium. Strategic union conversations were undertaken in the United States by the nine-church Consultation on Church Union (1960) and by such uniting churches as the United Church of Christ (1957), the Presbyterian Church, U.S.A. (1983), and the Evangelical Lutheran Church in America (1988).

Spiritual disciplines play a key role in ecumenism, a movement steeped in prayer for unity. During the Week of Prayer for Christian Unity, celebrated every year (January 18–25), Christians from many traditions engage in prayer, Bible study, worship, and fellowship in anticipation of the unity Christ wills. (P.A.C.)

## The Christian church and non-Christian religions

A spiritual encounter and discussion of Christianity with other world religions has begun only during the 20th century as a consequence of change in the general religious, political, and economic situation of the world. The global spread of Christianity through the activity of the European and American churches in the 18th and 19th centuries led to Christianity's immediate encounter with all other existing religions. Until the beginning of the 19th century there were still places on Earth where non-Christian religions never came into contact with Christianity. Since then, Christianity has entered into a direct contact with all living non-Christian religions in the world. The close connection between Christian world missions and political, economic, technical, and cultural expansion has, at the same time, been loosened.

Recent Christianity's contacts with non-Christian religions

After World War II, the former mission churches were transformed into independent churches in the newly autonomous Asian and African states. The concern for a responsible cooperation of the members of Christian minority churches and its non-Christian fellow citizens became the more urgent with a renaissance of the Asian higher religions in numerous Asian states. Since World War II Hinduism, Buddhism, and Islām have been trying to regain their former position of leadership in intellectual and spiritual life, mainly in the educational systems of their countries in the Asian states and—in the case of Islām—in some African states.

All Asian higher religions have also turned to activities in world missions in Christian countries in Europe, the Americas, and Australia. Hinduism, for example, has founded numerous Vedānta centres in North America and Europe within the framework of the Ramakrishna and Vivekananda missions. South Asian Theravāda (Way of the Elders) Buddhism and the Mahāyāna (Greater Vehicle) Buddhism of Japan (mainly Zen Buddhism, an intuitive-meditative sect) have begun world missionary activities under the influence of a Buddhist renaissance. This influence has penetrated Europe and North America not so much in the form of a directly organized mission as in the form of a spontaneously received flow of religious ideas and methods of meditation through literature, philosophy, psychology, and psychotherapy. As a result, Christianity in the latter part of the 20th century found itself forced to enter into a factual discussion with non-Christian religions, particularly because the constitutional privileges once enjoyed by certain religions had been rescinded in most states.

Modern history of religions, on the other hand, has caused a general transformation of religious consciousness in the West since the middle of the 19th century. Until the beginning of the 20th century, the knowledge of non-Christian higher religions was still the privilege of a few specialists. In the meantime, in a second wave of enlightenment, a wide range of people have studied the results of research in the form of translations of source materials from the non-Christian religions. The spreading of the religious art of Tibet, India, and the Far East through touring exhibitions and the possibility of a direct participation in non-Christian religious ceremonies through radio and television has created a new attitude toward the other religions in the broad public of Europe and North America. The knowledge of the plurality of the world religions characterizes the religious consciousness of the 20th century in a way that was unknown in former centuries. In recognition of this fact, numerous Christian institutions for the study of non-Christian religions have been founded: *e.g.*, in Bangalore, India; in Rangoon, Burma; in Bangkok, Thailand; in Kyōto, Japan; and in Hong Kong. There are also a number of more specialized centres in several countries.

The increase of knowledge of non-Western religions in the West

The readiness of encounter or even cooperation of Christianity with non-Christian religions is a phenomenon of modern times, with few precedents in the history of the struggle of Christianity and the non-Christian religions. Until the 18th century, Christianity showed little inclination to engage in a serious study of non-Christian religions. Four hundred years after the beginning of the struggle with the Muslims in Spain, almost half a century after the proclamation of the First Crusade against Islām, Peter the Venerable, abbot of Cluny, issued the first translation of the Qur'ān (the Islāmic scriptures) in 1141 in Toledo; but he was not understood by his contemporaries. Bernard of Clairvaux, the propagator of the Second Crusade, even refused to read it. Four hundred years later, in 1542/43, Theodor Bibliander, a theologian and successor of the Swiss Reformer Zwingli, edited the translation of the Qur'ān by Peter the Venerable again. He was subsequently arrested, and he and his publisher could be freed only through intervention by Luther.

Knowledge of Hinduism was sometimes deliberately delayed by the missionaries. August Hermann Francke, the supporter of the Lutheran Tranquebar mission in India, prevented the publication of the work of the missionary Bartholomäus Ziegenbalg about the religion of the non-Christian Malabarese of India. The name Buddha is mentioned for the first time in Christian literature—and there only once—by Clement of Alexandria about AD 200; and it vanished after that from Christian literature for a full 1,300 years. Pāli, the language of the Buddhist canon, remained unknown in the West until the beginning of the 19th century, when modern Buddhology was founded.

Reasons for the reticence of contacts with non-Christian religions

The reasons for such reticence toward contact with foreign religions were twofold: (1) The ancient church was significantly influenced by the Jewish attitude toward the pagan religions of its environment. Like Judaism, it viewed the pagan gods as "nothings" next to the true God, the Creator of the world and, in the case of the Christians, the Father of Jesus Christ; they were offsprings of human error that were considered to be identical with the wooden, stone, or bronze images that were made by humans. (2) Besides this, there was the tendency to degrade the pagan gods as demons, evil demonic forces engaged in mortal combat with the true God. The conclusion of the history of salvation, according to the Christian understanding, was to be a final struggle between Christ and his church on the one side and the forces, powers, and thrones of the Antichrist on the other, culminating finally with the victory of Christ.

### CONFLICTING CHRISTIAN ATTITUDES

The history of religion, however, continued even after Christ. During the 3rd and 4th centuries a new non-Christian world religion appeared in the form of Manichaeism, which countered the Christian Church with new holy books, a new institution, and a new universal claim of validity. The Christian Church never acknowledged Manichaeism as a new religion but considered it a Christian heresy and opposed it as such.

When Islām was founded in the 7th century as a new higher religion, it considered revelation as received by the Prophet Muḥammad to be superior to the former levels of Old and New Testament revelation. Christianity also fought Islām as a Christian heresy. This new threat was seen as the fulfillment of the eschatological prophecies of the Apocalypse concerning the coming of the "false prophet" (Revelation to John). The apocalyptic interpretation of Islām as the religion of the "false prophet" also coined the archetypal struggle of the Christian Church of the Middle Ages against foreign religions, namely, the crusade. The idea of the Crusades deeply influenced the self-consciousness of Western Christianity even in later centuries.

The peace of faith

The dialogue of the 15th-century German theologian Nicholas of Cusa on the peace of faith (1453) is the first Christian document that calls for the establishment of an eternal peace among world religions. In spite of this, the idea of the crusade remained the model for the fulfillment of the new missionary task that arose within the Roman Catholic Church with the discovery and exploration of the American continents by Spain and Portugal. Only the penetration of the Islāmic wall that had separated Europe spiritually and economically from the empires of the Asian higher religions and only the encounter with these higher religions in countries such as China and Japan—which could not be subjugated to the rule of Roman Catholic

kings by the sword—led to a gradual overcoming of the idea of the crusade. In China and Japan the missionaries saw themselves forced into an argument with the indigenous higher religions that could be carried on only with intellectual weapons. The old Logos theory prevailed in a new form that was founded on natural law, particularly among the Jesuit theologians who worked at the Chinese emperor's court in Peking.

The philosophy of the Enlightenment in the 18th century spread the acknowledgment of a plurality of higher religions among the educated in Europe, partly—as in the case of the German philosopher Gottfried Wilhelm Leibniz—in immediate connection with the theories of natural law of the Jesuit missionaries in China. This insight pointed to the striking convergence of non-Christian higher religions with Christianity and in that way prepared the development of the comparative study of religion. Only in the philosophy of the Enlightenment was the demand of tolerance, which thus far in Christian Europe had been applied solely as a postulate of behaviour toward the followers of another Christian denomination, extended to include the followers of different religions.

Pietistic missionary activities

The missions that were carried out in the late 18th and the 19th centuries by pietistically or fundamentally oriented churches ignored this knowledge or consciously fought against it. Simple lay Christianity of revivalist congregations demanded that a missionary denounce all pagan "idolatry." The spiritual and intellectual argument with non-Christian higher religions simply did not exist for this simplified fundamental theology, and in this view a real encounter of Christianity with non-Christian higher religions did not, on the whole, occur in the 18th and 19th centuries. (E.W.B./J.Hi.)

### MODERN VIEWS

The 20th century has seen an explosion of publicly available information concerning the wider religious life of humanity, as a result of which the older Western assumption of the manifest superiority of Christianity has lost plausibility in many minds. Early 20th-century thinkers such as Rudolf Otto, who saw religion throughout the world as a response to the Holy, and Ernst Troeltsch, who showed that socioculturally Christianity is one of a number of comparable traditions, opened up new ways of regarding the other major religions.

Exclusivism, inclusivism, and pluralism

Given that the central concern of both Christianity and the other great world faiths is salvation, Christians today adopt one of three main points of view. One is exclusivism, which holds that there is salvation only for Christians. This theology underlay much of the history outlined above, expressed both in the Roman Catholic dogma *extra ecclesiam nulla salus* ("outside the church no salvation") and in the assumption of the 18th- and 19th-century Protestant missionary movements that outside the proclaimed Gospel there is no salvation. The exclusivist outlook was eroded within advanced Roman Catholic thinking in the decades leading up to the second Vatican Council (1962–65) and was finally abandoned in the council's pronouncements. Within Protestant Christianity there is no comparable central authority, but most Protestant theologians, except within the extreme Fundamentalist constituencies, have also moved away from the exclusivist position.

The move, among both Roman Catholics and Protestants, has been toward inclusivism, the view that, although salvation is by definition Christian salvation, brought about by the atoning work of Christ, it is nevertheless in principle available to all human beings, whether Christian or not. The Roman Catholic theologian Karl Rahner expressed the inclusivist view by saying that good and devout people of other faiths may, even without knowing it, be regarded as "anonymous Christians." Others have expressed in different ways the thought that non-Christians also are included within the universal scope of Christ's salvific work and their religions fulfilled in Christianity.

The third position, to which a number of individual theologians have moved in recent years, is pluralism. According to this view, the great world faiths, including Christianity, are valid spheres of a salvation that takes characteristically different forms within each—though consisting in each case in the transformation of human existence from self-centredness to a new orientation toward the Divine Reality. The other religions are thus not secondary contexts of Christian redemption but independently authentic paths of salvation. The pluralist position is controversial in Christian theology because it affects the ways in which the doctrines of the person of Christ, atonement, and the Trinity are formulated.

Christians engage in dialogue with the other major religions through the World Council of Churches' subunit on Dialogue with People of Living Faiths and Ideologies and the Vatican's Secretariat for Non-Christians, as well as a variety of extra-ecclesiastical organizations, such as the World Congress of Faiths. A multitude of interreligious encounters takes place throughout the world, many initiated by Christian and others by non-Christian individuals and groups. (J.Hi.)

### BIBLIOGRAPHY

*General works:* Information on aspects treated in this article is available in DAVID B. BARRETT (ed.), *World Christian Encyclopedia: A Comparative Study of Churches and Religions in the Modern World, AD 1900–2000* (1982); F.L. CROSS and E.A. LIVINGSTONE (eds.), *The Oxford Dictionary of the Christian Church,* 2nd ed. (1974, reprinted 1983); J.D. DOUGLAS (ed.), *The New International Dictionary of the Christian Church,* 2nd ed. (1978); MIRCEA ELIADE (ed.), *The Encyclopedia of Religion,* 16 vol. (1987), with helpful bibliographies; and NEW CATHOLIC ENCYCLOPEDIA, 16 vol. (1967–79, reissued 1981), especially useful for the Roman Catholic Church.

*History of the Christian Church:* Broad overviews are found in KURT ALAND, *A History of Christianity,* 2 vol. (1985–86; originally published in German, 1980–82); ROLAND H. BAINTON, *The Horizon History of Christianity* (1964, reissued with the title *Christianity,* 1985); GEOFFREY BARRACLOUGH (ed.), *The Christian World: A Social and Cultural History* (1981); OWEN CHADWICK, *The Pelican History of the Church,* 6 vol. (1960–70, reprinted 1985–86); PAUL JOHNSON, *A History of Christianity* (1976, reprinted 1985); KENNETH SCOTT LATOURETTE, *A History of Christianity,* rev. ed., 2 vol. (1975); HENRY CHADWICK and G.R. EVANS (eds.), *Atlas of the Christian Church* (1987); and F. VAN DER MEER and CHRISTINE MOHRMANN, *Atlas of the Early Christian World,* trans. from Dutch (1958, reprinted 1966). See also the series edited by HUBERT JEDIN and JOHN DOLAN, *Handbook of Church History,* 10 vol. (1965–81; originally published in German, 7 vol. in 10, 1962–79), later vol. of which have the title *History of the Church.*

Guides to the first five centuries of the Christian Church include LOUIS DUCHESNE, *Early History of the Christian Church: From Its Foundation to the End of the Fifth Century,* trans. from French, 3 vol. (1909–24, reprinted 1957–60); W.H.C. FREND, *The Rise of Christianity* (1984, reprinted 1986); ROBERT M. GRANT, *Augustus to Constantine: The Thrust of the Christian Movement into the Roman World* (1970); ROWAN A. GREER, *Broken Lights and Mended Lives: Theology and Common Life in the Early Church* (1986); ADOLF HARNACK, *The Mission and Expansion of Christianity in the First Three Centuries,* 2nd enl. and rev. ed., 2 vol. (1908; originally published in German, 1902); HANS LIETZMANN, *A History of the Early Church,* 4 vol. (1949–52, reissued 4 vol. in 2, 1964; originally published in German, 4 vol., 1932–44); A.D. NOCK, *Conversion: The Old and the New in Religion from Alexander the Great to Augustine of Hippo* (1933, reprinted 1988); and A.D. NOCK (ed.), *Essays on Religion and the Ancient World,* ed. and comp. by ZEPH STEWART, 2 vol. (1972, reprinted 1986). See also J.M. WALLACE-HADRILL, *The Frankish Church* (1983).

For discussions of more specialized topics, see CECIL JOHN CADOUX, *The Early Church and the World: A History of the Christian Attitude to Pagan Society and the State Down to the Time of Constantine* (1925, reprinted 1955); HENRY CHADWICK, *Early Christian Thought and the Classical Tradition: Studies in Justin, Clement, and Origen* (1966, reprinted 1984); FRANCIS DVORNIK, *Early Christian and Byzantine Political Philosophy: Origins and Background,* 2 vol. (1966); ROBERT M. GRANT, *Early Christianity and Society* (1977); ROBIN LANE FOX, *Pagans and Christians* (1986, reissued 1988); WAYNE A. MEEKS, *The First Urban Christians: The Social World of the Apostle Paul* (1983); ARNALDO MOMIGLIANO, *The Conflict Between Paganism and Christianity in the Fourth Century* (1963, reprinted 1970); PETER RICHARDSON, *Israel in the Apostolic Church* (1969); and J.M. HUSSEY, *The Orthodox Church in the Byzantine Empire* (1986).

In addition to the relevant volumes of the histories cited above, the church in the Middle Ages is studied in HANS-GEORG BECK, *Kirche und theologische Literatur im byzantinis-*

*chen Reich* (1959, reprinted 1977); JOHN BOSSY, *Christianity in the West, 1400–1700* (1985); LOUIS DUCHESNE, *L'Église au VI^e siècle* (1925); JUDITH HERRIN, *The Formation of Christendom* (1987); and STEVEN OZMENT, *The Age of Reform (1250–1550): An Intellectual and Religious History of Late Medieval and Reformation Europe* (1980).

Modern church history is treated in the general histories cited above; in the sections below on roles of Christianity, on missions, and on ecumenism; and in KENNETH SCOTT LATOURETTE, *Christianity in a Revolutionary Age: A History of Christianity in the Nineteenth and Twentieth Centuries,* 5 vol. (1958–62, reissued 1973); JAMES HASTINGS NICHOLS, *History of Christianity, 1650–1950: Secularization of the West* (1956); JERALD C. BRAUER, *Protestantism in America: A Narrative History,* rev. ed. (1965, reprinted 1974); and CHARLES H. LIPPY and PETER W. WILLIAMS (eds.), *Encyclopedia of the American Religious Experience: Studies of Traditions and Movements,* 3 vol. (1988).

*Major traditional doctrinal issues:* (*The meaning of dogma*): KARL BARTH, *Church Dogmatics,* 4 vol. in 12 (1939–59; originally published in German, 4 vol. in 12, 1932–59); EMIL BRUNNER, *Dogmatics,* vol. 3, *The Christian Doctrine of the Church, Faith, and the Consummation* (1960, reissued in a new trans., 1979; originally published in German, 1960); YVES M.J. CONGAR, *A History of Theology,* trans. from French (1968); T.A. BURKILL, *The Evolution of Christian Thought* (1971); JAROSLAV PELIKAN, *The Christian Tradition: A History of the Development of Doctrine* (1971– ), of which 4 vol. had appeared by 1987; PAUL TILLICH, *A History of Christian Thought, from Its Judaic and Hellenistic Origins to Existentialism* (1972); J.N.D. KELLY, *Early Christian Doctrines,* 5th rev. ed. (1977, reprinted 1985); HUBERT CUNLIFFE-JONES (ed.), *A History of Christian Doctrine* (1978, reissued 1980); and WALTER KASPER, *An Introduction to Christian Faith* (1980; originally published in German, 1972).

(*God the Father*): W.R. MATTHEWS, *God in Christian Thought and Experience,* 3rd ed. (1963); H.P. OWEN, *Concepts of Deity* (1971); GORDON D. KAUFMAN, *The Theological Imagination: Constructing the Concept of God* (1981); and WALTER KASPER, *The God of Jesus Christ* (1984, reprinted 1986; originally published in German, 1982).

(*God the Son*): ALBERT SCHWEITZER, *The Quest of the Historical Jesus: A Critical Study of Its Progress from Reimarus to Wrede,* 3rd ed. (1954, reissued 1981; originally published in German, 1906); GÜNTHER BORNKAMM, *Jesus of Nazareth* (1960, reprinted 1975; originally published in German, 1956); EDWARD SCHILLEBEECKX, *Christ, the Sacrament of the Encounter with God* (1963, reprinted 1977; originally published in Dutch, 1960); FREDERICK HOUK BORSCH, *The Son of Man in Myth and History* (1967); EDWARD ROCHIE HARDY (ed.), *Christology of the Later Fathers* (1954, reprinted 1977); JOHN REUMANN, *Jesus in the Church's Gospels: Modern Scholarship and the Earliest Sources* (1968, reprinted 1973); and ALOYS GRILLMEIER, *Christ in Christian Tradition,* vol. 1, *From the Apostolic Age to Chalcedon (451),* trans. from German, 2nd rev. ed. (1975).

(*God the Holy Spirit*): CHARLES WILLIAMS, *The Descent of the Dove: A Short History of the Holy Spirit in the Church* (1939, reissued 1974); HENRY P. VAN DUSEN, *Spirit, Son and Father: Christian Faith in the Light of the Holy Spirit* (1958, reissued 1960); FREDERICK DALE BRUNER, *A Theology of the Holy Spirit: The Pentacostal Experience and the New Testament Witness* (1970, reprinted 1973); GEORGE T. MONTAGUE, *The Holy Spirit: Growth of a Biblical Tradition* (1976); KARL RAHNER, *The Spirit in the Church* (1979; originally published in German, 1977); YVES M.J. CONGAR, *I Believe in the Holy Spirit,* 3 vol. (1983; originally published in French, 1979–80); and ALASDAIR I.C. HERON, *The Holy Spirit* (1983).

(*The Holy Trinity*): JULES LEBRETON, *History of the Dogma of the Trinity: From Its Origins to the Council of Nicaea* (1939; originally published in French, 8th ed., 1927); KARL RAHNER, *The Trinity,* trans. from German (1970); EDMUND J. FORTMAN, *The Triune God: A Historical Study of the Doctrine of the Trinity* (1972); ROBERT W. JENSON, *The Triune Identity: God According to the Gospel* (1982); and MICHAEL O'CARROLL, *Trinitas: A Theological Encyclopedia of the Holy Trinity* (1987).

(*The concept of man*): EMIL BRUNNER, *Man in Revolt: A Christian Anthropology* (1939, reissued 1957; originally published in German, 1937); REINHOLD NIEBUHR, *The Nature and Destiny of Man: A Christian Interpretation,* 2 vol. (1941–43, reprinted 1964); H. WHEELER ROBINSON, *The Christian Doctrine of Man,* 4th ed. (1958, reprinted 1974); WERNER G. KÜMMEL, *Man in the New Testament,* rev. and enl. ed. (1963; originally published in German, 1948); ERNST BENZ, "The Concept of Man in Christian Thought," in S. RADHAKRISHNAN and P.T. RAJU (eds.), *The Concept of Man: A Study in Comparative Philosophy,* 2nd ed. (1966, reprinted 1972), pp. 394–451; and WOLFHART PANNENBERG, *What Is Man?: Contemporary Anthropology in Theological Perspective* (1970).

(*The church*): Works on various aspects of church doctrine include, on the church, GEORGE JOHNSTON, *The Doctrine of the Church in the New Testament* (1943); HANS KÜNG, *The Church* (1967, reissued 1976; originally published in German, 1967); DIETRICH BONHOEFFER, *Sanctorum Communio: Eine dogmatische Untersuchung zur Soziologie der Kirche,* 4th ed. (1969); and EINAR MOLLAND, *Christendom: The Christian Churches, Their Doctrines, Constitutional Forms, and Ways of Worship* (1971); on the formation of the biblical canon, HANS VON CAMPENHAUSEN, *The Formation of the Christian Bible* (1972, reissued 1977; originally published in German, 1968); HARRY Y. GAMBLE, *The New Testament Canon: Its Making and Meaning* (1985); and BRUCE M. METZGER, *The Canon of the New Testament: Its Origin, Development, and Significance* (1987); on Christian creeds and confessions, PHILIP SCHAFF, *Biblioteca symbolica ecclesiae universalis: The Creeds of Christendom,* 6th ed., 3 vol. (1919, reprinted 1985); B.A. GERRISH (ed.), *The Faith of Christendom: A Source Book of Creeds and Confessions* (1963); J.N.D. KELLY, *Early Christian Creeds,* 3rd ed. (1972, reprinted 1981), and a companion volume, *The Athanasian Creed* (1964); and JOHN H. LEITH (ed.), *Creeds of the Churches: A Reader in Christian Doctrine, from the Bible to the Present,* 3rd ed. (1982); on the apostolic succession, ERNST BENZ, *Bischofsamt und apostolische Sukzession im deutschen Protestantismus* (1953); on church polity and structure, JAMES VERNON BARTLET, *Church-Life and Church-Order During the First Four Centuries* (1943); on the liturgy, JOSEF A. JUNGMANN, *The Early Liturgy, to the Time of Gregory the Great,* trans. from German (1959); THEODOR KLAUSER, *A Short History of the Western Liturgy: An Account and Some Reflections,* 2nd ed. (1979; originally published in German, 1965); JAMES F. WHITE, *Introduction to Christian Worship* (1980); and HERMAN A.J. WEGMAN, *Christian Worship in East and West: A Study Guide to Liturgical History* (1985; originally published in Dutch, 1976); on Christian tradition, DANIEL T. JENKINS, *Tradition, Freedom, and the Spirit* (U.K. title, *Tradition and the Spirit,* 1951); and F.W. DILLISTONE (ed.), *Scripture and Tradition* (1955); on monasticism, CUTHBERT BUTLER, *Benedictine Monachism: Studies in Benedictine Life and Rule,* 2nd ed. (1924, reprinted 1962); DAVID KNOWLES, *Christian Monasticism* (1969); and JEAN LECLERCQ, *The Love of Learning and the Desire for God: A Study of Monastic Culture,* 3rd ed. (1982); and on Christian art and iconography, EMILE MÂLE, *Religious Art from the Twelfth to the Eighteenth Century* (1949, reissued 1970; originally published in French, 1945); JOHN G. DAVIES, *The Origin and Development of Early Christian Church Architecture* (1952); JANE DILLENBERGER, *Style and Content in Christian Art* (1965, reissued 1986); LEONID OUSPENSKY and VLADIMIR LOSSKY, *The Meaning of Icons,* 2nd ed. (1982; originally published in German, 1952); and ROBERT MILBURN, *Early Christian Art and Architecture* (1987).

(*Last things*): PAUL S. MINEAR, *Christian Hope and the Second Coming* (1954); RUDOLF BULTMANN, *History and Eschatology* (U.S. title, *The Presence of Eternity,* 1957, reissued 1975); OSCAR CULLMANN, *Christ and Time: The Primitive Christian Conception of Time and History,* rev. ed. (1962, reprinted 1964; originally published in German, 1946); JÜRGEN MOLTMANN, *Theology of Hope: On the Ground and the Implications of a Christian Eschatology* (1967, reprinted 1975; originally published in German, 1964), and *Hope and Planning,* trans. from German (1971); WILLIAM STRAWSON, *Jesus and the Future Life,* new and rev. ed. (1970); and GEOFFREY WAINWRIGHT, *Eucharist and Eschatology,* 2nd ed. (1978, reprinted 1981).

*Church year:* LOUIS DUCHESNE, *Christian Worship: Its Origin and Evolution,* 5th ed. (1919, reprinted 1956; originally published in French, 4th rev. and enl. ed., 1908), ch. 8, is fundamental but should be supplemented by later handbooks, such as J.A. JUNGMANN, *Public Worship* (1957, reissued 1966; originally published in German, 1955), ch. 9; and JOHN H. MILLER, *Fundamentals of the Liturgy* (1960, reprinted 1964), ch. 8. Good summary accounts are those of NOËLE M. DENIS-BOULET, *The Christian Calendar* (1960; originally published in French, 1959); A. ALAN MCARTHUR, *The Evolution of the Christian Year* (1953); and ADOLF ADAM, *The Liturgical Year: Its History & Its Meaning After the Reform of the Liturgy* (1981), basic for present Roman Catholic use. THOMAS J. TALLEY, *The Origins of the Liturgical Year* (1986), is a fresh reading of the early evidence. More popular treatments, valuable for their detail of popular observance, are the works of FRANCIS X. WEISER, *The Christmas Book* (1952, reissued 1954), *The Easter Book* (1954), *The Holyday Book* (1956), and *Handbook of Christian Feasts and Customs: The Year of the Lord in Liturgy and Folklore* (1958). See also SUE SAMUELSON, *Christmas: An Annotated Bibliography* (1982). Insights into primitive and non-Christian backgrounds of the church year are contained in MIRCEA ELIADE, *The Myth of the Eternal Return* (1954, reprinted 1974; originally published in French, 1949); and E.O. JAMES, *Seasonal Feasts and Festivals* (1961, reprinted 1963). For Jewish background, see ROLAND DE VAUX, *Ancient Israel: Its Life and Institutions* (1961, reissued 1973; originally published in

French, 2 vol., 1958–60), pt. 4, ch. 15–18. A standard monograph on the origin of the seven-day week is F.H. COLSON, *The Week* (1926, reprinted 1974). EVIATAR ZERUBAVEL, *The Seven Day Circle: The History and Meaning of the Week* (1985), is a comprehensive historical, cultural, and sociological study. More exhaustive and detailed is WILLY RORDORF, *Sunday: The History of the Day of Rest and Worship in the Earliest Centuries of the Christian Church* (1968; originally published in German, 1962). For historical discussions of Holy Week and Easter, see MASSEY H. SHEPHERD, JR., *The Paschal Liturgy and the Apocalypse* (1960); and JOHN WALTON TYRER, *Historical Survey of Holy Week: Its Services and Ceremonial* (1932). CLARENCE SEIDENSPINNER, *Great Protestant Festivals* (1952), defends nontraditional observances in modern Protestant churches.

*Canon law:* A short introduction is provided by LADISLAS M. ORSY, *From Vision to Legislation: From the Council to a Code of Laws* (1985). The most important modern works on the history of canon law are HANS E. FEINE, *Kirchliche Rechtsgeschichte: Die katholische Kirche,* 5th ed. (1972); WILLIBALD M. PLÖCHLE, *Geschichte des Kirchenrechts,* 2nd enl. ed., 5 vol. (1960–70); and GABRIEL LE BRAS (ed.), *Histoire du droit et des institutions de l'eglise en occident* (1955– ), with 12 vol. published by 1987. The articles in R. NAZ (ed.), *Dictionnaire de droit canonique,* 7 vol. (1935–65), although in need of updating, can still provide much historical and doctrinal information. A respected analytical commentary on the 1917 code is A. VERMEERSCH and J. CREUSEN, *Epitome Iuris Canonici,* 7th ed. rev. by R.P. CREUSEN, 3 vol. (1949–56), in Latin. Commentaries on the 1983 code include JAMES A. CORIDEN, THOMAS J. GREEN, and DONALD E. HEINTSCHEL (eds.), *The Code of Canon Law: A Text and Commentary* (1985), intended mainly for practitioners; JOSEPH LISTL, HUBERT MÜLLER, and HERIBERT SCHMITZ (eds.), *Handbuch des katholischen Kirchenrechts* (1983), a doctrinally elaborate, thematically organized work; and KLAUS LÜDICKE (ed.), *Münsterischer Kommentar zum Codex Iuris Canonici* (1985– ), kept up to date by frequent loose-leaf supplements concerning official pronouncements and current literature. Bibliographies are found in the series "Repertoire Bibliographique des Institutions Chrétiennes" (1969– ), with text in French, English, German, Italian, and Spanish.

*Patristic literature:* The most important texts on the Church Fathers are BERTOLD ALTANER, *Patrology* (1960; originally published in German, 1938); F.L. CROSS, *The Early Christian Fathers* (1960); and JOHANNES QUASTEN, *Patrology,* 3 vol. (1950–60, reprinted 1983), continued by ANGELO DI BERARDINO (ed.), *Patrology: The Golden Age of Latin Patristic Literature from the Council of Nicea to the Council of Chalcedon* (1986; originally published in Italian, 1978). Other works include HANS VON CAMPENHAUSEN, *The Fathers of the Greek Church* (1959, reissued 1963; originally published in German, 1955), and *The Fathers of the Latin Church* (1964, reprinted 1969; originally published in German, 1960); FRANCES M. YOUNG, *From Nicaea to Chalcedon: A Guide to the Literature and Its Background* (1983), on the chief Greek Fathers from 325 to 451; and BONIFACE RAMSEY, *Beginning to Read the Fathers* (1985). See also the volumes in the series "Message of the Fathers of the Church" (1983– ).

*Christian philosophy:* On the early period, CLAUDE TRESMONTANT, *The Origins of Christian Philosophy* (1962; originally published in French, 1962); A.H. ARMSTRONG and R.A. MARKUS, *Christian Faith and Greek Philosophy* (1960, reissued 1964); and ADAM FOX (ed. and trans.), *Plato and the Christians* (1957), cover the Hebraic and Greek sources of Christian thought. The classic works on the medieval period are ÉTIENNE GILSON, *The Spirit of Mediaeval Philosophy* (1936; originally published in French, 2 vol., 1932), *Reason and Revelation in the Middle Ages* (1938, reprinted 1966), and *History of Christian Philosophy in the Middle Ages* (1955, reissued 1980). See also PHILIPPE DELHAYE, *Medieval Christian Philosophy* (1960; originally published in French, 1959). J.V. LANGMEAD CASSERLEY, *The Christian in Philosophy* (1949, reissued 1955); and GEORGE F. THOMAS, *Religious Philosophies of the West* (1965), and *Philosophy and Religious Belief* (1970), provide broad surveys up to the contemporary period. For existentialism, see DAVID E. ROBERTS, *Existentialism and Religious Belief* (1957). Much of the contemporary discussion takes place in articles, such as those collected in BASIL MITCHELL (ed.), *The Philosophy of Religion* (1971, reprinted 1978), all of whose contributors are Christian philosophers grappling with current issues. Also useful are STUART C. BROWN (ed.), *Reason and Religion* (1977); RICHARD SWINBURNE, *The Existence of God* (1979); HANS KÜNG, *Does God Exist?: An Answer for Today* (1980; originally published in German, 1978); ALVIN PLANTINGA and NICHOLAS WOLTERSTORFF (eds.), *Faith and Rationality: Reason and Belief in God* (1983); and LEROY S. ROUNER (ed.), *Religious Pluralism* (1984). *Faith and Philosophy: Journal of the Society of Christian Philosophers* (quarterly) is another forum of contemporary discussion.

*Mysticism:* General descriptive approaches include EVELYN UNDERHILL, *Mysticism: A Study in the Nature and Development of Man's Spiritual Consciousness* (1911, reissued 1977); R.C. ZAEHNER, *Mysticism, Sacred and Profane: An Inquiry Into Some Varieties of Praeter-Natural Experience* (1957, reissued 1980); SIDNEY SPENCER, *Mysticism in World Religion* (1963, reissued 1971); and ANDREW LOUTH, *The Origins of the Christian Mystical Tradition from Plato to Denys* (1981, reprinted 1983). A good anthology is that of ELMER O'BRIEN, *Varieties of Mystic Experience* (1964). Many of the texts of the great Christian mystics have been published in new translations in "The Classics of Western Spirituality" series (1978– ). Helpful for the serious student are the classic works of WILLIAM JAMES, *The Varieties of Religious Experience: A Study in Human Nature* (1902, reissued 1985); FRIEDRICH VON HÜGEL, *The Mystical Element of Religion: As Studied in Saint Catherine of Genoa and Her Friends,* 2 vol. (1908, reprinted 1961); CUTHBERT BUTLER, *Western Mysticism: The Teaching of SS Augustine, Gregory, and Bernard on Contemplation and the Contemplative Life* (1922, reprinted 1975); JOSEPH MARÉCHAL, *Studies in the Psychology of the Mystics* (1927; originally published in French, 2 vol., 1924–37); ALBERT SCHWEITZER, *The Mysticism of Paul the Apostle* (1931, reissued 1968; originally published in German, 1930); HENRI BERGSON, *The Two Sources of Morality and Religion* (1935; originally published in French, 1932); RUDOLF OTTO, *Mysticism East and West: A Comparative Analysis of the Nature of Mysticism* (1932, reissued 1987; originally published in German, 1926); JACQUES MARITAIN, *Distinguish to Unite: or, The Degrees of Knowledge* (1959; originally published in French, 4th ed., 1946). For recent issues, see STEVEN T. KATZ (ed.), *Mysticism and Philosophical Analysis* (1978); and PHILLIP C. ALMOND, *Mystical Experience and Religious Doctrine: An Investigation of the Study of Mysticism in World Religions* (1982). The most important recent theological contributions have been those of KARL RAHNER, *The Practice of Faith: A Handbook of Contemporary Spirituality* (1983, reprinted 1986; originally published in German, 1982); and HANS URS VON BALTHASAR, *The Glory of the Lord: A Theological Aesthetics,* trans. from German (1981– ).

*Christian myth and legend:* On the nature of myth, see MIRCEA ELIADE, *Myth and Reality* (1963, reprinted 1975; originally published in French, 1963). WILLIAM G. DOTY, *Mythography: The Study of Myths and Rituals* (1986), analyzes a number of important approaches to the study of myth and, in addition, offers extensive bibliographies. Resistance to myth and legend in early Christianity is described in WALTER BAUER, *Orthodoxy and Heresy in Earliest Christianity* (1971, reprinted 1979; originally published in German, 1934). MIRCEA ELIADE, *A History of Religious Ideas,* 3 vol. (1978–85; originally published in French, 3 vol., 1976–83), discusses Christian myth and legend in several chapters.

On the androgyny of Christ, see WAYNE A. MEEKS, "The Image of the Androgyne: Some Uses of a Symbol in Earliest Christianity," *History of Religions,* 13(3):165–208 (Feb. 1974); and CAROLINE WALKER BYNUM, *Jesus as Mother: Studies in the Spirituality of the High Middle Ages* (1982, reprinted 1984). GEO WIDENGREN, *Mesopotamian Elements in Manichaeism (King and Saviour II): Studies in Manichaean, Mandaean, and Syrian-Gnostic Religion* (1946); and ROBERT MURRAY, *Symbols of Church and Kingdom: A Study in Early Syriac Tradition* (1975, reprinted 1977), discuss the origins of sacramental oils. HUGO RAHNER, *Greek Myths and Christian Mystery* (1963, reissued 1971; originally published in German, 1945), traces the theme of the Christian World Tree.

For apocryphal gospels, see MORTON SMITH, *The Secret Gospel: The Discovery and Interpretation of the Secret Gospel According to Mark* (1973, reissued 1982); JAMES M. ROBINSON (comp.), *The Nag Hammadi Library in English* (1977); JACQUES E. MÉNARD, *L'Évangile selon Philippe* (1964, reissued 1967), and *L'Évangile selon Thomas* (1975); and the series "Bibliothèque copte de Nag Hammadi: section textes" (1977– ). See also CHARLES W. HEDRICK and ROBERT HODGSON, JR. (eds.), *Nag Hammadi, Gnosticism, & Early Christianity* (1986). For the *Protogospel of James,* the *Chronicle of Zugnin,* and the *Opus Imperfectum in Matthaeum,* see UGO MONNERET DE VILLARD, *Le leggende orientali sui magi evangelici* (1952). On Bogomil and Cathar apocrypha, see ÉMILE TURDEANU, "Apocryphes bogomiles et apocryphes pseudo-bogomiles," *Revue de l'Histoire des Religions* 138:22–52, 176–218 (1950); and EDINA BOZÓKY (ed. and trans.), *Le Livre secret des cathares, interrogatio Iohannis: apocryphe d'origine bogomile* (1980).

On the cult of saints, see H. DELEHAYE, *The Legends of the Saints: An Introduction to Hagiography* (1907, reprinted 1974; originally published in French, 1905); LAWRENCE S. CUNNINGHAM, *The Meaning of Saints* (1980); PETER BROWN, *The Cult of the Saints: Its Rise and Function in Latin Christianity* (1981); DONALD WEINSTEIN and RUDOLPH M. BELL, *Saints & Society: The Two Worlds of Western Christendom, 1000–1700* (1982,

reprinted 1986); and STEPHEN WILSON (ed.), *Saints and Their Cults: Studies in Religious Sociology, Folklore, and History* (1983, reprinted 1985). See also JOHN J. DELANEY, *Dictionary of Saints* (1980); and DAVID HUGH FARMER, *The Oxford Dictionary of Saints,* 2nd ed. (1987), on Irish and English saints.

HENRY KAHANE and RENÉE KAHANE, *The Krater and the Grail: Hermetic Sources of the Parzival* (1965, reprinted 1984), analyzes Wolfram von Eschenbach's *Parzival.* For the Arthurian cycles, see P.B. GROUT *et al.* (eds.), *The Legend of Arthur in the Middle Ages* (1983); and ALFRED NUTT, *Studies on the Legend of the Holy Grail: With Especial References to the Hypothesis of Its Celtic Origin* (1888, reissued 1967).

Christian alchemy is described in C.G. JUNG, *Psychology and Alchemy,* 2nd ed. rev. (1968, reprinted 1980; originally published in German, 2nd rev. ed., 1952). Symbolic astronomy and the mutual refiguration of Christian and pagan legends are treated in a 15th-century text, IOAN P. COULIANO (CULIANU), *Eros and Magic in the Renaissance,* trans. from French (1987). On the alchemical researches of Enlightenment scientists, especially physicists and chemists, see FRANCES A. YATES, *The Rosicrucian Enlightenment* (1972, reissued 1986); and BETTY JO TEETER DOBBS, *The Foundations of Newton's Alchemy: or, "The Hunting of the Greene Lyon"* (1975, reprinted 1983).

Two examples of non-Western materials are ROGER BASTIDE, *The African Religions of Brazil: Toward a Sociology of the Interpenetration of Civilizations* (1978; originally published in French, 1960), especially pp. 260–84 on Afro-Brazilian Christianity; and MARC DE CIVRIEUX, *Watunna: An Orinoco Creation Cycle,* trans. from Spanish (1980).

*Roles of Christianity:* The relation of the Christian community to the world is discussed in ERNST TROELTSCH, *The Social Teaching of the Christian Churches,* 2 vol. (1931, reprinted 1981; originally published in German, 1912), dated in specifics but still one of the most comprehensive and influential studies of this topic; H. RICHARD NIEBUHR, *Christ and Culture* (1951, reprinted 1975); PAUL TILLICH, *Theology of Culture* (1959, reprinted 1978), essays on philosophy, art, literature, and science; and PETER L. BERGER, *The Sacred Canopy: Elements of a Sociological Theory of Religion* (1967, reprinted 1969; U.K. title, *The Social Reality of Religion,* 1969, reissued 1973). Works on various aspects of Christianity's intersection with the world include, on pastoral care, WILLIAM A. CLEBSCH and CHARLES R. JAEKLE, *Pastoral Care in Historical Perspective* (1964, reissued 1983), including excerpts from primary sources in the history of the church; and RONALD L. NUMBERS and DARREL W. AMUNDSEN (eds.), *Caring and Curing: Health and Medicine in the Western Religious Traditions* (1986), a unique and comprehensive presentation by scholars of various faith traditions, with bibliographies; on poverty, MICHEL MOLLAT (ed.), *Études sur l'histoire de la pauvreté,* 2 vol. (1974), a collection of essays on the history of the church's understanding of poverty from the early church to the modern period—each essay with an English abstract; and CARTER LINDBERG, "Through a Glass Darkly: A History of the Church's Vision of the Poor and Poverty, *The Ecumenical Review,* 33(1):37–52 (Jan. 1981); on birth control, JOHN T. NOONAN, JR., *Contraception: A History of Its Treatment by the Catholic Theologians and Canonists,* enl. ed. (1986); on the concept of love, ANDERS NYGREN, *Agape and Eros,* 2 vol. in 3 (1932–39, reissued 1982; originally published in Swedish, 2 vol., 1930–36); on black theology, GAYRAUD S. WILMORE and JAMES H. CONE (eds.), *Black Theology: A Documentary History, 1966–1979* (1979); and JAMES H. CONE, *For My People: Black Theology and the Black Church* (1984); on liberation theology, DEANE WILLIAM FERM, *Third World Liberation Theologies: An Introductory Survey* (1986); and LEONARDO BOFF and CLODOVIS BOFF, *Introducing Liberation Theology* (1987; originally published in Portuguese, 1986); and on feminist theology, ROSEMARY RADFORD RUETHER, *Sexism and God-Talk: Toward a Feminist Theology* (1983); and LETTY M. RUSSELL (ed.), *Feminist Interpretation of the Bible* (1985).

*Missions:* DAVID B. BARRETT (ed.), *World Christian Encyclopedia: A Comparative Study of Churches and Religions in the Modern World* (1982), is comprehensive and indispensable. See also STEPHEN NEILL, GERALD ANDERSON, and JOHN GOODWIN (eds.), *Concise Dictionary of the Christian World Mission* (1971); and DON M. MCCURRY, *World Christianity* (1979– ), with 5 volumes appearing by 1987 on the Middle East, eastern Asia, South Asia, Central America and the Caribbean, and Oceania. KENNETH SCOTT LATOURETTE, *A History of the Expansion of Christianity,* 7 vol. (1937–45, reprinted 1971), is a pioneering classic. STEPHEN NEILL, *A History of Christian Missions,* 2nd ed. rev. by OWEN CHADWICK (1986), is a lively, engaging work. WALTER M. ABBOTT (ed.), *The Documents of Vatican II* (1966), includes the relevant texts. POPE PAUL VI, *On Evangelization in the Modern World* (1975), addresses post-Vatican II debates. R. PIERCE BEAVER (ed.), *American Missions in Bicentennial Perspective* (1977), is a collection of interpretive essays. Volumes in the series "Mission Trends," ed. by GERALD H. ANDERSON and THOMAS F. STRANSKY (1974– ), include discussions of current issues, evangelization, Third World theologies, North American and European liberation theologies, and Christianity and other religions. See also MARCELLO DE CARVALHO AZEVEDO, *Inculturation and the Challenges of Modernity* (1982). Useful journals include *International Review of Mission* (quarterly); and *International Bulletin of Missionary Research* (quarterly), which annually updates the statistics in the *World Christian Encyclopedia.*

*Ecumenism:* Introductions to the topic are provided by PAUL A. CROW, JR., *Christian Unity: Matrix for Mission* (1982); NORMAN GOODALL, *The Ecumenical Movement: What It Is and What It Does,* 2nd ed. (1964); ERNST LANGE, *And Yet It Moves: Dream and Reality of the Ecumenical Movement,* trans. from Swedish (1979); JOHN T. MCNEILL, *Unitive Protestantism: The Ecumenical Spirit and Its Persistent Expression,* rev. ed. (1964); and CHARLES CLAYTON MORRISON, *The Unfinished Reformation* (1953). Historical overviews can be found in MARC BOEGNER, *The Long Road to Unity* (1970; originally published in French, 1968); ROBERT MCAFEE BROWN, *The Ecumenical Revolution: An Interpretation of the Catholic-Protestant Dialogue,* rev. and expanded ed. (1969); WILLIAM ADAMS BROWN, *Toward a United Church: Three Decades of Ecumenical Christianity* (1946); SAMUEL MCCREA CAVERT, *The American Churches in the Ecumenical Movement, 1900–1968* (1968), and *Church Cooperation and Unity in America: A Historical Review: 1900–1970* (1970); HAROLD E. FEY (ed.), *The Ecumenical Advance,* 2nd ed. (1986); AUSTIN FLANNERY (ed.), *Vatican Council II: The Conciliar and Post-Conciliar Documents* (1975, reissued 1984), and a companion volume, *Vatican Council II: More Postconciliar Documents* (1982); NORMAN GOODALL, *Ecumenical Progress: A Decade of Change in the Ecumenical Movement, 1961–71* (1972); WILLIAM RICHEY HOGG, *Ecumenical Foundations: A History of the International Missionary Council and Its Nineteenth Century Background* (1952); HARDING MEYER and LUKAS VISCHER (eds.), *Growth in Agreement: Reports and Agreed Statements of Ecumenical Conversations on a World Level* (1984); CONSTANTIN G. PATELOS (ed.), *The Orthodox Church in the Ecumenical Movement: Documents and Statements, 1902–1975* (1978); RUTH ROUSE and STEPHEN NEILL (eds.), *A History of the Ecumenical Movement, 1517–1948,* 3rd ed. (1986); BARRY TILL, *The Churches Search for Unity* (1972); THOMAS F. TORRANCE, "Ecumenism: A Reappraisal of Its Significance, Past, Present and Future," in his *Theology in Reconciliation: Essays Towards Evangelical and Catholic Unity in East and West* (1975); HENRY PITNEY VAN DUSEN, *One Great Ground of Hope: Christian Missions and Christian Unity* (1961); MAURICE VILLAIN, *Unity: A History and Some Reflections* (1963; originally published in French, 3rd rev. and augmented ed., 1961); W.A. VISSER 'T HOOFT, *Memoirs* (1973), and *The Genesis and Formation of the World Council of Churches* (1982); and HANS-RUEDI WEBER, *Asia and the Ecumenical Movement, 1895–1961* (1966).

*The Christian Church and non-Christian religions:* The most comprehensive and up-to-date survey of Christian attitudes toward the world religions is PAUL F. KNITTER, *No Other Name?* (1985). A wide range of views is reflected in JOHN HICK and BRIAN HEBBLETHWAITE (eds.), *Christianity and Other Religions: Selected Readings* (1980); and GERALD H. ANDERSON and THOMAS F. STRANSKY (eds.), *Christ's Lordship and Religious Pluralism* (1981). The classic modern statement of a conservative position is that of HENDRICK KRAEMER, *The Christian Message in a Non-Christian World,* 3rd ed. (1956, reprinted 1969). S.J. SAMARTHA (ed.), *Faith in the Midst of Faiths: Reflections on Dialogue in Community* (1977), was produced by the World Council of Churches. ARNULF CAMPS, *Partners in Dialogue: Christianity and Other World Religions* (1983; originally published in Dutch, 3 vol., 1976–78); and HANS KÜNG *et al., Christianity and the World Religions: Paths of Dialogue with Islam, Hinduism, and Buddhism* (1986; originally published in German, 1984), represent different contemporary Roman Catholic standpoints. The pluralistic option is expressed in, for example, WILFRED CANTWELL SMITH, *Towards a World Theology: Faith and the Comparative History of Religion* (1981); JOHN HICK, *God Has Many Names* (1982); and JOHN HICK and PAUL F. KNITTER (eds.), *The Myth of Christian Uniqueness: Toward a Pluralistic Theology of Religions* (1987).

(M.E.M./H.Cha./J.J.Pe./E.W.B./M.H.S./L.M.Ö/J.N.D.K./J.Hi./B.J.McG./L.E.S./C.H.L./W.R.H./P.A.C.)

# Chungking

Chungking (Ch'ung-ch'ing in Wade–Giles; Chongqing in Pinyin), the largest city in Szechwan Province and the leading river port and industrial centre in southwestern China, is 1,400 miles (2,250 kilometres) from the sea, at the confluence of the Yangtze and Chia-ling rivers. During World War II it was the capital of Nationalist China. The city was named Ch'ung-ch'ing ("Double-Blessed") in 1188 under the Nan (Southern) Sung dynasty (AD 1127–1279) because of its commanding position between the cities of Shun-ch'ing (modern Nan-ch'ung) to the north and Shao-ch'ing (modern P'eng-shui) to the south. The city proper includes the Old City and adjacent areas, while the much larger Chungking Municipality (Ch'ung-ch'ing Shih) comprises several counties and a number of lesser cities.

This article is divided into the following sections:


Physical and human geography 367
- The landscape 367
  - The city site
  - Climate
  - The city plan
- The people 368
- The economy 368
  - Industry
  - Trade
  - Transportation
- Administration and social conditions 369
  - Government
  - Public utilities
  - Health
  - Education
- Cultural life 369

History 369
- The early period 369
- The modern period 370

Bibliography 370


## Physical and human geography

### THE LANDSCAPE

**The city site.** Chungking is built on and around Mount Chin-pi (Chin-pi Shan), a hilly promontory of red Jurassic sandstone and shale, which reaches a maximum elevation of about 900 feet (275 metres) above sea level. The promontory is bounded on the north by the Chia-ling River (Chia-ling Chiang), with the industrial area of Chiang-pei on its north bank, and on the east and south by the Yangtze. Other hills, southern offshoots of the Hua-ying Mountains (Hua-ying Shan), rise in the city's outskirts and suburbs.

**Climate.** Chungking is noted for its mild and intensely humid climate. It is shielded from the cold northern winds by the Tsinling Mountains and has little or no frost or ice in winter; the mean temperatures in January and February, the only cool months, are 47° F (8° C) and 50° F (10° C), respectively. Summer, which lasts from May through September, is hot and humid; the August mean temperature is 84° F (29° C), and on many days the high temperature exceeds 100° F (38° C). The remaining months are warm, with mean temperatures ranging between 58° and 67° F (14° and 19° C).

The bulk of rain falls from April through October; the average annual total is about 43 inches (1,087 millimetres). Because of the high humidity, fog and mist are particularly heavy. From October to April the city is perpetually blanketed by fog, which hampers inland navigation, aviation, and local traffic. Chungking's climate has earned the city the nickname "furnace of the Yangtze." The aptness of this name has only increased under the conditions of modernity: contaminated by soot, carbon dioxide, and acid rain, the atmosphere of Chungking is among the most polluted of any city in China.

**The city plan.** *Layout.* The Old City of Chungking (formerly surrounded by a city wall and gates, of which only the names now remain) occupies the eastern third of the rocky promontory and covers an area of about 28 square miles (73 square kilometres). The south and east slopes facing the waterfront form the "lower city," while the remainder is the "upper city." An east–west avenue runs through the middle of each of these areas, and a third runs atop the spine of the promontory's ridge. Cross streets are narrower and often winding; following the topography of the hill, some of them go up and down in flights of hundreds of steps. Chungking's main business district is located around the Liberation Monument (Chieh-fang Pei) in the centre of the Old City.

New districts and suburbs

The new sections of the city on the western part of the promontory spread far along the banks of the two rivers, covering an area considerably larger than the Old City. During World War II the offices of the Nationalist government were located there, and they are now the sites of government office buildings and of museums and exhibition halls. The city has grown so much that the incorporation of numerous industrial towns and suburban communities has extended the city limits to Pei-p'ei in the north and to Pai-shih-i in the southwest. Equally important are the suburban areas on the south shore of the Yangtze. In former times, ferries were the only means by which the rivers could be crossed; now they also may be crossed by way of the Chia-ling Bridge (1966) to the northwest and the Chungking Yangtze Bridge (1980) to the south. A cableway across the Chia-ling River links the Old City with Chiang-pei. The spacious gardens and beautiful residences of the suburban areas contribute much to relieving the crowded conditions of the Old City.

*Housing.* Before World War II Chungking was a city of narrow streets and crowded housing. Streets and lanes followed the contours of the hills. The houses were constructed of bamboo, wood, or thatch in the poorer residential areas and of brick in the wealthier areas. In all areas there was a high degree of congestion. A vigorous modernization program was introduced when the city became the seat of the Nationalist government. Part of the city wall was demolished to make way for new streets, and existing streets were graded and widened. The tremendous demand for housing created by an influx of government workers and refugees led to the rapid expansion of the sections west of the Old City.

From 1938 to 1942 Chungking was heavily bombarded by the Japanese, causing massive destruction in the city. Parts of the wall and virtually all of the city's historic monuments and temples were damaged or destroyed. Because of the destruction the new Communist government (which came to power in 1949) had little difficulty in carrying forward the tasks of modernization and expansion after the war. Modern buildings now stand throughout the city. In the northern suburbs and adjacent areas large buildings provide living quarters for workers and accommodations for factories and workshops. The large brick apartments, generally four to six stories high, are surrounded by trees and vegetable gardens. Yet Chungking remains a city of striking contrasts. Houses of traditional design, blackened by weather, are still to be found on steep hills and along the highways to the suburbs. In many places bamboo structures still line the river bluffs.



Residential area and (right) the town hall in Chungking.
Liu Shilin—New China Pictures Co./Eastfoto

*Outlying suburbs.* In contrast to the congested conditions in the city and the industrial districts, the outlying suburbs have a number of delightful resorts and spas. Among the scenic spots on the south shore are the temple in honour of Empress Yü, consort of the Hsia dynasty emperor of the same name, on Mount Tu; the wooded summer resorts of Ch'ing-shui-ch'i ("Clear Water Creek") and Yang-t'ien-wo ("Sky-gazing Hollow") on Mount Huang; and Nan-wen-ch'üan ("South Hot Springs"), which has delightful retreats at Hua-ch'i ("Flower Creek") and Hu-hsiao-k'ou ("Tiger Roar Gap"). A short distance north of the city are the springs of Ko-lo-shan. Farther up the Chia-ling River at Pei-p'ei are the Chin-yün Shih Temple, the celebrated retreat of the Sung dynasty savant Feng Chin-yün, and Pei-wen-ch'üan ("North Hot Springs"), reputedly superior to the South Hot Springs because its water is warmer in winter and cooler in summer.

Resorts and spas

## THE PEOPLE

Before the war with Japan, Chungking had fewer than 250,000 inhabitants. From 1938 onward, people from the Japanese-occupied coastal provinces flocked to the wartime capital at an astonishing rate. A part of Chungking's population increase since 1938 has consisted of government workers, factory personnel, and refugees from other provinces. (In the late 1940s, however, the city's population decreased temporarily with the return of people to the coastal provinces.) The influx of people from downriver has contributed to turning formerly parochial Chungking into a cosmopolitan city; the number of people living in the city proper alone is now more than 10 times greater than the population of the Old City before the war. The Szechwan dialect, despite its heavy accent and many regional slang words, is quite intelligible to speakers of Mandarin.

## THE ECONOMY

**Industry.** As early as the middle of the Ming dynasty (1368–1644) workshops for spinning, weaving, silk reeling, and brewing were established in Chungking. The city was opened to foreign trade in 1890, and two metal mills were set up a year later. By 1905 Chungking had spinning and weaving mills, silk-reeling mills, and glassmaking and cigarette plants.

The foundations of Chungking's modern industry were laid between 1938 and 1945, when factories transplanted from the coastal provinces began production under the aegis of the Nationalist government; and because coal, iron, and other resources were in such close proximity, industry rapidly expanded. Considerable industrial development was undertaken by the Communist government after 1949. By the late 20th century Chungking had become one of the largest and fastest growing industrial centres in southwestern China.

Industrial growth

The city's enormous complex of integrated iron and steel plants is among China's largest facilities. Ore is mined at Ch'i-chiang (in the southern part of the municipality) and at Wei-yüan (a short distance west of Nei-chiang). Coal is mined at several locations in the municipality, including Chiang-pei, Pei-p'ei, Pa-hsien, Pi-shan, and Ho-ch'uan. Oil transported from a major oil field just to the west at Tzu-kung and by pipeline from a field to the north supplies Chungking's oil refinery. Power-generating capacity was greatly enlarged with the completion of the Shih-tze-t'an hydroelectric station on the Lung-ch'i River, northeast of the city.

Other important heavy industries include machine, farm tool, and munitions factories; truck and motor-coach manufacturing plants; and chemical and fertilizer plants (manufacturing soap, candles, acid and caustic soda, fertilizers, plastics, and chemical fibres). The city also has a copper refinery, alcohol plants (making gasoline substitutes), and rubber reconditioning plants. In light industries Chungking leads the entire southwest of the nation. Noteworthy are the cotton, silk, paper, and leather industries, as well as flour mills, dyeing factories, and vegetable-oil and food-processing plants. Chungking is also noted for its handicrafts, especially lacquer ware.

In the 1980s the city became the scene of major experiments in industrial economics. Factories in Chungking, for the first time in China, were allowed to channel some of their profits into expansion; and factory managers were given increased decision-making power.

**Trade.** Chungking is the focal point of trade and transport not only of Szechwan but also of the hinterland provinces of Shensi, Yunnan, and Kweichow and of the autonomous region of Tibet. Since 1979 its port—along with several others on the Yangtze—has been open for direct foreign trade, increasing the city's importance as an international trade centre. Chungking is a major oil port, and other exports to downriver provinces and abroad include rolled steel, chemicals, raw silk, goatskin, wool, hides, hog bristles, salt, sugar, tobacco, tung oil, jute, wax, canned foods, medicinal herbs, rhubarb, and musk. Before World War II Chungking imported large quantities of consumer goods from downriver or from abroad, but rapid industrialization brought self-sufficiency in consumer goods to Szechwan and the interior provinces.

**Transportation.** Since 1949 bicycles, buses, and motorbikes have replaced chairs on bamboo poles and rickshas as the principal means of transport in Chungking. Cable tramways provide cheap and convenient transport over the steep hills.

Chungking is served by two great rivers, the Yangtze and the Chia-ling, and is the leading port of southwestern China. As a result of extensive work carried out in the 1950s—including dredging, clearing shoals, and installing buoys and signals—navigation through the Yangtze Gorges has been rendered easy and safe. Steamers now make a round trip between Hankow (Wu-han) in eastern Hupeh Province and Chungking in less than a week. Above Chungking smaller steamers are able to sail

up to I-pin on the Yangtze (and beyond to Chia-ting on the Min River) and up to Nan-ch'ung on the Chia-ling. Above these points, junks can navigate beyond Ch'eng-tu to Kuan-hsien and Mao-hsien on the Min and to Lüeh-yang in southern Shensi on the Chia-ling.

Chungking's railroad system developed rapidly after 1949. The Chungking–Ch'eng-tu railroad, completed in 1952, is the vital link between the Ch'eng-tu Plain and the Yangtze; a southern spur extends through Tzu-kung and I-pin. The Ch'eng-tu–Pao-chi line, completed four years later, connects the city with the Lunghai Railroad and the entire Northwest, as well as with Wu-han in Hupeh Province and a major north–south line; the Chungking–An-k'ang railway also directly links the city with Wu-han. The Chungking–Kuei-yang railroad not only connects Chungking with the province of Kweichow to the south but also joins other lines in Yunnan and Kwangsi running to the Vietnamese border.

The first roads for wheeled traffic in the city were built in 1933. As a result of work begun during World War II, Chungking is now the hub of an extensive network of highways. Major arterials lead south to Kuei-yang (303 miles), northeast to Wan-hsien (258 miles), and northwest to Ch'eng-tu (275 miles). The Chungking Yangtze Bridge carries highway traffic across the river from the southern Kuei-yang highway to the northern part of the city. An air terminal, located at P'ai-shih-i about 17 miles west of the city, provides regular flights to major cities throughout China.

### ADMINISTRATION AND SOCIAL CONDITIONS

**Government.** Chungking's municipal government is part of the hierarchical structure of the Chinese government—and the parallel structure of the Chinese Communist Party—that extends from the national organization, through the provincial apparatus, to the municipal and, ultimately, neighbourhood levels. The principal responsibilities of the Chungking Municipal People's Congress, the major decision-making body, include issuing administrative orders, collecting taxes, determining the budget, and implementing economic plans. A standing committee selected from its members recommends policy decisions and oversees the operation of municipal government. Executive authority rests with the Chungking People's Government, the officers of which are elected by the Chungking Municipal People's Congress; it consists of a mayor, vice mayors, and numerous bureaus in charge of public security, the judicial system, and other civil, economic, social, and cultural affairs.

Administratively the city is divided into a number of districts (*ch'ü*), each under a district mayoralty. At the next lower level are police substations and street mayoralties that handle civil affairs in the same subareas. Neighbourhood street committees perform the auxiliary functions of mediating disputes, propagating legal orders, and promoting sanitation and welfare. These committees are quasi-official administrations, consolidated under Chinese Communist Party leadership and covering blocks of streets of varying sizes. Chungking Municipality has considerably extended the territorial limits of the municipal area to include a series of urban–rural units surrounding the city proper. Since 1980 the municipal government has allowed farmers to engage in industry, commerce, and transportation in addition to cultivation.

Policy of urban–rural integration

**Public utilities.** Although an electric-light plant was established in the early 1900s, it was not until the late 1920s and early 1930s that a modernization drive was launched by local leaders in Chungking to improve living conditions. Demolition of the city walls was initiated, streets were widened, and a piped water system and a telephone exchange were introduced. Yet even during the 1940s sanitation and public hygiene were still poor. The city had a large rat population, opium smoking in homes and inns was widespread, and lice-ridden waifs and beggars were a familiar sight. But because of energetic measures carried out since 1949, including the installation of a modern sewer system, these conditions belong to the past. Chungking has now achieved a high degree of cleanliness, although air pollution has become a problem.

**Health.** Chungking has a considerable number of hospitals. The major share of medical care, however, is provided by clinics and health stations that are operated by neighbourhood street committees. These clinics are equipped with a limited number of beds and are staffed by physicians. Western-style medicine is combined with traditional herb medicine and acupuncture. Family planning is practiced and contraceptives are free. Because the government recognizes that there still is a shortage of medical facilities, it places great emphasis on the drive for physical fitness.

**Education.** Since 1949 the number of schools at all levels—kindergartens, primary schools, middle schools, and colleges—has increased. The growth of kindergartens, which were little known in prewar years, has enabled many women to obtain proper care for their children and thus become part of the work force. The government has attached great importance to the establishment of teacher-training schools, vocational-technical schools, and part-time agricultural middle schools.

Institutions of higher learning

The Chungking University (founded in 1929) offers its students a comprehensive range of studies. Other institutions of higher learning include the Southwest Political Science and Law College, the Chungking Institute of Medicine, the Chungking Construction Engineering College, the Southwestern Institute of Agriculture, the Szechwan Institute of Fine Arts, and the Southwest College of Education. The Chungking Library and the Chungking Municipal Museum are among the leading cultural centres in the city.

### CULTURAL LIFE

Sports and recreation are dominant features of Chungking's cultural life. The Ta-t'ien-wan Stadium, the city's main sports centre, offers a football field; volleyball, basketball, and tennis courts; a track-and-field playground; and a parachute tower. The stadium has a capacity of 100,000. Numerous parks, both in the Old City and in outlying areas, attract large numbers of visitors. Of particular appeal are the hot springs, which are open year-round. South of the city, among well-kept gardens with lakes and pavilions, are the sulfurous springs of Nan-wen-ch'üan Park. To the north of the city are the well-known hot springs of Pei-wen-ch'üan Park along the Chia-ling River. Visitors come to relax, often soaking for hours in one of the numerous baths filled with warm mineral water, or they swim in one of the three Olympic-sized pools, which are also fed by the hot springs.

A noteworthy feature of Chungking's cultural life is its distinctive Szechwan cuisine. This highly spiced food is characterized mainly by the use of hot peppers as well as by such delicacies as tree ears, black mushrooms, and fresh bamboo shoots and peanuts.

## History

### THE EARLY PERIOD

According to ancient accounts Chungking was the birthplace of the consort of Emperor Yü of the legendary Hsia dynasty, about 4,000 years ago. In the 11th century BC, under the Western Chou dynasty, the region surrounding Chungking became a feudal state known as Pa. In the 5th century BC, Pa established relations with the mid-Yangtze kingdom of Ch'u. It was later incorporated into the Ch'in empire. By the mid-3rd century BC the region became part of the kingdom of Shu and was totally independent of northern and central China.

The swing of the historical pendulum—in which the city's status alternated between being ruled by an empire in northern China, forming part of an empire in central China, and detaching itself to become independent of both northern and central China—continued throughout subsequent centuries. The city finally became an integral part of the unified Chinese empire, first under the Ming dynasty (1368–1644), and then under the Ch'ing, or Manchu, dynasty (1644–1911/12).

The city walls

The first substantial city wall was constructed around 250 BC. It was repaired and expanded during the 3rd century AD and rebuilt with solid stone in 1370. In the 1630s,

at the end of the Ming dynasty, the rebellion of Chang Hsien-chung subjected Chungking to plunder, slaughter, and destruction. The city wall was restored in 1663. Some five miles in circumference, it had a total of 17 gates: eight gates remained closed on the advice of geomancers (practitioners of divination by means of figures or lines), while nine were open to traffic. Additional work was done to strengthen the city wall in 1760.

### THE MODERN PERIOD

Chungking was opened to British trade in 1890, but navigational difficulties on the Yangtze delayed steamer traffic for more than a decade. Meanwhile, the Treaty of Shimonoseki (1895), which concluded the first Sino-Japanese War (1894–95), gave Japan the right to establish a concession. Accordingly, in 1901, when British trade opened, a Japanese concession was established at Wang-chia-to on the south shore of the Yangtze. This concession lasted until 1937, when it was abandoned by Japan on the outbreak of war.

On the eve of the revolution of 1911 Chungking, along with the provincial capital, Ch'eng-tu, played a major role in bringing about the overthrow of the Manchus; many patriots of the region joined the revolutionary party of the Chinese Nationalist leader Sun Yat-sen. Despite such progressive trends and a nominal allegiance to the central government, Chungking was unable to break away from the grip of regional separatism.

Capital of the Nationalist government

Yet in 1938, when war again broke out with Japan, Chungking became the capital of the Nationalist government. Hundreds of government offices were moved to the city from Nanking, along with the diplomatic missions of foreign powers; and tens of thousands of people came from coastal provinces, bringing with them arsenals, factories, and schools. Friendly powers, too, rushed supplies to Chungking to bolster its war effort. Despite the Japanese bombings, the morale of its population—at the time more than 1,000,000—was high. Chiang Kai-shek's failure to control inflation and corruption, however, caused the war effort to falter from 1942 onward. In 1946, on the eve of the renewed civil war against the Communists, the Nationalist capital returned to Nanking. Three years later, in April 1949, Nanking fell. The Nationalist government fled to Canton and then once again—for less than two months—to Chungking (October to December 1949). When the Nationalists fled to Taiwan in December, the Communist victory on the mainland was complete.

Shortly after the Communist takeover in 1949, repair of the war damage began, and expansion of the city's industrial base, established in the early 20th century, was vigorously pursued. Even though energies were temporarily deflected during the Great Leap Forward (1958–60) and the Cultural Revolution (1966–76), the city, nonetheless, succeeded in carrying out extensive modernization projects and significantly raised the standard of living.

**BIBLIOGRAPHY.** Comprehensive general references are FREDRIC KAPLAN, JULIAN SOBIN, and ARNE DE KEIJZER, *The China Guidebook,* 6th ed. (1985), revised annually; and NAGEL PUBLISHERS, *China,* English version by ANNE L. DESTENAY, 4th ed. (1982). References to Chungking's role in Chinese history can be found in IMMANUEL C.Y. HSÜ, *The Rise of Modern China,* 3rd ed. (1983); and an analysis of a specific epoch is ROBERT A. KAPP, *Szechwan and the Chinese Republic, 1911–1938* (1973). Chungking during World War II is portrayed in THEODORE H. WHITE and ANNALEE JACOBY, *Thunder out of China* (1946, reprinted 1980); and PAUL M.A. LINEBARGER, *The China of Chiang K'ai-shek: A Political Study* (1941, reprinted 1973). EDGAR SNOW, *Red China Today,* rev. ed. (1971), contains one of the best descriptions of the modern scene. For geography, see GEORGE BABCOCK CRESSEY, *China's Geographic Foundations* (1934); and for an economic geography, see T.R. TREGEAR, *China, a Geographical Survey* (1980). FREDERICA M. BUNGE and RINN-SUP SHINN (eds.), *China, a Country Study,* 3rd ed. (1981), discusses several aspects of Chungking's industry, trade, and transportation. Articles in *China Reconstructs* (monthly) and *Far Eastern Economic Review* (weekly) contain useful information on contemporary developments.

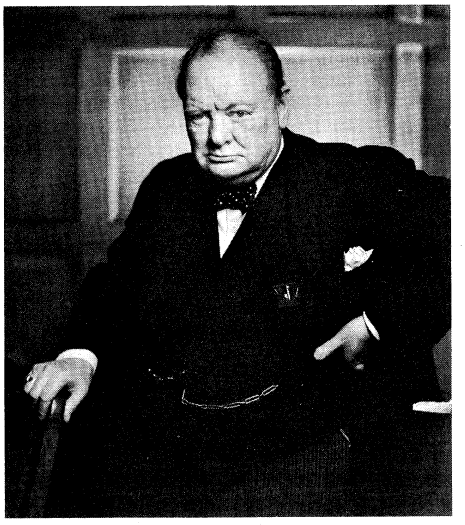
(P.-c.K./Wa.M.)

# Churchill

Sir Winston Churchill—author, orator, and statesman—led Great Britain from the brink of defeat to victory as wartime prime minister from 1940 to 1945. After a sensational rise to prominence in national politics before World War I, he acquired a reputation for erratic judgment in the war itself and in the decade that followed. Politically suspect in consequence, he was a lonely figure until his response to Adolf Hitler's challenge brought him to leadership of a national coalition in 1940. With Franklin D. Roosevelt and Joseph Stalin he shaped Allied strategy in World War II, and after the breakdown of the alliance he alerted the West to the expansionist threat of the Soviet Union. He led the Conservative Party back to office in 1951 and remained prime minister until 1955, when ill health forced his resignation.

Churchill was born on Nov. 30, 1874, prematurely, at Blenheim Palace, Oxfordshire, and was christened Winston Leonard Spencer Churchill. In his veins ran the blood of both of the English-speaking peoples whose unity, in peace and war, it was to be a constant purpose of his to promote. Through his father, Lord Randolph Churchill, the meteoric Tory politician, he was directly descended from John Churchill, 1st duke of Marlborough, the hero of the wars against Louis XIV of France in the early 18th century. His mother, Jennie Jerome, a noted beauty, was the daughter of a New York financier and horse racing enthusiast, Leonard W. Jerome.

© Karsh—Woodfin Camp and Associates



Churchill, photographed by Yousuf Karsh, 1941.

Early career as a soldier and journalist

The young Churchill passed an unhappy and sadly neglected childhood, redeemed only by the affection of Mrs. Everest, his devoted nurse. At Harrow his conspicuously poor academic record seemingly justified his father's decision to enter him into an army career. It was only at the third attempt that he managed to pass the entrance examination to the Royal Military College, now Academy, Sandhurst, but, once there, he applied himself seriously and passed out (graduated) 20th in a class of 130. In 1895, the year of his father's tragic death, he entered the 4th Hussars. Initially the only prospect of action was in Cuba, where he spent a couple of months of leave reporting the Cuban war of independence from Spain for the *Daily Graphic* (London). In 1896 his regiment went to India, where he saw service as both soldier and journalist on the North-West Frontier (1897). Expanded as *The Story of the Malakand Field Force* (1898), his dispatches attracted such wide attention as to launch him on the career of authorship that he intermittently pursued throughout his life. In 1897–98 he wrote *Savrola* (1900), a Ruritanian romance, and got himself attached to Lord Kitchener's Nile expeditionary force in the same dual role of soldier and correspondent. *The River War* (1899) brilliantly describes the campaign.

## POLITICAL CAREER BEFORE 1939

The five years after Sandhurst saw Churchill's interests expand and mature. He relieved the tedium of army life in India by a program of reading designed to repair the deficiencies of Harrow and Sandhurst, and in 1899 he resigned his commission to enter politics and make a living by his pen. He first stood as a Conservative at Oldham, where he lost a by-election by a narrow margin, but found quick solace in reporting the South African War for *The Morning Post* (London). Within a month after his arrival in South Africa he had won fame for his part in rescuing an armoured train ambushed by Boers, though at the price of himself being taken prisoner. But this fame was redoubled when less than a month later he escaped from military prison. Returning to Britain a military hero, he laid siege again to Oldham in the election of 1900. Churchill succeeded in winning by a margin as narrow as that of his previous failure. But he was now in Parliament and, fortified by the £10,000 his writings and lecture tours had earned for him, was in a position to make his own way in politics.

A self-assurance redeemed from arrogance only by a kind of boyish charm made Churchill from the first a notable House of Commons figure, but a speech defect, which he never wholly lost, combined with a certain psychological inhibition to prevent him from immediately becoming a master of debate. He excelled in the set speech, on which he always spent enormous pains, rather than in the impromptu; Lord Balfour, the Conservative leader, said of him that he carried "heavy but not very mobile guns." In matter as in style he modeled himself on his father, as his admirable biography, *Lord Randolph Churchill* (1906; revised edition 1952), makes evident, and from the first he wore his Toryism with a difference, advocating a fair, negotiated peace for the Boers and deploring military mismanagement and extravagance.

**As Liberal minister.** In 1904 the Conservative government found itself impaled on a dilemma by Colonial Secretary Joseph Chamberlain's open advocacy of a tariff. Churchill, a convinced free trader, helped to found the Free Food League. He was disavowed by his constituents and became increasingly alienated from his party. In 1904 he joined the Liberals and won renown for the audacity of his attacks on Chamberlain and Balfour. The radical elements in his political makeup came to the surface under the influence of two colleagues in particular, John Morley, a political legatee of W.E. Gladstone, and David Lloyd George, the rising Welsh orator and firebrand. In the ensuing general election in 1906 he secured a notable victory in Manchester and began his ministerial career in the new Liberal government as undersecretary of state for the colonies. He soon gained credit for his able defense of the policy of conciliation and self-government in South Africa. When the ministry was reconstructed under Prime Minister Herbert H. Asquith in 1908, Churchill was promoted to president of the Board of Trade, with a seat in the Cabinet. Defeated at the ensuing by-election in Manchester, he won an election at Dundee. In the same year he married the beautiful Clementine Hozier; it was a marriage of unbroken affection that provided a secure and happy background for his turbulent career.

Attacks on Conservatives

At the Board of Trade, Churchill emerged as a leader in the movement of Liberalism away from laissez-faire toward social reform. He completed the work begun by his predecessor, Lloyd George, on the bill imposing an

eight-hour maximum day for miners. He himself was responsible for attacking the evils of "sweated" labour by setting up trade boards with power to fix minimum wages and for combating unemployment by instituting state-run labour exchanges.

When this Liberal program necessitated high taxation, which in turn provoked the House of Lords to the revolutionary step of rejecting the budget of 1909, Churchill was Lloyd George's closest ally in developing the provocative strategy designed to clip the wings of the upper chamber. Churchill became president of the Budget League, and his oratorical broadsides at the House of Lords were as lively and devastating as Lloyd George's own. Indeed Churchill, as an alleged traitor to his class, earned the lion's share of Tory animosity. His campaigning in the two general elections of 1910 and in the House of Commons during the passage of the Parliament Act of 1911, which curbed the House of Lords' powers, won him wide popular acclaim.

Home secretary

In the Cabinet his reward was promotion to the office of home secretary. Here, despite substantial achievements in prison reform, he had to devote himself principally to coping with a sweeping wave of industrial unrest and violent strikes. Upon occasion his relish for dramatic action led him beyond the limits of his proper role as the guarantor of public order. For this he paid a heavy price in incurring the long-standing suspicion of organized labour.

In 1911 the provocative German action in sending a gunboat to Agadir, the Moroccan port to which France had claims, convinced Churchill that in any major Franco-German conflict Britain would have to be at France's side. When transferred to the Admiralty in October 1911, he went to work with a conviction of the need to bring the navy to a pitch of instant readiness. His first task was the creation of a naval war staff. To help Britain's lead over steadily mounting German naval power, Churchill successfully campaigned in the Cabinet for the largest naval expenditure in British history. Despite his inherited Tory views on Ireland, he wholeheartedly embraced the Liberal policy of Home Rule, moving the second reading of the Irish Home Rule Bill of 1912 and campaigning for it in the teeth of Unionist opposition. Although, through his friendship with F.E. Smith (later 1st earl of Birkenhead) and Austen Chamberlain, he did much to arrange the compromise by which Ulster was to be excluded from the immediate effect of the bill, no member of the government was more bitterly abused—by Tories as a renegade and by extreme Home Rulers as a defector.

**During World War I.** War came as no surprise to Churchill. He had already held a test naval mobilization. Of all the Cabinet ministers he was the most insistent on the need to resist Germany. On Aug. 2, 1914, on his own responsibility, he ordered the naval mobilization that guaranteed complete readiness when war was declared. The war called out all of Churchill's energies. In October 1914, when Antwerp was falling, he characteristically rushed in person to organize its defense. When it fell the public saw only a disillusioning defeat, but in fact the prolongation of its resistance for almost a week enabled the Belgian Army to escape and the crucial Channel ports to be saved. At the Admiralty, Churchill's partnership with Adm. Sir John Fisher, the first sea lord, was productive both of dynamism and of dissension. In 1915, when Churchill became an enthusiast for the Dardanelles expedition as a way out of the costly stalemate on the Western Front, he had to proceed against Fisher's disapproval. The campaign aimed at forcing the straits and opening up direct communications with Russia. When the naval attack failed and was called off on the spot by Adm. J.M. de Robeck, the Admiralty war group and Asquith both supported de Robeck rather than Churchill. Churchill came under heavy political attack, which intensified when Fisher resigned. Preoccupied with departmental affairs, Churchill was quite unprepared for the storm that broke about his ears. He had no part at all in the maneuvers that produced the first coalition government and was powerless when the Conservatives, with the sole exception of Sir William Maxwell Aitken (soon Lord Beaverbrook), insisted on his being demoted from the Admiralty to the duchy of Lancaster. There he was given special responsibility for the Gallipoli Campaign (a land assault at the straits) without, however, any powers of direction. Reinforcements were too few and too late; the campaign failed and casualties were heavy; evacuation was ordered in the autumn.

The Dardanelles Campaign

In November 1915 Churchill resigned from the government and returned to soldiering, seeing active service in France as lieutenant colonel of the 6th Royal Scots Fusiliers. Although he entered the service with zest, army life did not give full scope for his talents. In June 1916, when his battalion was merged, he did not seek another command but instead returned to Parliament as a private member. He was not involved in the intrigues that led to the formation of a coalition government under Lloyd George, and it was not until 1917 that the Conservatives would consider his inclusion in the government. In March 1917 the publication of the Dardanelles commission report demonstrated that he was at least no more to blame for the fiasco than his colleagues.

Even so, Churchill's appointment as minister of munitions in July 1917 was made in the face of a storm of Tory protest. Excluded from the Cabinet, Churchill's role was almost entirely administrative, but his dynamic energies thrown behind the development and production of the tank (which he had initiated at the Admiralty) greatly speeded up the use of the weapon that broke through the deadlock on the Western Front. Paradoxically, it was not until the war was over that Churchill returned to a service department. In January 1919 he became secretary of war. As such he presided with surprising zeal over the cutting of military expenditure. The major preoccupation of his tenure in the War Office was, however, the Allied intervention in Russia. Churchill, passionately anti-Bolshevik, secured from a divided and loosely organized Cabinet an intensification and prolongation of the British involvement beyond the wishes of any major group in Parliament or the nation—and in the face of the bitter hostility of labour. And in 1920, after the last British forces had been withdrawn, Churchill was instrumental in having arms sent to the Poles when they invaded the Ukraine.

Appointment to the Colonial Office

In 1921 Churchill moved to the Colonial Office, where his principal concern was with the mandated territories in the Middle East. For the costly British forces in the area he substituted a reliance on the air force and the establishment of rulers congenial to British interests; for this settlement of Arab affairs he relied heavily on the advice of T.E. Lawrence. For Palestine, where he inherited conflicting pledges to Jews and Arabs, he produced in 1922 the White Paper that confirmed Palestine as a Jewish national home while recognizing continuing Arab rights. Churchill never had departmental responsibility for Ireland, but he progressed from an initial belief in firm, even ruthless, maintenance of British rule to an active role in the negotiations that led to the Irish treaty of 1921. Subsequently, he gave full support to the new Irish government.

In the autumn of 1922 the insurgent Turks appeared to be moving toward a forcible reoccupation of the Dardanelles neutral zone, which was protected by a small British force at Chanak (now Çanakkale). Churchill was foremost in urging a firm stand against them, but the handling of the issue by the Cabinet gave the public impression that a major war was being risked for an inadequate cause and on insufficient consideration. A political debacle ensued that brought the shaky coalition government down in ruins, with Churchill as one of the worst casualties. Gripped by a sudden attack of appendicitis, he was not able to appear in public until two days before the election, and then only in a wheelchair. He was defeated humiliatingly by more than 10,000 votes. He thus found himself, as he said, all at once "without an office, without a seat, without a party, and even without an appendix."

**In and out of office, 1922–29.** In convalescence and political impotence Churchill turned to his brush and his pen. His painting never rose above the level of a gifted amateur's, but his writing once again provided him with the financial base his independent brand of politics required. His autobiographical history of the war, *The World Crisis,* netted him the £20,000 with which he purchased Chartwell, henceforth his country home in Kent. When

he returned to politics it was as a crusading anti-Socialist, but in 1923, when Stanley Baldwin was leading the Conservatives on a protectionist program, Churchill stood, at Leicester, as a Liberal free trader. He lost by approximately 4,000 votes. Asquith's decision in 1924 to support a minority Labour government moved Churchill farther to the right. He stood as an "Independent Anti-Socialist" in a by-election in the Abbey division of Westminster. Although opposed by an official Conservative candidate—who defeated him by a hairbreadth of 43 votes—Churchill managed to avoid alienating the Conservative leadership and indeed won conspicuous support from many prominent figures in the party. In the general election in November 1924 he won an easy victory at Epping under the thinly disguised Conservative label of "Constitutionalist." Baldwin, free of his flirtation with protectionism, offered Churchill, the "constitutionalist free trader," the post of chancellor of the Exchequer. Surprised, Churchill accepted; dumbfounded, the country interpreted it as a move to absorb into the party all the right-of-centre elements of the former coalition.

Chancellor of the Exchequer under Baldwin

In the five years that followed, Churchill's early liberalism survived only in the form of advocacy of rigid laissez-faire economics; for the rest he appeared, repeatedly, as the leader of the diehards. He had no natural gift for financial administration, and though the noted economist John Maynard Keynes criticized him unsparingly, most of the advice he received was orthodox and harmful. His first move was to restore the gold standard, a disastrous measure, from which flowed deflation, unemployment, and the miners' strike that led to the general strike of 1926. Churchill offered no remedy except the cultivation of strict economy, extending even to the armed services. Churchill viewed the general strike as a quasi-revolutionary measure and was foremost in resisting a negotiated settlement. He leaped at the opportunity of editing the *British Gazette,* an emergency official newspaper, which he filled with bombastic and frequently inflammatory propaganda. The one relic of his earlier radicalism was his partnership with Neville Chamberlain as minister of health in the cautious expansion of social services, mainly in the provision of widows' pensions.

In 1929, when the government fell, Churchill, who would have liked a Tory-Liberal reunion, deplored Baldwin's decision to accept a minority Labour government. The next year an open rift developed between the two men. On Baldwin's endorsement of a Round Table Conference with Indian leaders, Churchill resigned from the shadow cabinet and threw himself into a passionate, at times almost hysterical, campaign against the Government of India bill (1935) designed to give India dominion status.

**Exclusion from office, 1929–39.** Thus, when in 1931 the National Government was formed, Churchill, though a supporter, had no hand in its establishment or place in its councils. He had arrived at a point where, for all his abilities, he was distrusted by every party. He was thought to lack judgment and stability and was regarded as a guerrilla fighter impatient of discipline. He was considered a clever man who associated too much with clever men—Birkenhead, Beaverbrook, Lloyd George—and who despised the necessary humdrum associations and compromises of practical politics.

The biography of Marlborough

In this situation he found relief, as well as profit, in his pen, writing, in *Marlborough: His Life and Times,* a massive rehabilitation of his ancestor against the criticisms of the 19th-century historian Thomas Babington Macaulay. But overriding the past and transcending his worries about India was a mounting anxiety about the growing menace of Hitler's Germany. Before a supine government and a doubting opposition, Churchill persistently argued the case for taking the German threat seriously and for the need to prevent the Luftwaffe from securing parity with the Royal Air Force. In this he was supported by a small but devoted personal following, in particular the gifted, curmudgeonly Oxford physics professor Frederick A. Lindemann (later Lord Cherwell), who enabled him to build up at Chartwell a private intelligence centre the information of which was often superior to that of the government. When Baldwin became prime minister in 1935, he persisted in excluding Churchill from office but gave him the exceptional privilege of membership in the secret committee on air-defense research, thus enabling him to work on some vital national problems. But Churchill had little success in his efforts to impart urgency to Baldwin's administration. The crisis that developed when Italy invaded Ethiopia in 1935 found Churchill ill prepared, divided between a desire to build up the League of Nations around the concept of collective security and the fear that collective action would drive Benito Mussolini into the arms of Hitler. The Spanish Civil War (1936–39) found him convinced of the virtues of nonintervention, first as a supporter and later as a critic of Francisco Franco. Such vagaries of judgment in fact reflected the overwhelming priority he accorded to one issue—the containment of German aggressiveness. At home there was one grievous, characteristic, romantic misreading of the political and public mood, when, in Edward VIII's abdication crisis of 1936, he vainly opposed Baldwin by a public championing of the King's cause.

When Neville Chamberlain succeeded Baldwin, the gulf between the Cassandra-like Churchill and the Conservative leaders widened. Repeatedly the accuracy of Churchill's information on Germany's aggressive plans and progress was confirmed by events; repeatedly his warnings were ignored. Yet his handful of followers remained small; politically, Chamberlain felt secure in ignoring them. As German pressure mounted on Czechoslovakia, Churchill without success urged the government to effect a joint declaration of purpose by Great Britain, France, and the Soviet Union. When the Munich Agreement with Hitler was made in September 1938, sacrificing Czechoslovakia to the Nazis, Churchill laid bare its implications, insisting that it represented "a total and unmitigated defeat." In March 1939 Churchill and his group pressed for a truly national coalition, and, at last, sentiment in the country, recognizing him as the nation's spokesman, began to agitate for his return to office. As long as peace lasted, Chamberlain ignored all such persuasions.

Churchill versus Chamberlain

## LEADERSHIP DURING WORLD WAR II

In a sense, the whole of Churchill's previous career had been a preparation for wartime leadership. An intense patriot; a romantic believer in his country's greatness and its historic role in Europe, the empire, and the world; a devotee of action who thrived on challenge and crisis; a student, historian, and veteran of war; a statesman who was master of the arts of politics, despite or because of long political exile; a man of iron constitution, inexhaustible energy, and total concentration, he seemed to have been nursing all his faculties so that when the moment came he could lavish them on the salvation of Britain and the values he believed Britain stood for in the world.

On Sept. 3, 1939, the day Britain declared war on Germany, Chamberlain appointed Churchill to his old post in charge of the Admiralty. The signal went out to the fleet: "Winston is back." On September 11 Churchill received a congratulatory note from Pres. Franklin D. Roosevelt and replied over the signature "Naval Person"; a memorable correspondence had begun. At once Churchill's restless energy began to be felt throughout the administration, as his ministerial colleagues as well as his own department received the first of those pungent minutes that kept the remotest corners of British wartime government aware that their shortcomings were liable to detection and penalty. All his efforts, however, failed to energize the torpid Anglo-French entente during the so-called "phony war," the period of stagnation in the European war before the German seizure of Norway in April 1940. The failure of the Narvik and Trondheim expeditions, dependent as they were on naval support, could not but evoke some memories of the Dardanelles and Gallipoli, so fateful for Churchill's reputation in World War I. This time, however, it was Chamberlain who was blamed, and it was Churchill who endeavoured to defend him.

**As prime minister.** The German invasion of the Low Countries, on May 10, 1940, came like a hammer blow on top of the Norwegian fiasco. Chamberlain resigned. He wanted Lord Halifax, the foreign secretary, to succeed him, but Halifax wisely declined. It was obvious that

Churchill alone could unite and lead the nation, since the Labour Party, for all its old distrust of Churchill's anti-Socialism, recognized the depth of his commitment to the defeat of Hitler. A coalition government was formed that included all elements save the far left and right. It was headed by a war Cabinet of five, which included at first both Chamberlain and Halifax—a wise but also magnanimous recognition of the numerical strength of Chamberlainite conservatism—and two Labour leaders, Clement Attlee and Arthur Greenwood. The appointment of Ernest Bevin, a tough trade-union leader, as minister of labour guaranteed cooperation on this vital front. Offers were made to Lloyd George, but he declined them. Churchill himself took, in addition to the leadership of the House of Commons, the Ministry of Defence. The pattern thus set was maintained throughout the war despite many changes of personnel. The Cabinet became an agency of swift decision, and the government that it controlled remained representative of all groups and parties. The Prime Minister concentrated on the actual conduct of the war. He delegated freely but also probed and interfered continuously, regarding nothing as too large or too small for his attention. The main function of the chiefs of the armed services became that of containing his great dynamism, as a governor regulates a powerful machine; but, though he prodded and pressed them continuously, he never went against their collective judgment. In all this, Parliament played a vital part. If World War II was strikingly free from the domestic political intrigues of World War I, it was in part because Churchill, while he always dominated Parliament, never neglected it or took it for granted. For him, Parliament was an instrument of public persuasion on which he played like a master and from which he drew strength and comfort.

Churchill and Parliament

On May 13 Churchill faced the House of Commons for the first time as prime minister. He warned members of the hard road ahead—"I have nothing to offer but blood, toil, tears and sweat"—and committed himself and the nation to all-out war until victory was achieved. Behind this simplicity of aim lay an elaborate strategy to which he adhered with remarkable consistency throughout the war. Hitler's Germany was the enemy; nothing should distract the entire British people from the task of effecting its defeat. Anyone who shared this goal, even a Communist, was an acceptable ally. The indispensable ally in this endeavour, whether formally at war or not, was the United States. The cultivation and maintenance of its support was a central principle of Churchill's thought. Yet whether the United States became a belligerent partner or not, the war must be won without a repetition for Britain of the catastrophic bloodlettings of World War I; and Europe at the conflict's end must be reestablished as a viable, self-determining entity, while the Commonwealth should remain as a continuing, if changing, expression of Britain's world role. Provided these essentials were preserved, Churchill, for all his sense of history, was surprisingly willing to sacrifice any national shibboleths—of orthodox economics, of social convention, of military etiquette or tradition—on the altar of victory. Thus, within a couple of weeks of this crusading anti-Socialist's assuming power, Parliament passed legislation placing all "persons, their services and their property at the disposal of the Crown"—granting the government in effect the most sweeping emergency powers in modern British history.

The effort was designed to match the gravity of the hour. After the Allied defeat and the evacuation of the battered British forces from Dunkirk, Churchill warned Parliament that invasion was a real risk to be met with total and confident defiance. Faced with the swift collapse of France, Churchill made repeated personal visits to the French government in an attempt to keep France in the war, culminating in the celebrated offer of Anglo-French union on June 16, 1940. When all this failed, the Battle of Britain began. Here Churchill was in his element, in the firing line—at fighter headquarters, inspecting coast defenses or antiaircraft batteries, visiting scenes of bomb damage or victims of the "blitz," smoking his cigar, giving his V sign, or broadcasting frank reports to the nation, laced with touches of grim Churchillian humour and splashed with Churchillian rhetoric. The nation took him to its heart; he and they were one in "their finest hour."

The Battle of Britain

Other painful and more debatable decisions fell to Churchill. The French fleet was attacked to prevent its surrender intact to Hitler. A heavy commitment was made to the concentrated bombing of Germany. At the height of the invasion threat, a decision was made to reinforce British strength in the eastern Mediterranean. Forces were also sent to Greece, a costly sacrifice; the evacuation of Crete looked like another Gallipoli, and Churchill came under heavy fire in Parliament.

In these hard days the exchange of U.S. overage destroyers for British Caribbean bases and the response, by way of lend-lease, to Churchill's boast "Give us the tools and we'll finish the job" were especially heartening to one who believed in a "mixing-up" of the English-speaking democracies. The unspoken alliance was further cemented in August 1941 by the dramatic meeting between Churchill and Roosevelt in Placentia Bay, Newfoundland, which produced the Atlantic Charter, a statement of common principles between the United States and Britain.

**Formation of the "grand alliance."** When Hitler launched his sudden attack on the Soviet Union, Churchill's response was swift and unequivocal. In a broadcast on June 22, 1941, while refusing to "unsay" any of his earlier criticisms of Communism, he insisted that "the Russian danger . . . is our danger" and pledged aid to the Russian people. Henceforth, it was his policy to construct a "grand alliance" incorporating the Soviet Union and the United States. But it took until May 1942 to negotiate a 20-year Anglo-Soviet pact of mutual assistance.

The Japanese attack on Pearl Harbor (Dec. 7, 1941) altered, in Churchill's eyes, the whole prospect of the war. He went at once to Washington, D.C., and, with Roosevelt, hammered out a set of Anglo-American accords: the pooling of both countries' military and economic resources under combined boards and a combined chiefs of staff; the establishment of unity of command in all theatres of war; and agreement on the basic strategy that the defeat of Germany should have priority over the defeat of Japan. The grand alliance had now come into being. Churchill could claim to be its principal architect. Safeguarding it was the primary concern of his next three and a half years.

The accord with Roosevelt

In protecting the alliance, the respect and affection between him and Roosevelt were of crucial importance. They alone enabled Churchill, in the face of relentless pressure from Stalin and ardent advocacy by the U.S. chiefs of staff, to secure the rejection of the "second front" in 1942, a project he regarded as premature and costly. In August 1942 Churchill himself flew to Moscow to advise Stalin of the decision and to bear the brunt of his displeasure. At home, too, he came under fire in 1942: first in January after the reverses in Malaya and the Far East and later in June when Tobruk in North Africa fell to the Germans, but on neither occasion did his critics muster serious support in Parliament. The year 1942 saw some reconstruction of the Cabinet in a "leftward" direction, which was reflected in the adoption in 1943 of Lord Beveridge's plan for comprehensive social insurance, endorsed by Churchill as a logical extension of the Liberal reforms of 1911.

**Military successes and political problems.** The Allied landings in North Africa necessitated a fresh meeting between Churchill and Roosevelt, this time in Casablanca in January 1943. There Churchill argued for an early, full-scale attack on "the under-belly of the Axis" but won only a grudging acquiescence from the Americans. There too was evolved the "unconditional surrender" formula of debatable wisdom. Churchill paid the price for his intensive travel (including Tripoli, Turkey, and Algeria) by an attack of pneumonia, for which, however, he allowed only the briefest of respites. In May he was in Washington again, arguing against persistent American aversion to his "under-belly" strategy; in August he was at Quebec, working out the plans for Operation Overlord, the cross-Channel assault. When he learned that the Americans were planning a large-scale invasion of Burma in 1944, his fears that their joint resources would not be adequate

Operation Overlord

for a successful invasion of Normandy were revived. In November 1943 at Cairo he urged on Roosevelt priority for further Mediterranean offensives, but at Tehrān in the first "Big Three" meeting, he failed to retain Roosevelt's adherence to a completely united Anglo-American front. Roosevelt, though he consulted in private with Stalin, refused to see Churchill alone; for all their friendship there was also an element of rivalry between the two Western leaders that Stalin skillfully exploited. On the issue of Allied offensive drives into southern Europe, Churchill was outvoted. Throughout the meetings Churchill had been unwell, and on his way home he came down again with pneumonia. Though recovery was rapid, it was mid-January 1944 before convalescence was complete. By May he was proposing to watch the D-Day assaults from a battle cruiser; only the King's personal plea dissuaded him.

Insistence on military success did not, for Churchill, mean indifference to its political implications. After the Quebec conference in September 1944, he flew to Moscow to try to conciliate the Russians and the Poles and to get an agreed division of spheres of influence in the Balkans that would protect as much of them as possible from Communism. In Greece he used British forces to thwart a Communist takeover and at Christmas flew to Athens to effect a settlement. Much of what passed at the Yalta Conference in February 1945, including the Far East settlement, concerned only Roosevelt and Stalin, and Churchill did not interfere. He fought to save the Poles but saw clearly enough that there was no way to force the Soviets to keep their promises. Realizing this, he urged the United States to allow the Allied forces to thrust as far into eastern Europe as possible before the Russian armies should fill the vacuum left by German power, but he could not win over Roosevelt, Vice Pres. Harry S. Truman, or their generals to his views. He went to Potsdam in July in a worried mood. But in the final decisions of the conference he had no part; halfway through, when news came of his government's defeat in parliamentary elections, he had to return to England and tender his resignation.

**Electoral defeat.** Already in 1944, with victory in prospect, party politics had revived, and by May 1945 all parties in the wartime coalition wanted an early election. But whereas Churchill wanted the coalition to continue at least until Japan was defeated, Labour wished to resume its independence. Churchill as the popular architect of victory seemed unbeatable, but as an election campaigner he proved to be his own worst enemy, indulging, seemingly at Beaverbrook's urging, in extravagant prophecies of the appalling consequences of a Labour victory and identifying himself wholly with the Conservative cause. His campaign tours were a triumphal progress, but it was the war leader, not the party leader, whom the crowds cheered. Labour's careful but sweeping program of economic and social reform was a better match for the nation's mood than Churchill's flamboyance. Though personally victorious at his Essex constituency of Woodford, Churchill saw his party reduced to 213 seats in a Parliament of 640.

## POSTWAR POLITICAL CAREER

**As opposition leader and world statesman.** The shock of rejection by the nation fell heavily on Churchill. Indeed, though he accepted the role of leader of the parliamentary opposition, he was never wholly at home in it. The economic and social questions that dominated domestic politics were not at the centre of his interests. Nor, with his imperial vision, could he approve of what he called Labour's policy of "scuttle," as evidenced in the granting of independence to India and Burma (though he did not vote against the necessary legislation). But in foreign policy a broad identity of view persisted between the front benches, and this was the area to which Churchill primarily devoted himself. On March 5, 1946, at Fulton, Mo., he enunciated, in the presence of President Truman, the two central themes of his postwar view of the world: the need for Britain and the United States to unite as guardians of the peace against the menace of Soviet Communism, which had brought down an "iron curtain" across the face of Europe; and with equal fervour he emerged as an advocate of European union. At Zürich, on Sept. 19, 1946, he urged the formation of "a council of Europe" and himself attended the first assembly of the council at Strasbourg in 1949. Meanwhile, he busied himself with his great history, *The Second World War,* six volumes (1948–53).

The "iron curtain" speech

The general election of February 1950 afforded Churchill an opportunity to seek again a personal mandate. He abstained from the extravagances of 1945 and campaigned with his party rather than above it.

The electoral onslaught shook Labour but left them still in office. It took what Churchill called "one more heave" to defeat them in a second election, in October 1951. Churchill again took a vigorous lead in the campaign. He pressed the government particularly hard on its handling of the crisis caused by Iran's nationalization of British oil companies and in return had to withstand charges of warmongering. The Conservatives were returned with a narrow majority of 26, and Churchill became prime minister for the second time. He formed a government in which the more liberal Conservatives predominated, though the Liberal Party itself declined Churchill's suggestion of office. A prominent figure in the government was R.A. Butler, the progressive-minded chancellor of the Exchequer. Anthony Eden was foreign secretary. Some notable Churchillians were included, among them Lord Cherwell, who, as paymaster general, was principal scientific adviser with special responsibilities for atomic research and development.

**As prime minister again.** The domestic labours and battles of his administration were far from Churchill's main concerns. Derationing, decontrolling, rehousing, safeguarding the precarious balance of payments—these were relatively noncontroversial policies; only the return of nationalized steel and road transport to private hands aroused excitement. Critics sometimes complained of a lack of prime ministerial direction in these areas and, indeed, of a certain slackness in the reins of government. Undoubtedly Churchill was getting older and reserving more and more of his energies for what he regarded as the supreme issues, peace and war. He was convinced that Labour had allowed the transatlantic relationship to sag, and one of his first acts was to visit Washington (and also Ottawa) in January 1952 to repair the damage he felt had been done. The visit helped to check U.S. fears that the British would desert the Korean War, harmonized attitudes toward German rearmament and, distasteful though it was to Churchill, resulted in the acceptance of a U.S. naval commander in chief of the eastern Atlantic. It did not produce that sharing of secrets of atom bomb manufacture that Churchill felt had unfairly lapsed after the war. To the disappointment of many, Churchill's advocacy of European union did not result in active British participation; his government confined itself to endorsement from the sidelines, though in 1954, faced with the collapse of the European Defense Community, Churchill and Eden came forward with a pledge to maintain British troops on the Continent for as long as necessary.

Advocacy of European unity

The year 1953 was in many respects a gratifying one for Churchill. It brought the coronation of Queen Elizabeth II, which drew out all his love of the historic and symbolic. He personally received two notable distinctions, the Order of the Garter and the Nobel Prize for Literature. However, his hopes for a revitalized "special relationship" with Pres. Dwight D. Eisenhower during his tenure in the White House, beginning in 1953, were largely frustrated. A sudden stroke in June, which caused partial paralysis, obliged Churchill to cancel a planned Bermuda meeting at which he hoped to secure Eisenhower's agreement to summit talks with the Russians. By October, Churchill had made a remarkable recovery and the meeting was held in December. But it did not yield results commensurate with Churchill's hopes. The two leaders, for all their amity, were not the men they once were; their subordinates, John Foster Dulles and Anthony Eden, were antipathetic; and, above all, the role and status of each country had changed. In relation to the Far East in particular there was a persistent failure to see eye to eye. Though Churchill and Eden visited Washington, D.C., in June 1954 in hopes of securing U.S. acceptance of the Geneva Accords designed to bring an end to the war in Indochina, their success was limited. Over Egypt, however, Churchill's conversion to

an agreement permitting a phased withdrawal of British troops from the Suez base won Eisenhower's endorsement and encouraged hopes, illusory as it subsequently appeared, of good Anglo-American cooperation in this area. In 1955, "arming to parley," Churchill authorized the manufacture of a British hydrogen bomb while still striving for a summit conference. Age, however, robbed him of this last triumph. His powers were too visibly failing. His 80th birthday, on Nov. 30, 1954, had been the occasion of a unique all-party ceremony of tribute and affection in Westminster Hall. But the tribute implied a pervasive assumption that he would soon retire. On April 5, 1955, his resignation took place, only a few weeks before his chosen successor, Sir Anthony Eden, announced plans for a four-power conference at Geneva.

**Retirement and death.** Although Churchill laid down the burdens of office amid the plaudits of the nation and the world, he remained in the House of Commons (declining a peerage) to become "father of the house" and even, in 1959, to fight and win yet another election. He also published another major work, *A History of the English-Speaking Peoples,* four volumes (1956–58). But his health declined, and his public appearances became rare. On April 9, 1963, he was accorded the unique distinction of having an honorary U.S. citizenship conferred on him by an act of Congress. His death at his London home on Jan. 24, 1965, was followed by a state funeral at which almost the whole world paid tribute. He was buried in the family grave in Bladon churchyard, Oxfordshire.

**Assessment.** In any age and time a man of Churchill's force and talents would have left his mark on events and society. A gifted journalist, a biographer and historian of classic proportions, an amateur painter of talent, an orator of rare power, a soldier of courage and distinction, Churchill, by any standards, was a man of rare versatility. But it was as a public figure that he excelled. His experience of office was second only to Gladstone's, and his gifts as a parliamentarian hardly less, but it was as a wartime leader that he left his indelible imprint on the history of Britain and on the world. In this capacity, at the peak of his powers, he united in a harmonious whole his liberal convictions about social reform, his deep conservative devotion to the legacy of his nation's history, his unshakable resistance to tyranny from the right or from the left, and his capacity to look beyond Britain to the larger Atlantic community and the ultimate unity of Europe. A romantic, he was also a realist, with an exceptional sensitivity to tactical considerations at the same time as he unswervingly adhered to his strategical objectives. A fervent patriot, he was also a citizen of the world. An indomitable fighter, he was a generous victor. Even in the transition from war to peace, a phase in which other leaders have often stumbled, he revealed, at an advanced age, a capacity to learn and to adjust that was in many respects superior to that of his younger colleagues.

**MAJOR WORKS**

HISTORY: *The Story of the Malakand Field Force* (1898); *The River War* (1899); *The World Crisis* (1923–29); *The Unknown War: The Eastern Front* (1931); *The Second World War* (1948–53); *A History of the English-Speaking Peoples* (1956–58).

BIOGRAPHY AND AUTOBIOGRAPHY: *Lord Randolph Churchill* (1906); *My African Journey* (1908); *My Early Life* (1930); *Marlborough: His Life and Times* (1933–38).

SPEECHES: *Into Battle* (1941); *The Unrelenting Struggle* (1942); *The End of the Beginning* (1943); *Onwards to Victory* (1944); *The Dawn of Liberation* (1945); *Victory* (1946); *Secret Session Speeches* (1946); *The Sinews of Peace* (1948); *Europe Unite* (1950); *In the Balance* (1951); *Stemming the Tide* (1953); *The Unwritten Alliance* (1961). The speeches have been collected in *Winston S. Churchill: His Complete Speeches, 1897–1963,* 8 vol. (1974).

OTHER WORKS: *Savrola* (1900); *Thoughts and Adventures* (1932); *Painting As a Pastime* (1948).

**BIBLIOGRAPHY.** The official biography, *Winston S. Churchill* (1966– ), was begun by Churchill's son, RANDOLPH S. CHURCHILL, and continued by MARTIN GILBERT, each volume covering a successive span of years and supported by companion volumes of documents. Churchill's own writings are an indispensable autobiographical source; see *While England Slept: A Survey of World Affairs, 1932–1938,* with preface and notes by RANDOLPH S. CHURCHILL (1938, reprinted 1971; U.K. title, *Arms and the Covenant: Speeches,* 1938, reissued 1975). VIOLET BONHAM CARTER, *Winston Churchill: An Intimate Portrait* (U.K. title, *Winston Churchill As I Knew Him,* 1965), is a vivid memoir. LORD CHARLES MORAN, *Churchill: The Struggle for Survival, 1940–1965* (1966, reissued 1976), written by his physician, gives intimate glimpses of his late years. See also HENRY PELLING, *Winston Churchill* (1974, reissued 1977), a comprehensive biography; JOSEPH P. LASH, *Roosevelt and Churchill, 1939–1941: The Partnership That Saved the West* (1976), a study that illustrates the importance of Churchill's strong personality and the force of his ideas; FRANÇOIS KERSAUDY, *Churchill and de Gaulle* (1981, reissued 1983); and WARREN F. KIMBALL (ed.), *Churchill and Roosevelt: The Complete Correspondence,* 3 vol. (1984), both studies of their wartime relationships; and JOHN COLVILLE, *The Fringes of Power: Downing Street Diaries, 1939–1955* (1985), a portrait of Churchill by the civil servant who was his private secretary during most of World War II and again in 1951–55.

(H.G.N.)

# Circulation and Circulatory Systems

Within any living organism, the process of circulation involves all of the fluids of the body and permits a continuous integration among various tissues. Important in this process is the intake of metabolic materials, the conveyance of these materials throughout the organism, and the return of harmful by-products to the environment.

Invertebrate animals have a great variety of liquids, cells, and modes of circulation, though many invertebrates have what is called an open system, in which fluid passes more or less freely throughout the tissues or defined areas of tissue. All vertebrates, however, have a closed system—that is, their circulatory system transmits fluid through an intricate network of vessels. This system contains two fluids, blood and lymph, and functions by means of two interacting modes of circulation, the cardiovascular system and the lymphatic system; both the fluid components and the vessels through which they flow reach their greatest elaboration and specialization in the mammalian systems and, particularly, in the human body.

For this reason the human cardiovascular system is discussed in a separate section of this article and is preceded by a general survey of the circulatory systems of all other animals, from the single-celled to the vertebrate. Finally, the last section of the article is concerned with the diseases and disorders of the human cardiovascular system and includes a discussion of various surgical treatments of the heart.

A full treatment of human blood and its various components can be found in the article BLOOD. A discussion of how the systems of circulation, respiration, and metabolism work together within an animal organism is found in the article RESPIRATION AND RESPIRATORY SYSTEMS. (Ed.)

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 421 and 422, and the *Index.*

This article is divided into the following sections:


Main features of circulatory systems 378
- General features of circulation 378
- Body fluids 378
- Fluid compartments 378

Invertebrate circulatory systems 379
- Basic physicochemical considerations 379
- Animals without independent vascular systems 379
- Vascular systems 379
  - Blood
  - Hearts
- Acoelomates and pseudocoelomates 381
- Coelomates 382
  - Annelida
  - Echiura
  - Mollusca
  - Brachiopoda
  - Arthropoda
  - Echinodermata
  - Hemichordata
  - Chordata

The vertebrate circulatory system 384
- The basic vertebrate pattern 384
  - The plan
  - Evolutionary trends
- Modifications among the vertebrate classes 386
  - Fishes
  - Amphibians
  - Reptiles
  - Birds
  - Mammals
- Embryonic development of the circulatory system 388
- Biodynamics of vertebrate circulation 389
  - Blood pressure and blood flow
  - Electrical activity
  - Control of heartbeat and circulation

The human cardiovascular system 390
- The heart 390
  - Description
  - Structure and function
  - Heartbeat
- The blood vessels 395
  - The arteries
  - The veins
  - The capillaries
- Evaluating the cardiovascular system 399
  - Invasive techniques
  - Noninvasive techniques

Cardiovascular system diseases and disorders 400
- Congenital heart disease 400
  - Abnormalities of individual heart chambers
  - Abnormalities of the atrial septum
  - Abnormalities of the ventricular septum
  - Abnormal origins of the great arteries
  - Abnormalities of the valves
  - Abnormalities of the myocardium and endocardium
  - Abnormalities of the coronary arteries
  - Abnormalities of the aorta
  - Anomalous pulmonary venous return
  - Anomalies of the vena cava
- Acquired heart disease 402
  - Atherosclerosis
  - Coronary artery disease
  - Coronary heart disease
  - Rheumatic heart disease
  - The heart, the pulmonary artery, and the aorta
  - Diseases of the endocardium and valves
  - Diseases of the myocardium
  - Diseases of the pericardium
- Disturbances in rhythm and conduction 408
  - Atrial arrhythmias
  - Atrioventricular node mechanisms
  - Ventricular arrhythmias
- Heart failure 409
  - Left ventricular failure
  - Pulmonary edema
  - Right heart failure
  - Incipient heart failure
  - Mild to moderate congestive heart failure
  - Severe congestive heart failure
  - Cardiogenic shock
- Surgical treatment of the heart 410
  - Cardiopulmonary bypass
  - Congenital cardiac defects
  - Acquired heart defects
- Diseases of the arteries 412
  - Occlusive disease
  - Nonocclusive disease
  - Functional disease
- Diseases of the veins 414
  - Organic disease
  - Functional disease
- Diseases of the capillaries 414
- Hemodynamic disorders 414
  - Hypertension
  - Hypotension
  - Syncope
- Physiological shock 414
  - Shock due to inadequate blood volume
  - Shock due to inadequate cardiac output
  - Shock due to increased circulatory capacity
  - Refractory and irreversible shock
  - Identification of the causes of shock
  - Prognosis

Bibliography 416

## Main features of circulatory systems

### GENERAL FEATURES OF CIRCULATION

All living organisms take in molecules from their environments, use them to support the metabolism of their own substance, and release by-products back into the environment. The internal environment differs more or less greatly from the external environment, depending on the species. It is normally maintained at constant conditions by the organism so that it is subject to relatively minor fluctuations. In individual cells, either as independent organisms or as parts of the tissues of multicellular animals, molecules are taken in either by their direct diffusion through the cell wall or by the formation by the surface membrane of vacuoles that carry some of the environmental fluid containing dissolved molecules. Within the cell, cyclosis (streaming of the fluid cytoplasm) distributes the metabolic products.

Molecules are normally conveyed between cells and throughout the body of multicellular organisms in a circulatory fluid, called blood, through special channels, called blood vessels, by some form of pump, which, if restricted in position, is usually called a heart. In vertebrates blood and lymph (the circulating fluids) have an essential role in maintaining homeostasis (the constancy of the internal environment) by distributing substances to parts of the body when required and by removing others from areas in which their accumulation would be harmful.

The animal kingdom can be divided into two subkingdoms—Protozoa, which contains the unicellular animals; and Metazoa, which contains all the multicellular animals. The Metazoa may be further divided into the Parazoa, which contains the Porifera (sponges), which lack defined tissues and have no organs; and Eumetazoa, which contains all of the remaining multicellular (metazoan) animals.

The Eumetazoa

One eumetazoan phylum, Cnidaria (Coelenterata)—which includes sea anemones, jellyfish, and corals—has a diploblastic level of organization (*i.e.,* its members have two layers of cells). The outer layer, called the ectoderm, and the inner layer, called the endoderm, are separated by an amorphous, acellular layer called the mesoglea; for these animals, bathing both cellular surfaces with environmental fluid is sufficient to supply their metabolic needs. All other major eumetazoan phyla are triploblastic (*i.e.,* their members have three layers of cells), with the third cellular layer, called the mesoderm, developing between the endoderm and ectoderm. At its simplest, the mesoderm provides a network of packing cells around the animal's organs; this is probably best exhibited in the phylum Platyhelminthes (flatworms).

Nematoda, Rotifera, and a number of other smaller eumetazoan classes and phyla have a fluid-filled cavity, called the pseudocoelom, that arises from an embryonic cavity and contains the internal organs free within it. All other eumetazoans have a body cavity, the coelom, which originates as a cavity in the embryonic mesoderm. Mesoderm lines the coelom and forms the peritoneum, which also surrounds and supports the internal organs. While this increase in complexity allows for increase in animal size, it has certain problems. As the distances from metabolizing cells to the source of metabolites (molecules to be metabolized) increases, a means of distribution around the body is necessary for all but the smallest coelomates.

Many invertebrate animals are aquatic and the problem of supplying fluid is not critical. For terrestrial organisms, however, the fluid reaching the tissues comes from water that has been drunk, absorbed in the alimentary canal, and passed to the bloodstream. Fluid may leave the blood, usually with food and other organic molecules in solution, and pass to the tissues, from which it returns in the form of lymph. Especially in the vertebrates, lymph passes through special pathways, called lymphatic channels, to provide the lymphatic circulation.

In many invertebrates, however, the circulating fluid is not confined to distinct vessels, and it more or less freely bathes the organs directly. The functions of both circulating and tissue fluid are thus combined in the fluid, often known as hemolymph. The possession of a blood supply and coelom, however, does not exclude the circulation of environmental water through the body. Members of the phylum Echinodermata (starfishes and sea urchins, for example) have a complex water vascular system used mainly for locomotion.

Circulatory systems

An internal circulatory system transports essential gases and nutrients around the body of an organism, removes unwanted products of metabolism from the tissues, and carries these products to specialized excretory organs, if present. Although a few invertebrate animals circulate external water through their bodies for respiration, and, in the case of cnidarians, nutrition, most species circulate an internal fluid, called blood.

There may also be external circulation that sets up currents in the environmental fluid to carry it over respiratory surfaces and, especially in the case of sedentary animals, to carry particulate food that is strained out and passed to the alimentary canal. Additionally, the circulatory system may assist the organism in movement; for example, protoplasmic streaming in amoeboid protozoans circulates nutrients and provides pseudopodal locomotion. The hydrostatic pressure built up in the circulatory systems of many invertebrates is used for a range of whole-body and individual-organ movement.

### BODY FLUIDS

The fluid compartments of animals consist of intracellular and extracellular components. The intracellular component includes the body cells and, where present, the blood cells, while the extracellular component includes the tissue fluid, coelomic fluid, and blood plasma. In all cases the major constituent is water derived from the environment. The composition of the fluid varies markedly depending on its source and is regulated more or less precisely by homeostasis.

Blood and coelomic fluid are often physically separated by the blood-vessel walls; where a hemocoel (a blood-containing body cavity) exists, however, blood rather than coelomic fluid occupies the cavity. The composition of blood may vary from what is little more than the environmental water containing small amounts of dissolved nutrients and gases to the highly complex tissue containing many cells of different types found in mammals.

Lymph essentially consists of blood plasma that has left the blood vessels and has passed through the tissues. It is generally considered to have a separate identity when it is returned to the bloodstream through a series of vessels independent of the blood vessels and the coelomic space. Coelomic fluid itself may circulate in the body cavity. In most cases this circulation has an apparently random nature, mainly because of movements of the body and organs. In some phyla, however, the coelomic fluid has a more important role in internal distribution and is circulated by ciliary tracts.

### FLUID COMPARTMENTS

Blood is circulated through vessels of the blood vascular system. Blood is moved through this system by some form of pump. The simplest pump, or heart, may be no more than a vessel along which a wave of contraction passes to propel the blood. This simple, tubular heart is adequate where low blood pressure and relatively slow circulation rates are sufficient to supply the animal's metabolic requirements, but it is inadequate in larger, more active, and more demanding species. In the latter animals, the heart is usually a specialized, chambered, muscular pump that receives blood under low pressure and returns it under higher pressure to the circulation. Where the flow of blood is in one direction, as is normally the case, valves in the form of flaps of tissue prevent backflow.

Heartbeat

A characteristic feature of hearts is that they pulsate throughout life and any prolonged cessation of heartbeat is fatal. Contractions of the heart muscle may be initiated in one of two ways. In the first, the heart muscle may have an intrinsic contractile property that is independent of the nervous system. This myogenic contraction is found in all vertebrates and some invertebrates. In the second, the heart is stimulated by nerve impulses from outside the heart muscle. The hearts of other invertebrates exhibit this neurogenic contraction.

The cardiac cycle

Chambered hearts, as found in vertebrates and some larger invertebrates, consist of a series of interconnected muscular compartments separated by valves. The first chamber, the auricle, acts as a reservoir to receive the blood that then passes to the second and main pumping chamber, the ventricle. Expansion of a chamber is known as diastole and contraction as systole. As one chamber undergoes systole the other undergoes diastole, thus forcing the blood forward. The series of events during which blood is passed through the heart is known as the cardiac cycle.

Contraction of the ventricle forces the blood into the vessels under pressure, known as the blood pressure. As contraction continues in the ventricle, the rising pressure is sufficient to open the valves that had been closed because of attempted reverse blood flow during the previous cycle. At this point the ventricular pressure transmits a high-speed wave, the pulse, through the blood of the arterial system. The volume of blood pumped at each contraction of the ventricle is known as the stroke volume, and the output is usually dependent on the animal's activity.

Blood vessels

After leaving the heart, the blood passes through a series of branching vessels of steadily decreasing diameter. The smallest branches, only a few micrometres (there are about 25,000 micrometres in one inch) in diameter, are the capillaries, which have thin walls through which the fluid part of the blood may pass to bathe the tissue cells. The capillaries also pick up metabolic end products and carry them into larger collecting vessels that eventually return the blood to the heart. In vertebrates there are structural differences between the muscularly walled arteries, which carry the blood under high pressure from the heart, and the thinner walled veins, which return it at much reduced pressure. Although such structural differences are less apparent in invertebrates, the terms artery and vein are used for vessels that carry blood from and to the heart, respectively.

The closed circulatory system found in vertebrates is not universal; a number of invertebrate phyla have an "open" system. In the latter animals, the blood leaving the heart passes into a series of open spaces, called sinuses, where it bathes internal organs directly. Such a body cavity is called a hemocoel, a term that reflects the amalgamation of the blood system and the coelom.

## Invertebrate circulatory systems

### BASIC PHYSICOCHEMICAL CONSIDERATIONS

Optimal metabolism

To maintain optimum metabolism, all living cells require a suitable environment, which must be maintained within relatively narrow limits. An appropriate gas phase (*i.e.*, suitable levels of oxygen and other gases), an adequate and suitable nutrient supply, and a means of disposal of unwanted products are all essential.

Direct diffusion through the body surface supplies the necessary gases and nutrients for small organisms, but even some single-celled protozoa have a rudimentary circulatory system. Cyclosis in many ciliates carries food vacuoles—which form at the forward end of the gullet (cytopharynx)—on a more or less fixed route around the cell, while digestion occurs to a fixed point of discharge.

For most animal cells, the supply of oxygen is largely independent of the animal and therefore is a limiting factor in its metabolism and ultimately in its structure and distribution. The nutrient supply to the tissues, however, is controlled by the animal itself, and, because both major catabolic end products of metabolism—ammonia ($NH_3$) and carbon dioxide ($CO_2$)—are more soluble than oxygen ($O_2$) in water and the aqueous phase of the body fluids, they tend not to limit metabolic rates. The diffusion rate of $CO_2$ is less than that of $O_2$, but its solubility is 30 times that of oxygen. This means that the amount of $CO_2$ diffusing is 26 times as high as for oxygen at the same temperature and pressure.

The oxygen available to a cell depends on the concentration of oxygen in the external environment and the efficiency with which it is transported to the tissues. Dry air at atmospheric pressure contains about 21 percent oxygen, the percentage of which decreases with increasing altitude. Well-aerated water has the same percentage of oxygen as the surrounding air; however, the amount of dissolved oxygen is governed by temperature and the presence of other solutes. For example, seawater contains 20 percent less oxygen than fresh water under the same conditions.

Rate of diffusion

The rate of diffusion depends on the shape and size of the diffusing molecule, the medium through which it diffuses, the concentration gradient, and the temperature. These physicochemical constraints imposed by gaseous diffusion have a relationship with animal respiration. Investigations have suggested that a spherical organism larger than 0.5 millimetre (0.02 inch) radius would not obtain enough oxygen for the given metabolic rate, and so a supplementary transport mechanism would be required. Many invertebrates are small, with direct diffusion distances of less than 0.5 millimetre. Considerably larger species, however, still survive without an internal circulatory system.

### ANIMALS WITHOUT INDEPENDENT VASCULAR SYSTEMS

A sphere represents the smallest possible ratio of surface area to volume; modifications in architecture, reduction of metabolic rate, or both may be exploited to allow size increase. Sponges (Figure 1A) overcome the problem of oxygen supply and increase the chance of food capture by passing water through their many pores using ciliary action. The level of organization of sponges is that of a coordinated aggregation of largely independent cells with poorly defined tissues and no organ systems. The whole animal has a relatively massive surface area for gaseous exchange, and all cells are in direct contact with the passing water current.

Among the eumetazoan (multicellular) animals the cnidarians (sea anemones, corals, and jellyfish) are diploblastic, the inner endoderm and outer ectoderm being separated by an acellular mesoglea. Sea anemones and corals may also grow to considerable size and exhibit complex external structure that, again, has the effect of increasing surface area. Their fundamentally simple structure—with a gastrovascular cavity continuous with the external environmental water—allows both the endodermal and ectodermal cells of the body wall access to aerated water, permitting direct diffusion (Figure 1B).

Circulation in jellyfish

This arrangement is found in a number of other invertebrates, such as Ctenophora (comb jellies), and is exploited further by jellyfish, which also show a rudimentary internal circulatory system. The thick, largely acellular, gelatinous bell of a large jellyfish may attain a diameter of 40 centimetres (16 inches) or more. The gastrovascular cavity is modified to form a series of water-filled canals that ramify through the bell and extend from the central gastric pouches to a circular canal that follows the periphery of the umbrella. Ciliary activity within the canals slowly passes food particles and water, taken in through the mouth, from the gastric pouches (where digestion is initiated) to other parts of the body. Ciliary activity is a relatively inefficient means of translocating fluids, and it may take up to half an hour to complete a circulatory cycle through even a small species. To compensate for the inefficiency of the circulation, the metabolic rate of the jellyfish is low, and organic matter makes up only a small proportion of the total body constituents. The central mass of the umbrella may be a considerable distance from either the exumbrella surface or the canal system, and, while it contains some wandering amoeboid cells, its largely acellular nature means that its metabolic requirements are small.

### VASCULAR SYSTEMS

While ciliary respiratory currents are sufficient to supply the requirements of animals with simple epithelial tissues and low metabolic rates, most species whose bodies contain a number of organ systems require a more efficient circulatory system. Many invertebrates and all vertebrates have a closed vascular system in which the circulatory fluid is totally confined within a series of vessels consisting of arteries, veins, and fine linking capillaries. Insects, most crustaceans, and many mollusks, however, have an open system in which the circulating fluid passes somewhat freely among the tissues before being collected and recirculated.

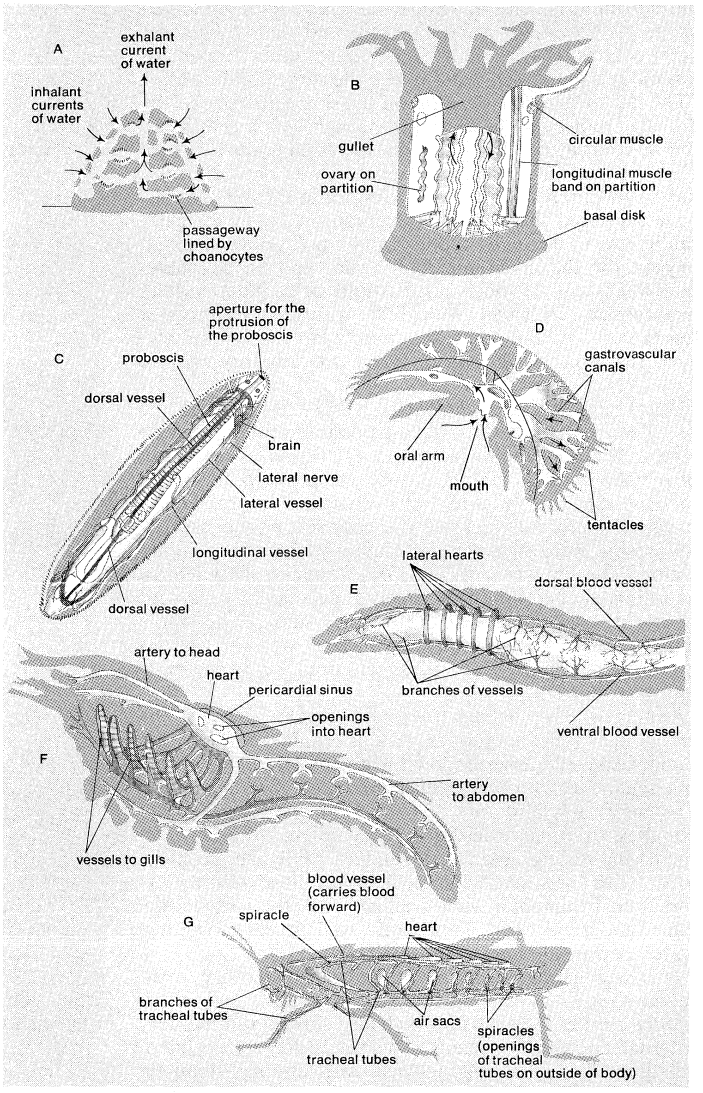



Figure 1: Vertical section through (A) sponge and (B) sea anemone, with water circulation indicated by arrows; (C) circulation in nemertean worm; (D) internal arrangement of ciliated endodermal canals in jellyfish; arrows show the direction of water currents; (E) circulation in earthworm; (F) circulation in decapod crustacean; (G) circulation in insect (grasshopper).

From (B,F) R. Buchsbaum, *Animals Without Backbones* (1948), University of Chicago Press; (C) T.J. Parker and W.A. Haswell, *Textbook of Zoology*, St. Martin's Press, Inc., Macmillan & Co., Ltd.; (E,G) Biological Sciences Curriculum Study, *Biological Science: An Inquiry Into Life*, second edition, 1968, Harcourt Brace Jovanovich

The distinction between open and closed circulatory systems may not be as great as was once thought; some crustaceans have vessels with dimensions similar to those of vertebrate capillaries before opening into tissue sinuses. The circulatory fluid in open systems is strictly hemolymph, but the term "blood" is commonly used to denote the transporting medium in both open and closed systems. Compared with closed systems, open circulatory systems generally work at lower pressures, and the rate of fluid return to the heart is slower. Blood distribution to individual organs is not regulated easily, and the open system is not as well-adapted for rapid response to change.

**Blood.** The primary body cavity (coelom) of triploblastic multicellular organisms arises from the central mesoderm, which emerges from between the endoderm and ectoderm during embryonic development. The fluid of the coelom containing free mesodermal cells constitutes the blood and lymph. The composition of blood varies between different organisms and within one organism at different stages during its circulation. Essentially, however, the blood consists of an aqueous plasma containing sodium, potassium, calcium, magnesium, chloride, and sulfate ions; some trace elements; a number of amino acids; and possibly a protein known as a respiratory pigment. If present in invertebrates, the respiratory pigments are normally dissolved in the plasma and are not enclosed in blood cells. The constancy of the ionic constituents of blood and their similarity to seawater have been used by some scientists as evidence of a common origin for life in the sea.

Composition of blood

An animal's ability to control its gross blood concentration (*i.e.*, the overall ionic concentration of the blood) largely governs its ability to tolerate environmental changes. In many marine invertebrates, such as echinoderms and some mollusks, the osmotic and ionic characteristics of the blood closely resemble those of seawater. Other aquatic, and all terrestrial, organisms, however, maintain blood concentrations that differ to some extent from their environments and thus have a greater potential range of habitats. In addition to maintaining the overall stability of the internal environment, blood has a range of other functions. It is the major means of transport of nutrients, metabolites, excretory products, hormones, and gases, and it may provide the mechanical force for such diverse processes as hatching and molting in arthropods and burrowing in bivalve mollusks.

Hemocytes

Invertebrate blood may contain a number of cells (hemocytes) arising from the embryonic mesoderm. Many different types of hemocytes have been described in different species, but they have been studied most extensively in insects, in which four major types and functions have been suggested: (1) phagocytic cells that ingest foreign particles and parasites and in this way may confer some nonspecific immunity to the insect; (2) flattened hemocytes that adhere to the surface of the invader and remove its supply of oxygen, resulting in its death; metazoan parasites that are too large to be engulfed by the phagocytic cells may be encapsulated by these cells instead; (3) hemocytes that assist in the formation of connective tissue and the secretion of mucopolysaccharides during the formation of basement membranes; they may be involved in other aspects of intermediate metabolism as well; and (4) hemocytes that are concerned with wound healing; the plasma of many insects does not coagulate, and either pseudopodia or secreted particles from hemocytes (cystocytes) trap other such cells to close the lesion until the surface of the skin regenerates.

While the solubility of oxygen in blood plasma is adequate to supply the tissues of some relatively sedentary invertebrates, more active animals with increased oxygen demands require an additional oxygen carrier. The oxygen carriers in blood take the form of metal-containing protein molecules that frequently are coloured and thus commonly known as respiratory pigments. The most widely distributed respiratory pigments are the red hemoglobins, which have been reported in all classes of vertebrates, in most invertebrate phyla, and even in some plants. Hemoglobins consist of a variable number of subunits, each containing an iron–porphyrin group attached to a protein. The distribution of hemoglobins in just a few members of a phylum and in many different phyla argues that the hemoglobin type of molecule must have evolved many times with similar iron–porphyrin groups and different proteins.

The green chlorocruorins are also iron–porphyrin pigments and are found in the blood of a number of families of marine polychaete worms. There is a close resemblance between chlorocruorin and hemoglobin molecules, and a number of species of a genus, such as those of *Serpula,* contain both, while some closely related species exhibit an almost arbitrary distribution. For example, *Spirorbis borealis* has chlorocruorin, *S. corrugatus* has hemoglobin, and *S. militaris* has neither.

The third iron-containing pigments, the hemerythrins, are violet. They differ structurally from both hemoglobin and chlorocruorin in having no porphyrin groups and containing three times as much iron, which is attached directly to the protein. Hemerythrins are restricted to a small number of animals, including some polychaete and sipunculid worms, the brachiopod *Lingula,* and some priapulids.

Hemocyanins

Hemocyanins are copper-containing respiratory pigments found in many mollusks (some bivalves, many gastropods,

and cephalopods) and arthropods (many crustaceans, some arachnids, and the horseshoe crab, *Limulus*). They are colourless when deoxygenated but turn blue on oxygenation. The copper is bound directly to the protein, and oxygen combines reversibly in the proportion of one oxygen molecule to two copper atoms.

The presence of a respiratory pigment greatly increases the oxygen-carrying capacity of blood; invertebrate blood may contain up to 10 percent oxygen with the pigment, compared with about 0.3 percent in the absence of the pigment. All respiratory pigments become almost completely saturated with oxygen even at oxygen levels, or pressures, below those normally found in air or water. The oxygen pressures at which the various pigments become saturated depend on their individual chemical characteristics and on such conditions as temperature, pH, and the presence of carbon dioxide.

In addition to their direct transport role, respiratory pigments may temporarily store oxygen for use during periods of respiratory suspension or decreased oxygen availability (hypoxia). They may also act as buffers to prevent large blood pH fluctuations, and they may have an osmotic function that helps to reduce fluid loss from aquatic organisms whose internal hydrostatic pressure tends to force water out of the body.

**Hearts.** All systems involving the consistent movement of circulating fluid require at least one repeating pump and, if flow is to be in one direction, usually some arrangement of valves to prevent backflow. The simplest form of animal circulatory pump consists of a blood vessel down which passes a wave of muscular contraction, called peristalsis, that forces the enclosed blood in the direction of contraction. Valves may or may not be present. This type of heart is widely found among invertebrates, and there may be many pulsating vessels in a single individual.

In the earthworm, the main dorsal (aligned along the back) vessel contracts from posterior to anterior 15 to 20 times per minute, pumping blood toward the head. At the same time, the five paired segmental lateral (side) vessels, which branch from the dorsal vessel and link it to the ventral (aligned along the bottom) vessel, pulsate with their own independent rhythms. Although unusual, it is possible for a peristaltic heart to reverse direction. After a series of contractions in one direction, the hearts of tunicates (sea squirts) gradually slow down and eventually stop. After a pause the heart starts again, with reverse contractions pumping the blood in the opposite direction.

An elaboration of the simple peristaltic heart is found in the tubular heart of most arthropods, in which part of the dorsal vessel is expanded to form one or more linearly arranged chambers with muscular walls. The walls are perforated by pairs of lateral openings (ostia) that allow blood to flow into the heart from a large surrounding sinus, the pericardium. The heart may be suspended by alary muscles, contraction of which expands the heart and increases blood flow into it. The direction of flow is controlled by valves arranged in front of the in-current ostia.

Chambered hearts

Chambered hearts with valves and relatively thick muscular walls are less commonly found in invertebrates but do occur in some mollusks, especially cephalopods (octopus and squid). Blood from the gills enters one to four auricles (depending on the species) and is passed back to the tissues by contraction of the ventricle. The direction of flow is controlled by valves between the chambers. The filling and emptying of the heart are controlled by regular rhythmical contractions of the muscular wall.

In addition to the main systemic heart, many species have accessory booster hearts at critical points in the circulatory system. Cephalopods have special muscular dilations, the branchial hearts, that pump blood through the capillaries, and insects may have additional ampullar hearts at the points of attachment of many of their appendages.

The control of heart rhythm may be either myogenic (originating within the heart muscle itself) or neurogenic (originating in nerve ganglia). The hearts of the invertebrate mollusks, like those of vertebrates, are myogenic. They are sensitive to pressure and fail to give maximum beats unless distended; the beats become stronger and more frequent with increasing blood pressure. Although under experimental conditions acetylcholine (a substance that transmits nerve impulses across a synapse) inhibits molluscan heartbeat, indicating some stimulation of the heart muscle by the nervous system, cardiac muscle contraction will continue in excised hearts with no connection to the central nervous system. Tunicate hearts have two noninnervated, myogenic pacemakers, one at each end of the peristaltic pulsating vessel. Separately, each pacemaker causes a series of normal beats followed by a sequence of abnormal ones; together, they provide periodic reversals of blood flow.

The control of heartbeat in most other invertebrates is neurogenic, and one or more nerve ganglia with attendant nerve fibres control contraction. Removal of the ganglia stops the heart, and the administration of acetylcholine increases its rate. Adult heart control may be neurogenic but not necessarily in all stages in the life cycle. The embryonic heart may show myogenic peristaltic contractions prior to innervation.

Heart rate differs markedly among species and under different physiological states of a given individual. In general it is lower in sedentary or sluggish animals and faster in small ones. The rate increases with internal pressure but often reaches a plateau at optimal pressures. Normally, increasing the body temperature 10° C (50° F) causes an increase in heart rate of two to three times. Oxygen availability and the presence of carbon dioxide affect the heart rate, and during periods of hypoxia the heart rate may decrease to almost a standstill to conserve oxygen stores.

Circulatory cycles

The time it takes for blood to complete a single circulatory cycle is also highly variable but tends to be much longer in invertebrates than in vertebrates. For example, in isolation, the circulation rate in mammals is about 10 to 30 seconds, for crustaceans about one minute, for cockroaches five to six minutes, and for other insects almost 30 minutes.

## ACOELOMATES AND PSEUDOCOELOMATES

At the simplest levels of metazoan organization, where there are at most two cell layers, the tissues are arranged in sheets. The necessity for a formal circulatory system does not exist, nor are the mesodermal tissues, normally forming one, present. The addition of the mesodermal layer allows greater complexity of organ development and introduces further problems in supplying all cells with their essential requirements.

Invertebrate phyla have developed a number of solutions to these problems; most but not all involve the development of a circulatory system: as described above, sponges and cnidarians (Figure 1D) permit all cells direct access to environmental water. Among the acoelomate phyla, the members of Platyhelminthes (flatworms) have no body cavity, and the space between the gut and the body wall, when present, is filled with a spongy organ tissue of mesodermal cells through which tissue fluids may percolate. Dorsoventral (back to front) flattening, ramifying gut ceca (cavities open at one end), and, in the endoparasitic flatworm forms, glycolytic metabolic pathways (which release metabolic energy in the absence of oxygen) reduce diffusion distances and the need for oxygen and allow the trematodes and turbellarians of this phylum to maintain their normal metabolic rates in the absence of an independent circulatory system. The greatly increased and specialized body surface of the cestodes (tapeworms) of this phylum has allowed them to dispense with the gut as well. Most of the other acoelomate invertebrate animals are small enough that direct diffusion constitutes the major means of internal transport.

One acoelomate phylum, Nemertea (proboscis worms), contains the simplest animals possessing a true vascular system (Figure 1C). In its basic form there may be only two vessels situated one on each side of the straight gut. The vessels unite anteriorly by a cephalic space and posteriorly by an anal space lined by a thin membrane. The system is thus closed, and the blood does not directly bathe the tissues. The main vessels are contractile, but blood flow is irregular and it may move backward or forward within an undefined circuit. The blood is usually colourless, although some species contain pigmented blood cells

whose function remains obscure; phagocytic amoebocytes are usually also present. Although remaining fundamentally simple, the system can grow more elaborate with the addition of extra vessels.

The pseudocoelom

Pseudocoelomate metazoans have a fluid-filled body cavity, the pseudocoelom, which, unlike a true coelom, does not have a cellular peritoneal lining. Most of the pseudocoelomates (*e.g.*, the classes Nematoda and Rotifera) are small and none possess an independent vascular system. Muscular body and locomotor movements may help to circulate nutrients within the pseudocoelom between the gut and the body wall. The lacunar system of channels within the body wall of the gutless acanthocephalans (spiny-headed worms) may represent a means of circulation of nutrients absorbed through the body wall. Hemoglobin has been found in the pseudocoelomic fluid of a number of nematodes, but its precise role in oxygen transport is not known.

## COELOMATES

Despite their greater potential complexity, many of the minor coelomate phyla (*e.g.*, Pogonophora, Sipuncula, and Bryozoa) contain small animals that rely on direct diffusion and normal muscular activity to circulate the coelomic fluid. All of the major and some of the minor phyla have well-developed blood vascular systems, often of open design.

**Annelida.** While some small segmented worms of the phylum Annelida have no separate circulatory system, most have a well-developed closed system. The typical arrangement is for the main contractile dorsal vessel to carry blood anteriorly while a number of vertical segmental vessels, often called hearts, carry it to the ventral vessel, in which it passes posteriorly (Figure 1E). Segmental branches supply and collect blood from the respiratory surfaces, the gut, and the excretory organs.

There is, however, great scope for variation on the basic circulatory pattern. Many species have a large intestinal sinus rather than a series of vessels supplying the gut, and there may be differences along the length of a single individual. The posterior blood may flow through an intestinal sinus, the medial flow through a dense capillary plexus, and the anterior flow through typical segmental capillaries. Much modification and complication may occur in species in which the body is divided into more or less distinct regions with specific functions.

Many polychaete worms (class Polychaeta), especially the fanworms but also representatives of other families, have many blind-ending contractile vessels. Continual reversals of flow within these vessels virtually replace the normal continuous-flow capillary system.

In most leeches (class Hirudinea), much of the coelomic space is filled with mesodermal connective tissue, leaving a series of interconnecting coelomic channels. A vascular system comparable to other annelids is present in a few species, but in most the coelomic channels containing blood (strictly coelomic fluid) have taken over the function of internal transport, with movement induced by contraction of longitudinal lateral channels.

Annelid blood

The blood of many annelids contains a respiratory pigment dissolved in the plasma, and the coelomic fluid of others may contain coelomic blood cells containing hemoglobin. The most common blood pigments are hemoglobin and chlorocruorin, but their occurrence does not fit any simple evolutionary pattern. Closely related species may have dissimilar pigments, while distant relatives may have similar ones. In many species the pigments function in oxygen transport, but in others they are probably more important as oxygen stores for use during periods of hypoxia.

In addition to internal circulation, many polychaete worms also set up circulatory currents for feeding and respiration. Tube-dwelling worms may use muscular activity to pass a current of oxygenated water containing food through their burrows, while filter-feeding fanworms use ciliary activity to establish complicated patterns of water flow through their filtering fans.

**Echiura.** The phylum Echiura (spoonworms) contains a small number of marine worms with a circulatory system of similar general pattern to that of the annelids. Main dorsal and ventral vessels are united by contractile circumintestinal vessels that pump the colourless blood. Coelomic fluid probably aids in oxygen transport and may contain some cells with hemoglobin.

**Mollusca.** With the exception of the cephalopods, members of the phylum Mollusca have an open circulatory system (Figure 2A). The chambered, myogenic heart normally has a pair of posterior auricles draining the gills and an anterior ventricle that pumps the blood through the anterior aorta to the tissue sinuses, excretory organs, and gills. Many gastropods lack a second set of gills, and in these the right auricle is vestigial or absent. The heart is enclosed within the coelomic cavity, which also surrounds part of the intestine. The single aorta branches, and blood is delivered into arterial sinuses, where it directly bathes the tissues. It is collected in a large venous cephalopedal sinus and, after passing through the excretory organs, returns to the gills. The hydrostatic pressure that develops in the blood sinuses of the foot, especially of bivalve mollusks, is used in locomotion. Blood flow into the foot is controlled by valves: as the pressure increases, the foot elongates and anchors into the substratum; muscular contraction then pulls the animal back down to the foot. This type of locomotion is seen most commonly in burrowing species, who move through the substratum almost exclusively by this means.

Like the annelids, many mollusks, especially the more sedentary bivalves, set up local feeding and respiratory currents. Fluid movement through the mantle cavity normally depends on muscular pumping through inhalant and exhalant siphons. Within the cavity itself, however, ciliary activity maintains continuous movement across the gill surfaces, collecting food particles and passing them to the mouth.

The cephalopods are more active than other mollusks

From (A,C) *Life: An Introduction to Biology*, 2nd ed., by George Gaylord Simpson and William S. Beck, © 1957, 1965 by Harcourt Brace Jovanovich, Inc., and reproduced with their permission; (B,D) H. Curtis, *Biology*, Worth Publishers, New York, 1968, pages 275, 297; (E) L. Hyman, *Comparative Vertebrate Anatomy* (1942), University of Chicago Press

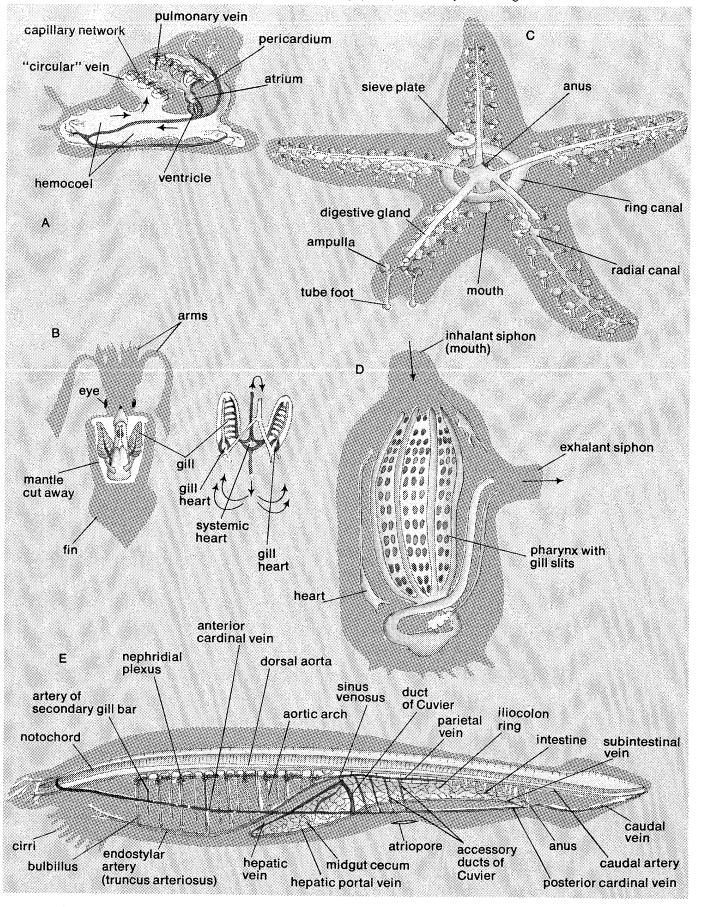



Figure 2: Circulatory system of (A) snail, (B) cephalopod, (C) starfish, (D) tunicate, (E) amphioxus.

Cephalopods

and consequently have higher metabolic rates and circulatory systems of a higher order of organization (Figure 2B). These systems are closed with distinct arteries, veins, and capillaries; the blood (6 percent of body weight) remains distinct from the interstitial fluid (15 percent of body weight). These relative percentages of body weight to blood volume are similar to those of vertebrates and differ markedly from those of species with open circulatory systems, in which hemolymph may constitute 40 to 50 percent of body weight.

The cephalopod heart usually consists of a median ventricle and two auricles. Arterial blood is pumped from the ventricle through anterior and posterior aortas that supply the head and body, respectively. It is passed through the capillary beds of the organs, is collected, and is returned to the heart through a major venous vessel, the vena cava. The vena cava bifurcates (divides into two branches) near the excretory organs, and each branch enters the nephridial sac before passing to the accessory hearts situated at the base of the gills. Veins draining the anterior and posterior mantle and the gonads merge with the branches of the vena cava before reaching the branchial hearts. Contraction of the branchial hearts increases the blood pressure and forces blood through the gill capillaries. The auricles then drain the gills of oxygenated blood.

The blood of most mollusks, including cephalopods, contains hemocyanin, although a few gastropods use hemoglobin. In the cephalopods the pigment unloads at relatively high oxygen pressures, indicating that it is used to transport rather than store oxygen.

Rapid cephalopod locomotion depends almost entirely on water pressure. During inhalation, muscular activity within the mantle wall increases the volume of the mantle cavity and water rushes in. Contraction of the circular mantle muscles closes the edge of the mantle and reduces its volume, forcing the enclosed water through the mobile funnel at high pressure. The force of water leaving the funnel propels the animal in the opposite direction.

**Brachiopoda.** Members of the phylum Brachiopoda (lamp shells) superficially resemble the mollusks but are not related. The circulatory system of brachiopods is open and consists of a small contractile heart situated over the gut, from which anterior and posterior channels supply sinuses in the wall of the gut, the mantle wall, and the reproductive organs.

**Arthropoda.** The blood vascular system of arthropods is open. The coelom is much reduced, and most of the spaces in the arthropod body are hemocoels. The tubular heart is dorsal and contained in a pericardial sinus. Blood is pumped from the heart through a series of vessels (arteries) that lead to the tissue sinuses. Although the blood flows freely through the tissues it may, especially in the larger species, be directed by membranes along a more or less constant pathway. The blood collects in a ventral sinus from which it is conducted back to the heart through one or more venous channels.

Variations in the circulatory patterns of the different classes of the phylum Arthropoda largely reflect the method of respiratory exchange and consequent function of the blood vascular system. Most of the aquatic species of the class Crustacea (Figure 1F) have gills with a well-developed circulatory system, including accessory hearts to increase blood flow through the gills. A small number of species lack gills and a heart, and oxygen is absorbed through the body surface; bodily movements or peristaltic gut contractions circulate the blood within the tissue spaces.

Oxygen transport in insects

In the mainly terrestrial class Insecta, the role of oxygen transport has been removed from the blood and taken over by the ramifying tracheal system that carries gaseous atmospheric oxygen directly to the consuming tissues (Figure 1G). Insects are able to maintain the high metabolic rates necessary for flight while retaining a relatively inefficient circulatory system.

Among the chelicerate (possessing fanglike front appendages) arthropods (for example, scorpions, spiders, ticks, and mites), the horseshoe crab, *Limulus,* has a series of book gills (gills arranged in membranous folds) on either side of the body into which blood from the ventral sinus passes for oxygenation prior to return to the heart. The largely terrestrial arachnids may have book lungs that occupy a similar position in the circulatory pathway, a tracheal system comparable to that of insects, or, in the case of smaller species, reduced tracheal and vascular systems in which contractions of the body muscles cause blood circulation through the sinus network.

The legs of spiders are unusual because they lack extensor muscles and because blood is used as hydraulic fluid to extend the legs in opposition to flexor muscles. The blood pressure of a resting spider is equal to that of a human being and may double during activity. The high pressure is maintained by valves in the anterior aorta and represents an exception to the general rule that open circulatory systems only function at low pressure.

**Echinodermata.** The circulatory systems of echinoderms (sea urchins, starfishes, and sea cucumbers) are complicated as they have three largely independent fluid systems. The large fluid-filled coelom that surrounds the internal organs constitutes the major medium for internal transport. Circulatory currents set up by the ciliated cells of the coelomic lining distribute nutrients from the gut to the body wall. Phagocytic coelomocytes are present, and in some species these contain hemoglobin. The coelomic fluid has the same osmotic pressure as seawater, and the inability to regulate that pressure has confined the echinoderms to wholly marine habitats.

The blood-vascular (hemal) system is reduced and consists of small, fluid-filled sinuses that lack a distinct lining. The system is most highly developed in the holothurians (sea cucumbers), in which it consists of an anterior hemal ring and radial hemal sinuses. The most prominent features are the dorsal and ventral sinuses, which accompany the intestine and supply it through numerous smaller channels. The dorsal sinus is contractile, and fluid is pumped through the intestinal sinuses into the ventral sinus and thence to the hemal ring. Most members of the class Holothuroidea have a pair of respiratory trees, located in the coelom on either side of the intestine, which act as the major sites for respiratory exchange. Each tree consists of a main tubular trunk with numerous side branches, each ending in a small vesicle. Water is passed through the tubules by the pumping action of the cloaca. The branches of the left tree are intermingled with the intestinal hemal sinuses and provide a means of oxygenating the blood via the coelomic fluid. The right tree is free in the coelomic fluid and has no close association with the hemal system. Respiratory exchange in other echinoderms is through thin areas of the body wall, and the hemal system tends to be reduced.

Sea cucumbers

The water vascular system of echinoderms is best developed in the starfishes and functions as a means of locomotion and respiratory exchange. The entire system consists of a series of fluid-filled canals lined with ciliated epithelium and derived from the coelom. The canals connect to the outside through a porous, button-shaped plate, called the madreporite, which is united via a duct (the stone canal) with a circular canal (ring canal) that circumvents the mouth (Figure 2C). Long canals radiate from the water ring into each arm. Lateral canals branch alternately from the radial canals, each terminating in a muscular sac (or ampulla) and a tube foot (podium), which commonly has a flattened tip that can act as a sucker. Contraction of the sac results in a valve in the lateral canal closing as the contained fluid is forced into the podium, which elongates. On contact with the substratum, the centre of the distal end of the podium is withdrawn, resulting in a partial vacuum and adhesion that is aided by the production of a copious adhesive secretion. Withdrawal results from contraction of the longitudinal muscles of the podia.

**Hemichordata.** Among the phylum Hemichordata are the enteropneusts (acornworms), which are worm-shaped inhabitants of shallow seas and have a short, conical proboscis, which gives them their common name. The vascular system of the Enteropneusta is open, with two main contractile vessels and a system of sinus channels. The colourless blood passes forward in the dorsal vessel, which widens at the posterior of the proboscis into a space, the contractile wall of which pumps the blood into the

glomerulus, an organ formed from an in-tucking of the hind wall of the proboscis cavity. From the glomerulus the blood is collected into two channels that lead backward to the ventral longitudinal vessel. This vessel supplies the body wall and gut with a network of sinuses that eventually drain back into the dorsal vessel.

**Chordata.** The phylum Chordata contains all animals that possess, at some time in their life cycles, a stiffening rod (the notochord), as well as other common features. The subphylum Vertebrata is a member of this phylum and will be discussed later (see below *The vertebrate circulatory system*). All other chordates are called protochordates and are classified into two groups: Tunicata and Cephalochordata.

The blood-vascular system of the tunicates, or sea squirts (Figure 2D), is open, the heart consisting of no more than a muscular fold in the pericardium. There is no true heart wall or lining and the whole structure is curved or U-shaped, with one end directed dorsally and the other ventrally. Each end opens into large vessels that lack true walls and are merely sinus channels. The ventral vessel runs along the ventral side of the pharynx and branches to form a lattice around the slits in the pharyngeal wall through which the respiratory water currents pass. Blood circulating through this pharyngeal grid is provided with a large surface area for gaseous exchange. The respiratory water currents are set up by the action of cilia lining the pharyngeal slits and, in some species, by regular muscular contractions of the body wall. Dorsally, the network of pharyngeal blood vessels drains into a longitudinal channel that runs into the abdomen and breaks up into smaller channels supplying the digestive loop of the intestine and the other visceral organs. The blood passes into a dorsal abdominal sinus that leads back to the dorsal side of the heart. The circulatory system of the sea squirt is marked by periodic reversals of blood flow caused by changes in the direction of peristaltic contraction of the heart.

Blood of sea squirts

Sea squirt blood has a slightly higher osmotic pressure than seawater and contains a number of different types of amoebocytes, some of which are phagocytic and actively migrate between the blood and the tissues. The blood of some sea squirts also contains green cells, which have a unique vanadium-containing pigment of unknown function.

Amphioxus (*Branchiostoma lanceolatum*) is a cephalochordate that possesses many typical vertebrate features but lacks the cranial cavity and vertebral column of the true vertebrate. Its circulatory pattern differs from that of most invertebrates as the blood passes forward in the ventral and backward in the dorsal vessels (Figure 2E). A large sac, the sinus venosus, is situated below the posterior of the pharynx and collects blood from all parts of the body. The blood passes forward through the subpharyngeal ventral aorta, from which branches carry it to small, accessory, branchial hearts that pump it upward through the gill arches. The oxygenated blood is collected into two dorsal aortas that continue forward into the snout and backward to unite behind the pharynx. The single median vessel thus formed branches to vascular spaces and the intestinal capillaries. Blood from the gut collects in a median subintestinal vein and flows forward to the liver, where it passes through a second capillary bed before being collected in the hepatic vein and passing to the sinus venosus. Paired anterior and posterior cardinal veins collect blood from the muscles and body wall. These veins lead, through a pair of common cardinal veins (duct of Cuvier), to the sinus venosus.

There is no single heart in the amphioxus, and blood is transported by contractions that arise independently in the sinus venosus, branchial hearts, subintestinal vein, and other vessels. The blood is nonpigmented and does not contain cells; oxygen transport is by simple solution in the blood. (B.E.Ma.)

## The vertebrate circulatory system

### THE BASIC VERTEBRATE PATTERN

**The plan.** All vertebrates have circulatory systems based on a common plan (Figure 3A), and so vertebrate systems show much less variety than do those of invertebrates. Although it is impossible to trace the evolution of the circulatory system by using fossils (because blood vessels do not fossilize as do bones and teeth), it is possible to theorize on its evolution by studying different groups of vertebrates and their developing embryos. Many of the variations from the common plan are related to the different requirements of living in water and on land.

*The heart.* The vertebrate heart lies below the alimentary canal in the front and centre of the chest, housed in its own section of the body cavity. During the development of an embryo, the heart first appears below the pharynx, and although it may also be in this position in adult animals, the heart often moves posteriorly as the animal grows and matures.

Development of the heart

The heart is basically a tube made of special muscle (cardiac muscle) that is not found anywhere else in the body. This cardiac muscle beats throughout life with its own automatic rhythm. Deoxygenated blood from the body is brought by veins into the most posterior part of the heart tube, the sinus venosus (Figure 3B). From there it passes forward into the atrium, the ventricle, and the conus arteriosus (called the bulbus cordis in embryos), and eventually to the arterial system. The blood is pushed through the heart because the various parts of the tube contract in sequence. As the heart develops from embryo to adult, each part of the tube becomes a chamber, separated from the others by valves, so that blood can neither flow backward in the system nor reenter the heart from the arteries. As the heart grows, it bends into an "S" shape, so that the sinus venosus and atrium lie above the ventricle and conus arteriosus.

*The blood vessels.* Gill slits are a fundamental feature of all vertebrate embryos, including humans. With few exceptions, there are six gill slits on each side. Blood leaving the heart travels from the conus arteriosus into the ventral aorta, which branches to send six pairs of arteries between the gill slits. The arterial branches join the dorsal aorta above the alimentary canal. Anterior to the gill slits, the ventral aorta branches again, forming two external carotid arteries that supply the ventral part of the head.

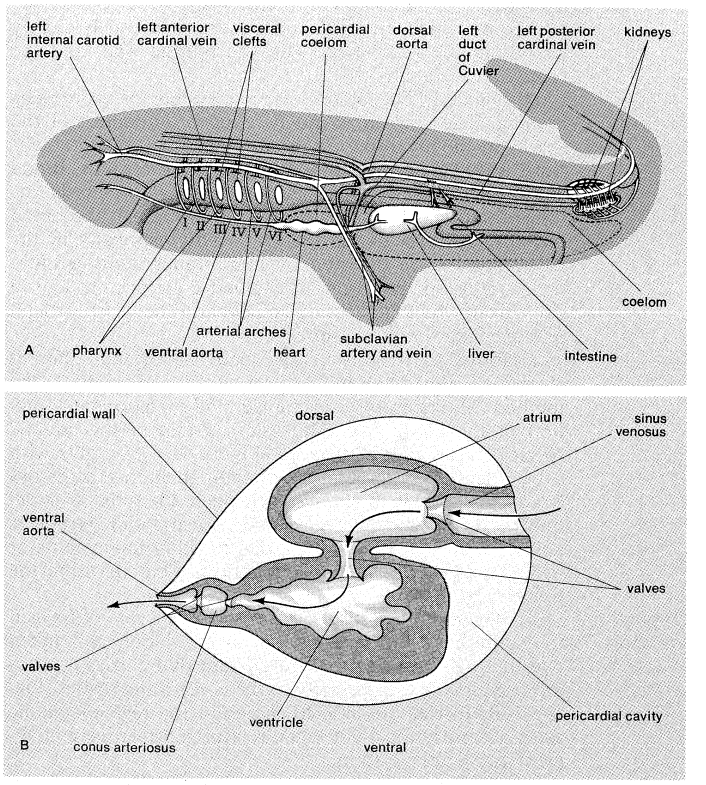



Figure 3: (A) Lateral view of a hypothetical primitive vertebrate showing arterial and venous systems in relation to each other and to the alimentary canal. (B) Left side longitudinal section of the basic form of vertebrate heart; arrow represents the direction of blood flow.

Two internal carotids, which are the anterior extensions of the dorsal aorta, supply the brain in the dorsal part of the head.

Deoxygenated blood collects in capillaries and then drains into larger and larger veins, which take it from various parts of the body to the heart. Of these, the anterior and posterior cardinal veins, each with left and right components, take blood to the heart from the front and rear of the body, respectively (Figure 3A). They lie dorsal to the alimentary canal, while the heart lies ventral to it. There is a common cardinal vein on each side, often called the duct of Cuvier, which carries blood ventrally into the sinus venosus. Various other veins join the cardinal veins from all over the body. The ventral jugular veins drain the lower part of the head and take blood directly into the common cardinal veins.

Lower vertebrates have two so-called portal systems, areas of the venous system that begin in capillaries in tissues and join to form veins, which divide to produce another capillary network en route to the heart. They are called the hepatic (liver) and renal (kidneys) portal systems. The hepatic system is important because it collects blood from the intestine and passes it to the liver, the centre for many chemical reactions concerned with the absorption of food into the body and the control of substances entering the general circulation. The function of the renal portal system is less clear, but it involves two veins that pass from the caudal vein to the kidneys, where they break up into capillaries.

Coronary circulation

The coronary circulation is that which supplies the heart muscle itself. It is of crucial importance, for the heart must never stop beating. Cardiac muscle needs oxygen from early in embryonic development until death. In mammals the coronary blood supply comes from the aorta, close to the heart. In evolutionary terms, this was not always so; many lower vertebrates, including agnathans and amphibians, have no specialized coronary arteries. The heart obtains its oxygen from blood passing through it. Fish have well-developed coronary vessels that arise from various sources, but ultimately from the efferent branchial system.

The introduction of lungs changed the site of oxygenation of the blood. In lungfishes coronary arteries arise from those anterior arterial arches receiving the most oxygenated blood from the heart. In reptiles coronary arteries branch from the systemic arch, but their position of origin varies. In some species they arise close to the heart, as in birds and mammals. Coronary veins generally run beside corresponding arteries but diverge from them to enter the main venous supply to the right atrium, or to the sinus venosus in fishes.

**Evolutionary trends.** Conventional classification divides vertebrates into two main groups—Gnathostomata, or vertebrates with jaws, and Agnatha, or those without jaws (the lampreys and hagfishes). This is a fundamental division, for agnathans also lack paired fins and scales. Agnathans are regarded as the most primitive group of vertebrates, not least because they appear first in the fossil record, before jawed fishes. Their circulatory systems differ in various ways from those of jawed vertebrates.

*Circulation in agnathans.* In the lamprey heart the atrium and ventricle are side by side, with the sinus venosus entering the atrium laterally. Nonmuscular valves prevent backflow of blood, and the conus arteriosus contains no cardiac muscle. There is no separate coronary blood supply, and the heart must obtain its oxygen from the blood as it goes through.

The arterial system in agnathans is most obviously modified because there are more than six sets of gills. Eight branches emerge from the ventral aorta, which splits into two, unlike the single vessel in most vertebrates with gill slits. Oxygenated blood from the gills is then collected into eight efferent vessels, which join to form a dorsal aorta, single for most of its length. Internal carotid arteries arise from the dorsal aorta, but the ventral part of the head is supplied from anterior efferent branchial (gill) vessels, not from the anterior part of the ventral aorta.

The venous system does not include a renal portal section, and there is asymmetry of the common cardinal veins, which take blood from the dorsal anterior and posterior cardinal veins down to the ventral heart. In embryos there are two of these, one on each side of the body; in lampreys, the left one disappears during development, while in hagfishes the right one disappears. Hagfishes also have accessory hearts in the venous system at several points. No other vertebrate has these structures.

It is not clear why the circulatory system of agnathans differs in these ways from the basic vertebrate plan. It is important to remember, however, that lampreys and hagfishes are specialized descendants of what was once a more diverse and widespread group of animals. Their circulatory systems may be less similar to the basic vertebrate plan than those of their ancestors because of their present mode of life.

*Circulation in jawed vertebrates.* Although clearly related to its mode of life, the blood system of a species also reflects its evolutionary history. The most significant change that occurred during early vertebrate evolution was the appearance of animals that could live and breathe on land. The first of these were the amphibians. Reptiles became even more independent of water because of their waterproof skins and shelled eggs, and from them evolved the most sophisticated land vertebrates, the mammals and birds. Obtaining oxygen entirely from air, instead of from water, involved drastic changes in the circulatory system.

Vertebrate evolution

Land vertebrates use their lungs to exchange carbon dioxide for oxygen from the air. Lungs may have evolved from a structure in fishes called the swim bladder, a sac that grows out from the anterior part of the gut. Fishes use it for buoyancy control, but it is possible that it was originally useful as an accessory for respiration. The problem is that lungs are found at a different site in the circulatory system from that of the gills, where oxygenation occurs in fishes. Instead of circulating around the body, as it does in fishes, oxygenated blood from the lungs returns to the heart. Therefore, in evolutionary terms, if mixing of oxygenated and deoxygenated blood was to be avoided in the heart, alterations to its structure had to occur. Land vertebrates developed lungs, a new vein (the pulmonary vein) to take blood from them to the heart, and a double circulation, whereby the heart is effectively divided into two halves—one-half concerned with pumping incoming deoxygenated blood from the body to the lungs and the

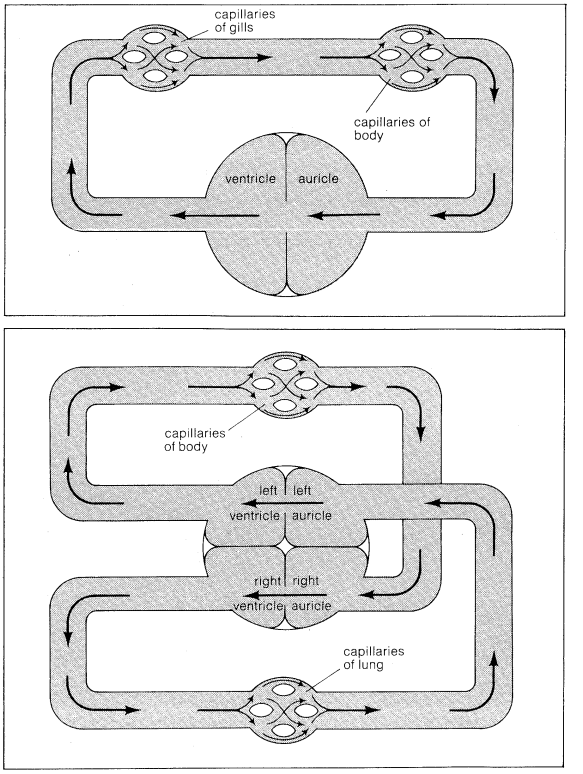



Figure 4: (Top) Single circulation of a fish; (bottom) double circulation of birds and mammals.

other with pumping oxygenated blood from the lungs around the body (Figure 4).

There are also modifications in the arterial and venous systems related to the appearance of lungs in the circulation. In the venous system, the paired posterior cardinal veins are replaced by a single posterior vena cava, and the renal portal system disappears. The main modifications to the basic arterial pattern involve what are the gill arteries of fishes. The anterior of these became responsible for carrying oxygenated blood to the head and to the brain; the intermediate arteries for carrying blood to the dorsal aorta, and so around the body; and the posterior arteries for carrying deoxygenated blood to the lungs.

Fishes

In fishes the four chambers of the heart are all well developed. Blood passes in sequence through the sinus venosus, atrium, ventricle, and conus arteriosus. The ventricle is the main pumping chamber, as it is in the hearts of all land vertebrates. During the evolution of the heart, the ventricle and atrium came to predominate; the sinus venosus became part of the atrium, while the conus arteriosus was incorporated into the ventricle. The atrium itself became a double structure—the two auricles—as did the ventricle, but the conversion of the ventricle into two chambers occurred later in evolution than the division of the atrium.

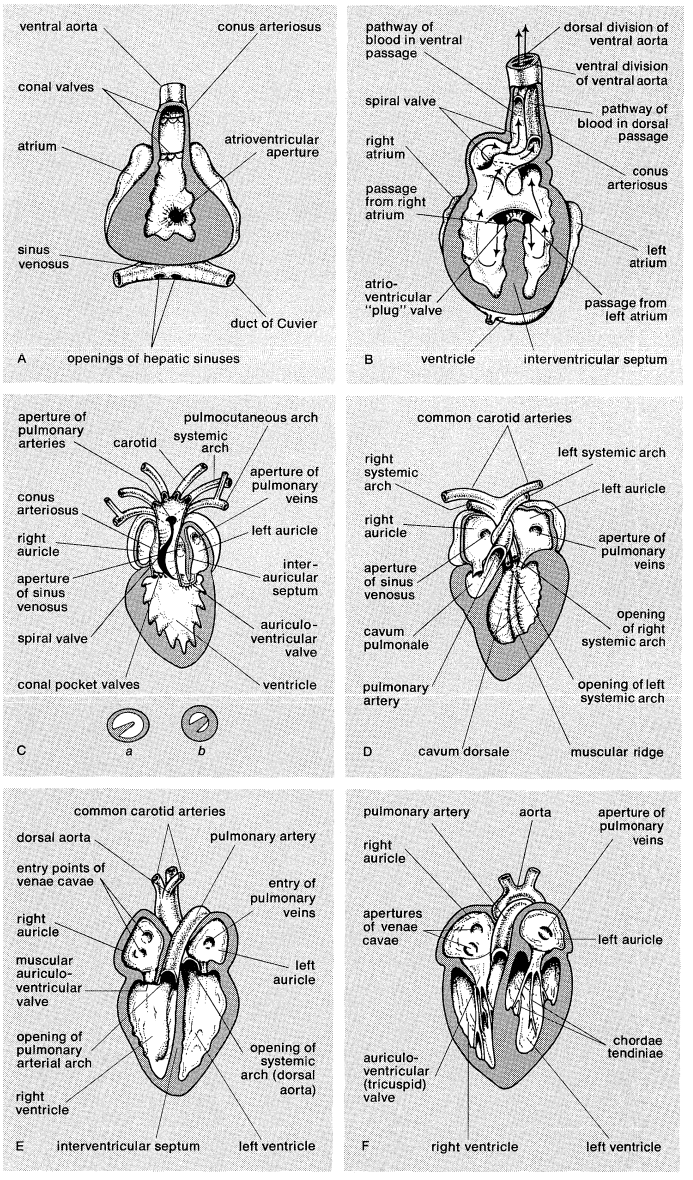



Figure 5: *Ventral dissections of vertebrate hearts.* (A) Dogfish (a shark); (B) lungfish; (C) frog (inset shows transverse sections of conus arteriosus in diastole [*a*] and systole [*b*]); (D) lizard; (E) bird; (F) mammal.

MODIFICATIONS AMONG THE VERTEBRATE CLASSES

**Fishes.** The hearts of fishes show little modification from the basic plan, except that lungfish hearts tend to become subdivided (Figure 5B). In them, the oxygenated blood carried by the pulmonary vein does not enter the sinus venosus along with the deoxygenated blood from the body. Instead, the oxygenated blood remains separate and enters the left side of the atrium. The atrium is partially divided into two auricles, and the ventricle also has a partial septum. Lungfishes show further signs of circulatory developments in their venous system (Figure 6). As in land vertebrates, there is a median posterior vena cava, and the posterior cardinal veins are reduced.

The arterial system of fishes is also altered from the basic plan. First there are the afferent (leading to) and efferent (leading from) parts of the gill (branchial) blood vessels. Each pair of blood vessels looping up between a pair of gills is called an arterial arch. During the development of embryos, the arterial arches become interrupted by capillaries in the gills. Thus, each arch consists of a ventral afferent section that brings blood to the gills from the heart and a dorsal efferent section that collects blood from the gill capillaries and carries it to the dorsal aorta. The whole circulatory system is a one-way arrangement, with the heart pumping only deoxygenated blood from the body forward to the gills to be oxygenated and redistributed to the body.

Oxygenation in fishes

Although six gill slits appear in embryos, few adult fishes retain all six. The first and most anterior gill slit in the series becomes the spiracle, and the first branchial arch is much modified; parts of it disappear altogether (Figure 6). The second branchial arch is variable in its presence in different fishes. In general, therefore, adult fishes often have only four of the six original arterial arches found in embryos. The external carotid arteries also show modifications. Instead of arising from the anterior part of the ventral aorta, they become connected with the efferent portion of the second branchial arch. This change ensures that, despite modifications to the most anterior of the arterial arches, blood just oxygenated in the gills will reach the head.

It may be that the prevalence of poorly oxygenated water in certain habitats explains the evolution of lungs and, hence, of land vertebrates. Fishes also have evolved accessory structures for obtaining oxygen from the air. These are often modified gill chambers, with dense capillary networks. Even the intestine may be involved, as in the loach *Haplosternum.*

The swim bladder

Except for sharks and their relatives (elasmobranchs), most fishes have a swim bladder, the structure from which lungs may have evolved. Although its prime function in fishes is to control buoyancy, the swim bladder may also act as an oxygen reserve, for the gas in it often contains a high concentration of oxygen derived from the blood's own supply. Blood to the swim bladder usually comes from the dorsal aorta. One African fish, *Polypterus,* uses its swim bladder for respiration, and the veins from it join the posterior cardinal veins close to the heart. These swim bladder veins are almost where pulmonary veins would be expected to be, if they were bringing oxygenated blood from lungs straight to the heart.

The lungfishes have gone further in adapting their circulatory systems to the presence of lungs, although the different species do not breathe air to the same extent. Some of their modifications foreshadow the changes that have taken place in amphibians. The divided atrium of the lungfish heart (Figure 5B) receives blood from the body on the right side and from the lungs on the left. The conus is large and is divided by a complex system of valves arranged in a spiral pattern and called the spiral valve. The ventral aorta is also subdivided internally. The result is that oxygenated blood from the left side of the ventricle is directed into the ventral division of the ventral aorta and passes to the anterior of the arterial arches, while deoxygenated blood from the right side of the ventricle is directed into the two most posterior arterial arches and passes mainly to the lungs.

Four arterial arches are present even in the lungfish species most dependent on breathing air (*Lepidosiren;*

Figure 6), where gills still exist. These are arches three to six of the original series of six present in fish embryos. Their arrangement is largely unaffected by the presence of lungs, except that the gills may be reduced and the arteries may pass straight through without intervening capillaries. Arches five and six, however, join together before entering the dorsal aorta and give rise to a large pulmonary artery to the lungs. Thus, in lungfishes, lungs and gills can be seen working side by side.

The circulatory systems of lungfishes are strikingly similar to those of amphibians, and although lungfishes do not seem to have been amphibian ancestors, they are related to fishes that were. It is likely that several groups of ancient fishes had lungs, partially divided hearts, and ventral aortas, and from one of these groups arose the land vertebrates.

**Amphibians.** Modern amphibians are characterized by the flexibility of their gaseous exchange mechanisms. Amphibian skin is moistened by mucous secretions and is well supplied with blood vessels. It is used for respiration to varying degrees. When lungs are present, carbon dioxide may pass out of the body across the skin, but in some salamanders there are no lungs and all respiratory exchanges occur via the skin. Even in such animals as frogs, it seems that oxygen can be taken up at times by the skin, under water for example. Therefore, regulation of respiration occurs within a single species, and the relative contribution of skin and lungs varies during the life of the animal.

The amphibian heart

The amphibian heart is generally of a tripartite structure, with a divided atrium but a single ventricle (Figure 5C). The lungless salamanders, however, have no atrial septum, and one small and unfamiliar group, the caecilians, has signs of a septum in the ventricle. It is not known whether the original amphibians had septa in both atrium and ventricle. They may have, and the absence of septa in many modern forms may simply be a sign of a flexible approach to the use of skin or lung, or both, as the site of oxygen exchange. In addition, the ventricle is subdivided by muscular columns into many compartments that tend to prevent the free mixing of blood.

The conus arteriosus is muscular and contains a spiral valve. Again, as in lungfishes, this has an important role in directing blood into the correct arterial arches. In the frog, *Rana,* venous blood is driven into the right atrium of the heart by contraction of the sinus venosus, and it flows into the left atrium from the lungs. A wave of contraction then spreads over the whole atrium and drives blood into the ventricle, where blood from the two sources tends to remain separate. Separation is maintained in the spiral valve, and the result is similar to the situation in lungfishes. Blood from the body, entering the right atrium,

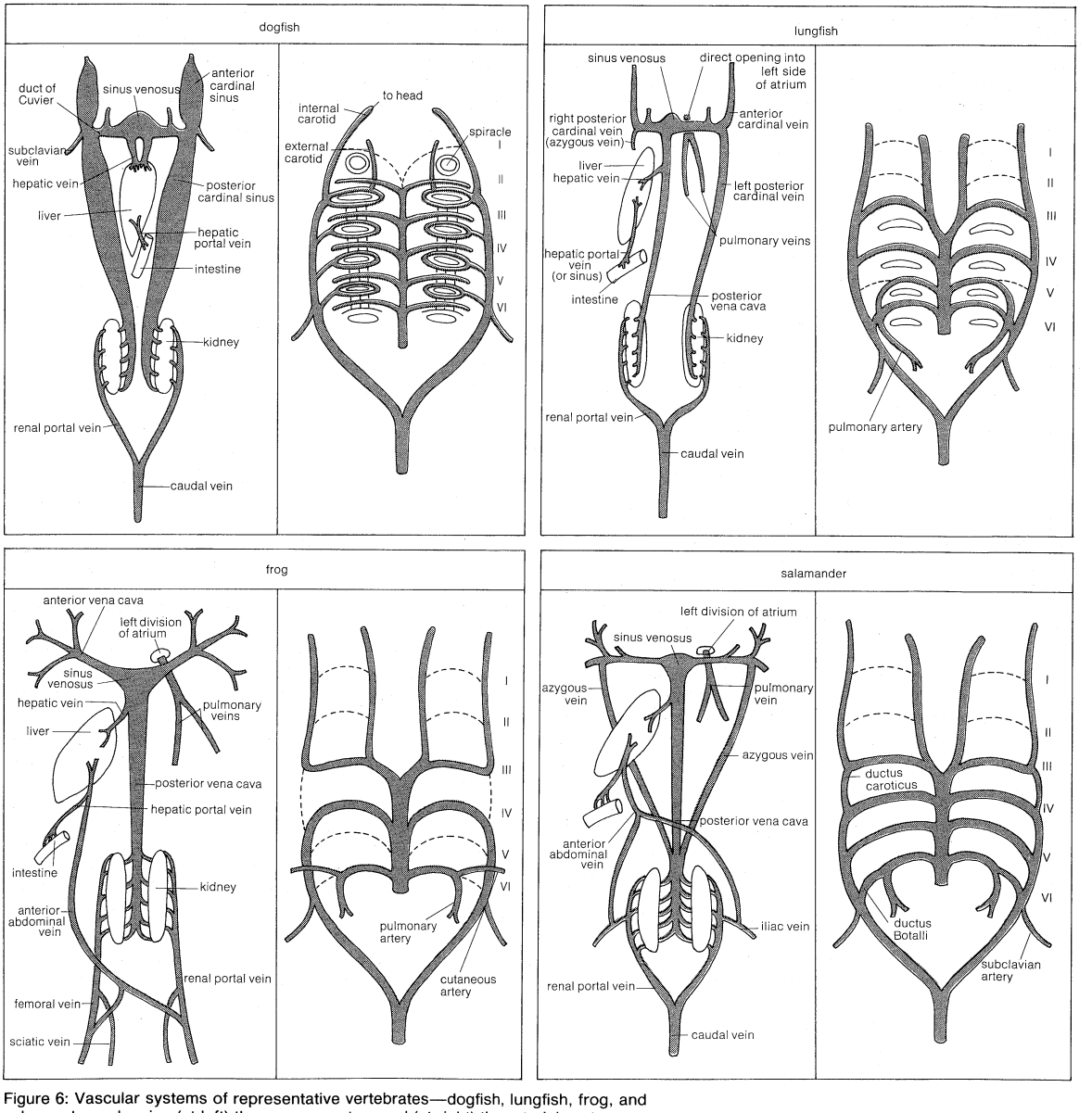



Figure 6: Vascular systems of representative vertebrates—dogfish, lungfish, frog, and salamander—showing (at left) the venous system and (at right) the arterial system as seen in ventral views.

tends to pass to the lungs and skin for oxygenation; that from the lungs, entering the left atrium, tends to go to the head. Some mixing does occur, and this blood tends to be directed by the spiral valve into the arterial arch leading to the body.

Blood returning from the skin does not enter the circulation at the same point as blood from the lungs. Thus, oxygenated blood arrives at the heart from two different directions—from the sinus venosus, to which the cutaneous (skin) vein connects, and from the pulmonary vein. Both right and left atria receive oxygenated blood, which must be directed primarily to the carotid arteries supplying the head and brain. It is likely that variable shunting of blood in the ventricle is important in ensuring this. A ventricular septum would inhibit shunting; it is at least possible that its absence in amphibians is not a primitive feature but a secondary adaptation to variable gas-exchange mechanisms.

The amphibian venous system shows various features that are characteristic of land vertebrates. The posterior cardinal veins are replaced by a posterior vena cava, but they are still visible in salamanders. There is a renal portal system, and an alternative route back to the heart from the legs is provided by an anterior abdominal vein that enters the hepatic portal vein to the liver (Figure 6).

Amphibian larvae and the adults of some species have gills. There are four arterial arches in salamanders (urodeles) and three in frogs (anurans; Figure 6). These are three through six of the original series, the fifth disappearing in adult frogs. There is no ventral aorta, and the arterial arches arise directly from the conus—an important feature, given that the conus and its spiral valve control the composition of blood reaching each arterial arch. The names given to the three arterial arches of frogs are those used in all land vertebrates, including mammals. They are the carotid (the third), systemic (the fourth), and pulmonary (the sixth) arches. Blood to the lungs (and skin in frogs) is always carried by the sixth arterial arch, which loses its connection to the dorsal aorta. All land vertebrates supply their lungs with deoxygenated blood from this source.

Lungs

**Reptiles.** Unlike lungfishes and amphibians, reptiles depend entirely on their lungs for respiration. Gills and skin do not provide additional sources of oxygen. Only the crocodiles, however, truly approach birds and mammals in their almost complete "double" circulation. Because of the development of a neck and relative elongation of that region of the body, the heart may be displaced posteriorly and the arrangement of arteries and veins may be altered accordingly. In general, however, the circulatory system resembles that in frogs (Figures 6 and 7).

The reptilian heart

Various changes can be seen in the reptilian heart. The left atrium is smaller than the right and always completely separate from it (Figure 5D). The sinus venosus is present but small. The ventricle is variously subdivided in different groups. Three arterial trunks arise directly from the ventricle, the conus having been partly incorporated into it. The three trunks are the right and left systemic arches and the pulmonary arch. The carotid arch is a branch of the right systemic arch. When the ventricle is actually beating, there is functional separation of blood from the two atria so that most oxygenated blood flows to the carotid arteries and hardly mixes with deoxygenated blood going to the lungs.

Crocodiles are the only living representatives of the archosaurian reptiles, the group that included the dinosaurs and from which birds evolved. Crocodiles have a complete ventricular septum, producing two equally sized chambers. The blood from the right and left atria is not mixed; despite this, there is an opening at the base of the right and left systemic arches, and blood can be shunted between the two. This is important during diving, when blood flow to the lungs is decreased. The crocodile heart is situated so posteriorly that the subclavian artery, which would normally arise from the dorsal aorta at the level of the systemic arch, arises from the carotid artery.

**Birds.** Bird circulatory systems have many similarities to those of reptiles, from which they evolved. The changes that have occurred are more of degree than of kind. The heart is completely divided into right and left sides (Figure 5E). The sinus venosus is incorporated into the right auricle and becomes the sinoauricular node. It is from this point that the heartbeat is initiated. There is no conus, and only two vessels leave the divided ventricle. These are the pulmonary artery from the right side and the systemic arch from the left. The systemic arch is asymmetrical—the main difference in this area between birds and lizards. Only the right part of the systemic arch is present, the left being suppressed (Figure 7). The arterial arches are no longer bilaterally symmetrical. Another difference between birds and lizards is found in the venous circulation: the renal portal system is reduced in birds.

**Mammals.** Mammals also evolved from reptiles, but not from the same group as did birds, and must have developed their double circulation independently from early reptiles. Nevertheless, several parallel changes occurred, such as the common incorporation of the sinus venosus into the right auricle. The most striking manifestation of different origins is seen in the mammalian aorta, which leaves the left ventricle and curves to the left (Figure 5F). The aorta corresponds to the left half of the systemic arch, while the right is missing (Figure 7). The carotid arteries arise from the left systemic arch (aorta), though their precise position varies among mammals. The arterial system is asymmetric, as in birds, but in the opposite way.

The heart of both mammals and birds is a double pump, powering two systems of vessels with different characteristics. The left ventricle has a thicker layer of muscle around it, a necessary adaptation for powering its beat against the high resistance of the extensive systemic circulation throughout the body. The right ventricle has a thinner wall, consistent with its role in pumping blood to the lungs against a much lower resistance.

## EMBRYONIC DEVELOPMENT OF THE CIRCULATORY SYSTEM

An embryo develops only with an adequate supply of oxygen and metabolites. In its early stages these may be provided by diffusion. Because the rate of diffusion becomes limiting beyond a certain size, however, the circulatory system becomes functional early in development, often before other organs and systems are obvious.

The heart develops from the middle embryonic tissue layer, the mesoderm, just below the anterior part of the gut. It begins as a tube that joins with blood vessels also forming in the mesoderm. Other mesodermal cells form a coat around the heart tube and become the muscular wall, or myocardium. The heart lies in its own section of body cavity, called the pericardial coelom, formed by partitions that cut it off from the main body cavity. From an original tube shape, the heart bends back on itself as it grows within the pericardial cavity. The sinus venosus and atrium lie above the ventricle and bulbus cordis (embryonic equivalent of the conus arteriosus). Septa gradually partition the heart into chambers.

In mammalian and bird embryos, the lungs are not used until birth. Oxygen is obtained in the former from the placenta and in the latter from embryonic membranes close to the porous eggshell.

The circulation has various modifications for diverting oxygenated blood from sources outside the embryo to the body of the embryo. In mammals blood from the placenta travels to the right auricle via the umbilical vein and posterior vena cava. It passes through an opening, the foramen ovale, into the left auricle, and then to the left ventricle and around the body. Deoxygenated blood entering the anterior vena cava fills the right ventricle; however, instead of passing to the lungs, it is shunted through the ductus arteriosus, between the pulmonary and systemic arches, and into the dorsal aorta. From the dorsal aorta the deoxygenated blood travels to the placenta, bypassing the lungs completely. At birth the foramen ovale closes, as does the ductus arteriosus, and the lungs become functional.

The development of the circulatory system in higher vertebrate embryos (*i.e.*, those of birds and mammals) generally follows a sequence of seven main events. Initially, a tubular heart bends into an "S" shape. Blood then flows from behind forward through the sinus venosus, atrium, ventricle, and bulbus cordis. There is then subdivision of the atrium and ventricle and of the opening between

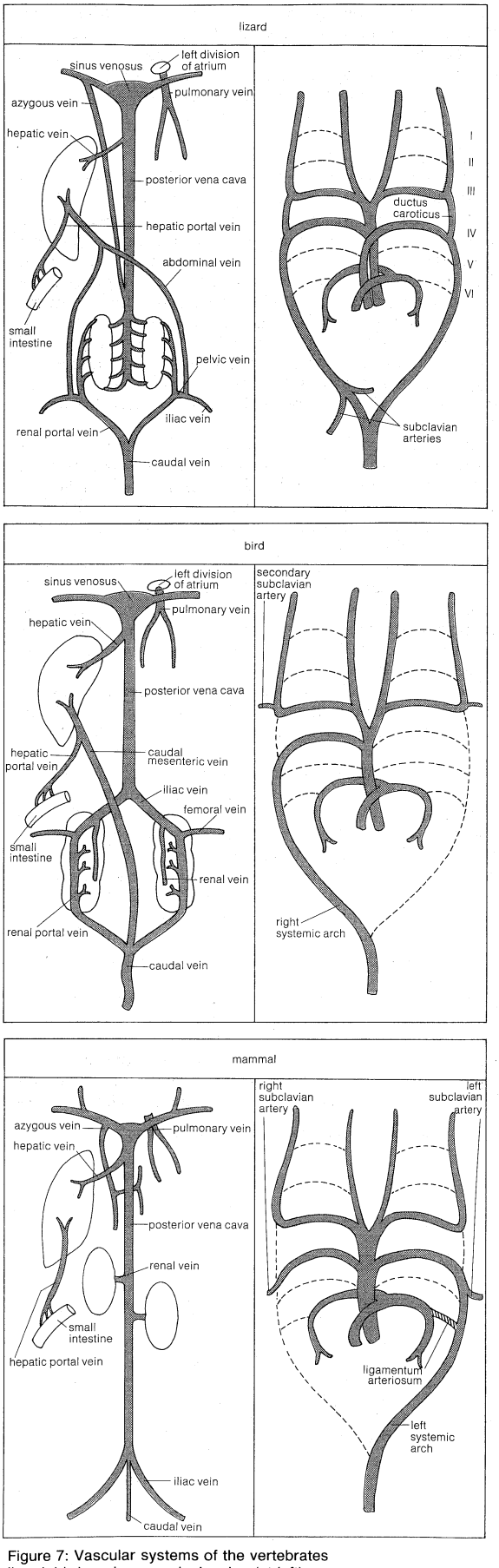



Figure 7: Vascular systems of the vertebrates lizard, bird, and mammal, showing (at left) the venous system and (at right) the arterial system as seen in ventral views.

them. The sinus venosus is incorporated into the right atrium. The pulmonary veins are segregated to open into the left atrium. The bulbus cordis is subdivided into a pulmonary trunk from the right ventricle and a systemic trunk from the left ventricle. Finally, an embryonic set of six arterial arches is reduced to three in adults, and their relationships are further complicated by asymmetrical loss of some parts and development of others.

## BIODYNAMICS OF VERTEBRATE CIRCULATION

**Blood pressure and blood flow.** The pressure that develops within the closed vertebrate circulatory system is highest at the pump—the heart—and decreases with distance away from the pump because of friction within the blood vessels. Because the blood vessels can change their diameter, blood pressure can be affected by both the action of the heart and changes in the size of the peripheral blood vessels. Blood is a living fluid—it is viscous and contains cells (45 percent of its volume in human beings)—and yet the effects of the cells on its flow patterns are small.

Blood enters the atrium by positive pressure from the venous system or by negative pressure drawing it in by suction. Both mechanisms operate in vertebrates. Muscular movements of the limbs and body, and gravity in land vertebrates, are forces propelling blood to the heart. In fishes and amphibians the atrium forces blood into the ventricle when it contracts. In birds and mammals the blood arrives at the heart with considerable residual pressure and passes through the auricles into the ventricles, apparently without much additional impetus from contraction of the auricles.

Two pressure levels

The ventricle is the main pumping chamber, but one of the features of double circulation is that the two circuits require different pressure levels. Although the shorter pulmonary circulation requires less pressure than the much longer systemic circuit, the two are connected to each other and must transport the same volume of fluid per unit time. The right and left ventricles in birds and mammals function as a volume and a pressure pump, respectively. The thick muscular wall of the left ventricle ensures that it develops a higher pressure during contraction in order to force blood through the body. It follows that pressures in the aorta and pulmonary artery may be very different. In human beings aortic pressure is about six times higher.

Valves throughout the system are crucial to maintain pressure. They prevent backflow at all levels; for example, they prevent flow from the arteries back into the heart as ventricular pressure drops at the end of a contraction cycle. Valves are important in veins, where the pressure is lower than in arteries.

Another impetus to blood flow is contraction of the muscles in the walls of vessels. This also prevents backflow of arterial blood toward the heart at the end of each contraction cycle. Input from nerves, sensory receptors in the vessels themselves, and hormones all influence blood vessel diameter, but responses differ according to position in the body and animal species.

Normally, the pressures that develop in a circulatory system vary widely in different animals. Body size can be an important factor. The closed circulation systems of vertebrates generally operate at higher pressures than the open blood systems of invertebrates; the systems of birds and mammals operate at the highest pressures of all.

**Electrical activity.** The vertebrate heart is myogenic (rhythmic contractions are an intrinsic property of the cardiac muscle cells themselves). Pulse rate varies widely in different vertebrates, but it is generally higher in small animals, at least in birds and mammals. Each chamber of the heart has its own contraction rate. In the frog, for example, the sinus venosus contracts fastest and is the pacemaker for the other chambers, which contract in sequence and at a decreasing rate, the conus being the slowest. In birds and mammals, where the sinus venosus is incorporated into the right atrium at the sinoauricular node, the latter is still the pacemaker and the heartbeat is initiated at that point. Thus, the evolutionary history of the heart explains the asymmetrical pattern of the heartbeat.

In the frog each contraction of the heart begins with a localized negative charge that spreads over the surface of

the sinus venosus. In lower vertebrates, the cardiac muscle cells themselves conduct the wave of excitation. In birds and mammals, however, special conducting fibres (arising from modified muscle cells) transmit the wave of excitation from the sinoauricular node to the septum between the auricles, and then, after a slight delay, down between and around the ventricles. The electrical activity of the heart can be recorded; the resulting pattern is called an electrocardiogram.

Factors

**Control of heartbeat and circulation.** Many factors, such as temperature, oxygen supply, or nervous excitement, affect heartbeat and circulation. Blood circulation is controlled mainly via nerve connections, sensory receptors, and hormones. These act primarily by varying the heart's pulse rate, amplitude, or stroke volume and by altering the degree of dilation or constriction of the peripheral blood vessels (*i.e.,* those blood vessels near the surface of the body).

Temperature has a direct effect on heart rate, and one of the ways in which mammals regulate their internal temperature is by controlling peripheral blood circulation. Mammals are endothermic (warm-blooded) vertebrates; their internal temperature is kept within narrow limits by using heat generated by the body's own metabolic processes. Lizards are ectothermic (cold-blooded); they obtain heat from the external environment by, for example, basking in the sun. The effects of oxygen concentration on the heart and blood vessels is rapid. Oxygen deficiency in the cardiac tissue causes dilation of the coronary capillaries, thereby increasing blood flow and oxygen supply.

Most effects on the circulation are indirect and complex. All vertebrate hearts receive input from nerves; for example, stimulation of a branch of the vagus nerve causes the release of acetylcholine at the nerve endings, which depresses the heart rate. Other nerve endings release norepinephrine, which increases the heart rate. Less directly, nervous stimulation brought about by stress causes the release of the hormones epinephrine and norepinephrine into the bloodstream. These substances not only make the heart beat faster and with a greater amplitude, but they also divert blood to the muscles by constricting the vessels in the skin and gut. This prepares the animal physiologically for physical exertion. Numerous other chemicals, such as nicotine, affect heart rate directly or indirectly.

Two other factors are important in the context of circulatory regulation—the concentrations of inorganic ions and sensory receptors in blood vessel walls. Sodium, potassium, and calcium ions are always involved in changes of electrical potential across cell membranes. A change in their concentrations, therefore, influences heartbeat profoundly. External calcium concentration can, for example, determine the conductance of sodium across the cardiac cell membranes. Sensory receptors in the walls of blood vessels register blood pressure. They are found in the aorta, carotid arteries, pulmonary artery, capillaries in the adrenal gland, and the tissues of the heart itself. Impulses from the receptors travel to the medulla of the brain, from where messages are sent via motor nerves to the heart and blood vessels. Regulation is thus achieved according to the body's needs. (M.E.Ro.)

## The human cardiovascular system

The human cardiovascular system is a closed tubular system in which blood, propelled by a muscular heart, flows through vessels to and from all parts of the body. Two circuits, the pulmonary and the systemic, consist of arterial, capillary, and venous components.

The primary function of the heart is to serve as a muscular pump propelling blood into and through vessels to and from all parts of the body. The arteries, which receive this blood at high pressure and velocity and conduct it throughout the body, are thickly walled with elastic fibrous tissue and muscle cells. The arterial tree—the branching system of arteries—terminates in short, narrow, muscular vessels called arterioles, from which blood enters simple endothelial tubes (*i.e.,* tubes formed of endothelial, or lining, cells) known as capillaries. These microscopically thin capillaries are permeable to vital cellular nutrients and waste products and distribute and receive nutrients and wastes. From the capillaries, the blood, now depleted of oxygen and burdened with waste products, moving more slowly and under low pressure, enters small vessels called venules, which converge to form veins, ultimately guiding the blood on its way back to the heart.

### THE HEART

**Description.** *Shape and location.* The human adult heart is normally slightly larger than a clenched fist with average dimensions of about 13 × 9 × 6 centimetres (5 × 3½ × 2½ inches) and weighing approximately 10.5 ounces (300 grams). It is cone-shaped in appearance, with the broad base directed upward and to the right and the apex pointing downward and to the left. It is located in the chest (thoracic) cavity behind the breastbone, or sternum in front of the windpipe, or trachea, the esophagus, and the descending aorta, between the lungs, and above the diaphragm (the muscular partition between the chest and abdominal cavities). About two-thirds of the heart lies to the left of the midline.

*Pericardium.* The heart is suspended in its own membranous sac, the pericardium. The strong outer portion of the sac, or fibrous pericardium, is firmly attached to the diaphragm below, the mediastinal pleura on the side, and the sternum in front, and gradually blends with the coverings of the superior vena cava and the pulmonary (lung) arteries and veins leading to and from the heart. (The space between the two lungs, the mediastinum, is bordered by the mediastinal pleura, a continuation of the membrane lining the chest. The superior vena cava is the principal channel for venous blood from the chest, the arms, the neck, and the head.)

Fibrous pericardium

Smooth, serous (moisture-exuding) membrane lines the fibrous pericardium, then bends back and covers the heart (Figure 8). The portion of membrane lining the fibrous pericardium is known as the parietal serous layer, or parietal pericardium, that covering the heart as the visceral serous layer, visceral pericardium, or epicardium.

The two layers of serous membrane are normally separated only by 10 to 15 millilitres (0.6 to 0.9 cubic inch) of pericardial fluid, which is secreted by the serous membranes. The slight space created by the separation is called the pericardial cavity. The pericardial fluid lubricates the two membranes with every beat of the heart as their surfaces glide over each other. Fluid is filtered into the pericardial space through both the visceral and parietal pericardia.

*Chambers of the heart.* The heart is divided by septa, or partitions, into right and left halves, and each half is subdivided into two chambers. The upper chambers, the atria, are separated by a partition known as the interatrial septum; the lower chambers, the ventricles, are separated by the interventricular septum (Figure 9). The atria re-

From S. Jacob and C. Francone, *Structure and Function in Man* (1970); W.B. Saunders Co.

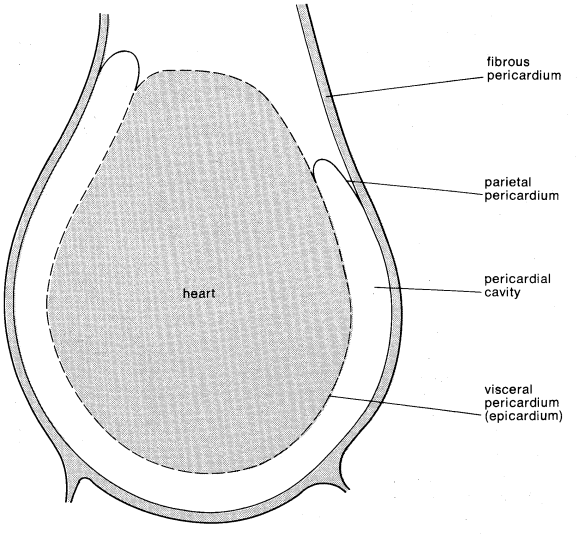



Figure 8: The pericardium.

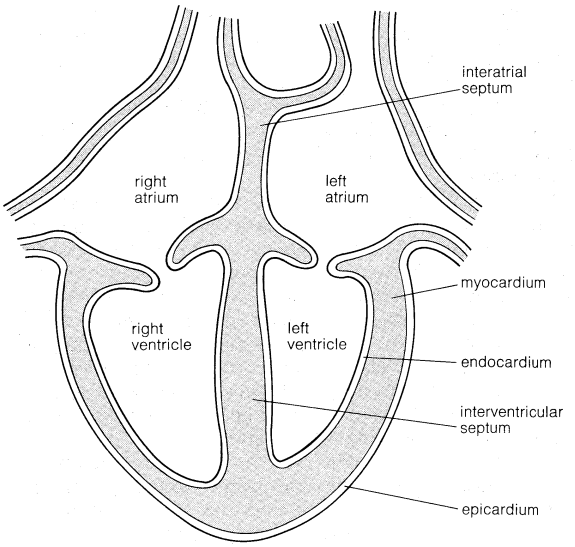



Figure 9: The walls and chambers of the heart.
From S. Jacob and C. Francone, *Structure and Function in Man* (1970); W.B. Saunders Co.

ceive blood from various parts of the body and pass it into the ventricles. The ventricles, in turn, pump blood to the lungs and to the remainder of the body.

The right atrium, or right superior portion of the heart, is a thin-walled chamber receiving blood from all tissues except the lungs. Three veins empty into the right atrium, the superior and inferior venae cavae, bringing blood from the upper and lower portions of the body, respectively, and the coronary sinus, draining blood from the heart itself. Blood flows from the right atrium to the right ventricle. The right ventricle, the right inferior portion of the heart, is the chamber from which the pulmonary artery carries blood to the lungs.

The left atrium, the left superior portion of the heart, is slightly smaller than the right atrium and has a thicker wall. The left atrium receives the four pulmonary veins, which bring oxygenated blood from the lungs. Blood flows from the left atrium into the left ventricle. The left ventricle, the left inferior portion of the heart, has walls three times as thick as those of the right ventricle. Blood is forced from this chamber through the aorta to all parts of the body except the lungs.

Interventricular sulci

*External surface of the heart.* Shallow grooves called the interventricular sulci, containing blood vessels, mark the separation between ventricles on the front and back surfaces of the heart. There are two grooves on the external surface of the heart. One, the atrioventricular groove, is along the line where the right atrium and the right ventricle meet; it contains a branch of the right coronary artery (the coronary arteries deliver blood to the heart muscle). The other, the anterior interventricular sulcus, runs along the line between the right and left ventricles and contains a branch of the left coronary artery.

On the posterior side of the heart surface, a groove called the posterior longitudinal sulcus marks the division between the right and left ventricles; it contains another branch of a coronary artery. A fourth groove, between the left atrium and ventricle, holds the coronary sinus, a channel for venous blood.

*Origin and development.* In the embryo, formation of the heart begins in the pharyngeal, or throat, region. The first visible indication of the embryonic heart occurs in the undifferentiated mesoderm, the middle of the three primary layers in the embryo, as a thickening of invading cells. An endocardial (lining) tube of flattened cells subsequently forms and continues to differentiate until a young tube with forked anterior and posterior ends arises.

Primary chambers in embryo

As differentiation and growth progress, this primitive tube begins to fold upon itself, and constrictions along its length produce four primary chambers. These are called, from posterior to anterior, the sinus venosus, atrium, ventricle, and truncus arteriosus (Figure 10). The characteristic bending of the tube causes the ventricle to swing first to the right and then behind the atrium, the truncus coming to lie between the sideways dilations of the atrium. It is during this stage of development and growth that the first pulsations of heart activity begin.

Endocardial cushions (local thickenings of the endocardium, or heart lining) "pinch" the single opening between the atrium and the ventricle into two portions, thereby forming two openings. These cushions are also responsible for the formation of the two atrioventricular valves (the valves between atria and ventricles), which regulate the direction of blood flow through the heart.

The atrium becomes separated into right and left halves first by a primary partition with a perforation and later by a secondary partition, which, too, has a large opening, called the foramen ovale, in its lower part. Even though the two openings do not quite coincide in position, blood still passes through, from the right atrium to the left. At birth, increased blood pressure in the left atrium forces the primary partition against the secondary one, so that the two openings are blocked and the atria are completely separated. The two partitions eventually fuse.

The ventricle becomes partially divided into two chambers by an indentation of myocardium (heart muscle) at its tip. This developing partition is largely muscular and is supplemented by membranous connective tissue that develops in conjunction with the subdivision of the truncus arteriosus by a spiral partition into two channels, one for systemic and one for pulmonary circulation (the aorta and the pulmonary artery, respectively). The greater portion of blood passing through the right side of the heart in the fetus is returned to the systemic circulation by the ductus arteriosus, a vessel connecting the pulmonary artery and the aorta. At birth this duct becomes closed by a violent contraction of its muscular wall. Thereafter the blood in the right side of the heart is driven through the pulmonary arteries to the lungs for oxygenation and returned to the left side of the heart for ejection into the systemic circulation. A distinct median furrow at the apex of the ventricles marks the external subdivision of the ventricle into right and left chambers.

**Structure and function.** *Valves of the heart.* To prevent backflow of blood, the heart is equipped with valves that permit the blood to flow in only one direction (Figure 11). There are two types of valves located in the heart: the atrioventricular valves (tricuspid and mitral) and the semilunar valves (pulmonary and aortic).

The atrioventricular valves are thin, leaflike structures located between the atria and the ventricles. The right atrioventricular opening is guarded by the tricuspid valve,

From S. Jacob and C. Francone, *Structure and Function in Man* (1970); W.B. Saunders Co.

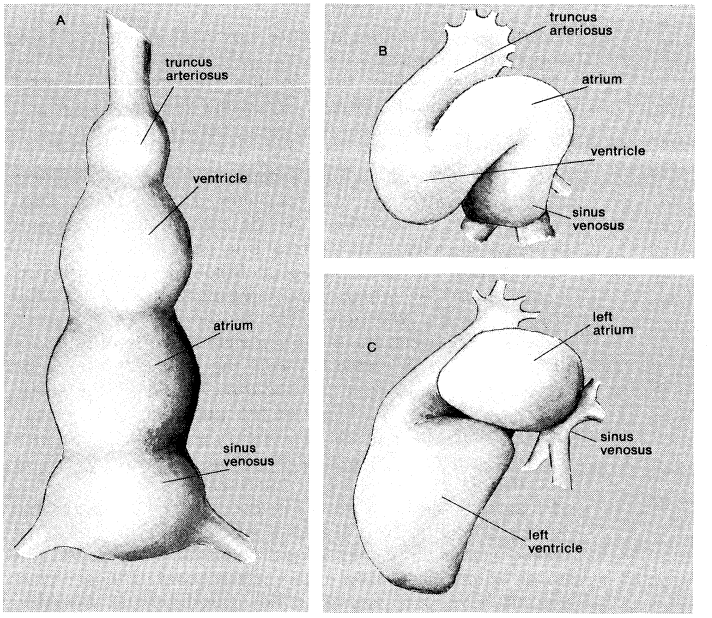



Figure 10: *Stages in embryonic development of the heart.* (A) Primitive heart tube. (B) Beginnings of flexion. (C) Complete flexion with separation of chambers.

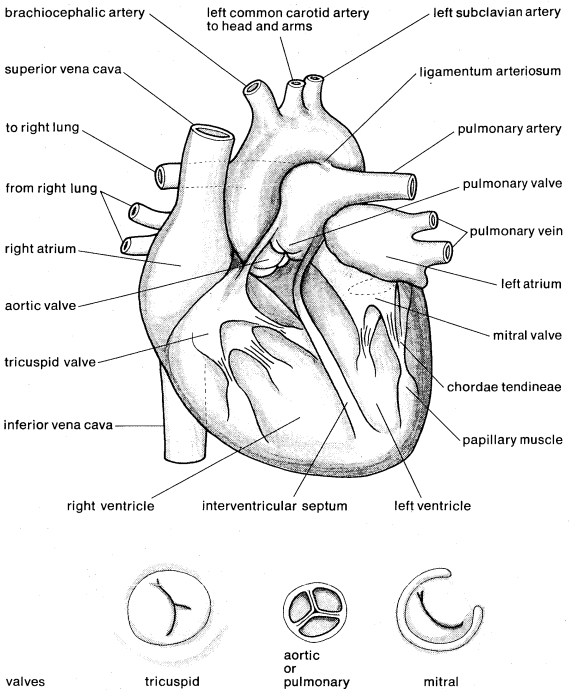



Figure 11: Schematic "transparent" drawing of the heart showing the relations of the various heart valves.
From S. Jacob and C. Francone, *Structure and Function in Man* (1970); W.B. Saunders Co.

so called because it consists of three irregularly shaped cusps, or flaps. The leaflets consist essentially of folds of endocardium (the membrane lining the heart) reinforced with a flat sheet of dense connective tissue. At the base of the leaflets, the middle supporting flat plate becomes continuous with that of the dense connective tissue of the ridge surrounding the openings.

Chordae tendineae and papillary muscles

Tendinous cords of dense tissue (chordae tendineae) covered by thin endocardium extend from the nipplelike papillary muscles to connect with the ventricular surface of the middle supporting layer of each leaflet. The chordae tendineae and the papillary muscles from which they arise limit the extent to which the portions of the valves near their free margin can billow toward the atria. The left atrioventricular opening is guarded by the mitral, or bicuspid, valve, so named because it consists of two flaps. The mitral valve is attached in the same manner as the tricuspid, but it is stronger and thicker because the left ventricle is by nature a more powerful pump working under high pressure.

Blood is propelled through the tricuspid and mitral valves as the atria contract. When the ventricles contract, blood is forced backward, passing between the flaps and walls of the ventricles. The flaps are thus pushed upward until they meet and unite, forming a complete partition between the atria and the ventricles. The expanded flaps of the valves are restrained by the chordae tendineae and papillary muscles from opening into the atria.

The semilunar valves are pocketlike structures attached at the point at which the pulmonary artery and the aorta leave the ventricles. The pulmonary valve guards the orifice between the right ventricle and the pulmonary artery. The aortic valve protects the orifice between the left ventricle and the aorta. The three leaflets of the aortic semilunar and two leaflets of the pulmonary valves are thinner than those of the atrioventricular valves, but they are of the same general construction with the exception that they possess no chordae tendineae.

Closure of the heart valves is associated with an audible sound. The first sound occurs when the mitral and tricuspid valves close, the second when the pulmonary and aortic semilunar valves close. These characteristic heart sounds have been found to be caused by the vibration of the walls of the heart and major vessels around the heart. The first heart sound, or "lubb," is heard when the ventricles contract, causing a sudden backflow of blood that closes the valves and causes them to bulge back. The elasticity of the valves then causes the blood to bounce backward into each respective ventricle. This effect sets the walls of the ventricles into vibration, and the vibrations travel away from the valves. When the vibrations reach the chest wall where the wall is in contact with the heart, sound waves are created that can be heard with the aid of a stethoscope.

The second heart sound results from vibrations set up in the walls of the pulmonary artery, the aorta, and, to a lesser extent, the ventricles as the blood reverberates back and forth between the walls of the arteries and the valves after the pulmonary and aortic semilunar valves suddenly close. These vibrations are then heard as the "dupp" sound as the chest wall transforms the vibrations into sound waves. The first heart sound is followed after a short pause by the second. A pause about twice as long comes between the second sound and the beginning of the next cycle. The opening of the valves is silent.

Three layers of the heart wall

*Wall of the heart.* The wall of the heart consists of three distinct layers—the epicardium (outer layer), the myocardium (middle layer), and the endocardium (inner layer). Coronary vessels supplying arterial blood to the heart penetrate the epicardium before entering the myocardium. This outer layer, or visceral pericardium, consists of a surface of flattened epithelial (covering) cells resting upon connective tissue.

The myocardium consists of interlacing bundles of cardiac muscle fibres (Figure 12) possessing the appearance of striated muscle (striped skeletal muscle) with intermittent dark plates crossing the fibres, but these highly specialized fibres differ fundamentally from those of skeletal muscle in the arrangement of nuclei and in the smaller calibre of the individual fibre. The nuclei are oval and situated along the central axis of the fibre, which may range in size from 12 to 21 micrometres in diameter. Each fibre consists of a bundle of smaller fibres, called myofibrils, each of which passes through the full length of the fibre and is covered by an external limiting membrane known as the sarcolemma.

From S. Jacob and C. Francone, *Structure and Function in Man* (1970); W.B. Saunders Co.

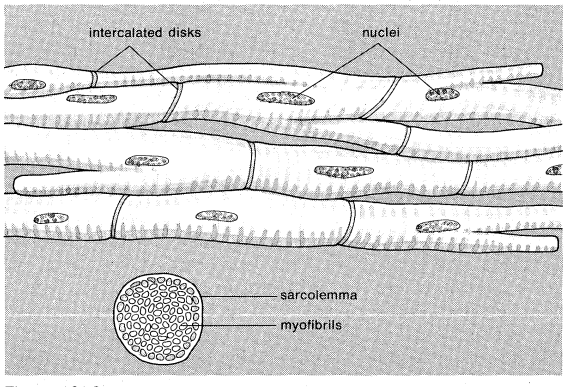



Figure 12: Cardiac muscle.

The individual cardiac muscle cells are striped crosswise throughout, with alternating dark bands that are opaque to light and with light bands that permit the passage of light. Prominent plates of condensed dark bands called intercalated disks, crossing the muscle fibre at uneven intervals, are perhaps the most conspicuous features unique to cardiac muscle.

It is the myocardial layer that causes the heart to contract; the bundles of the muscle fibres are so arranged as to result in a wringing type of movement that efficiently squeezes blood from the heart with each beat (Figure 13). The thickness of the myocardium varies according to the pressure generated to move blood to its destination. The myocardium of the left ventricle, which must drive blood out into the systemic circulation, is, therefore, thickest; the myocardium of the right ventricle, which propels blood to the lungs, is moderately thickened, while the atrial walls are relatively thin.

The endocardium

Forming the inner surface of the myocardial wall is a thin lining called the endocardium. This layer lines the

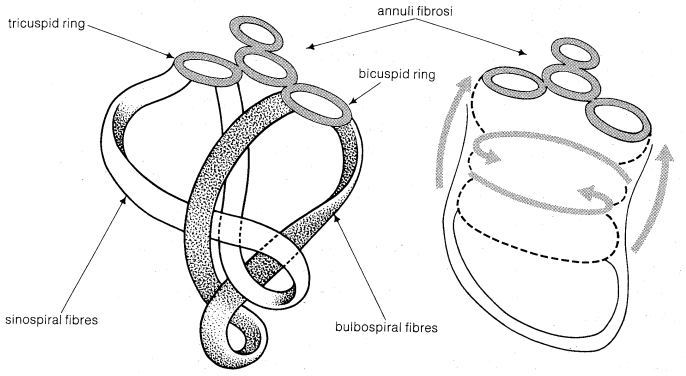



Figure 13: *The myocardium of the ventricles.*
(Left) Spiral muscles of ventricles. (Right) Dotted lines indicate changes in shape of heart due to contraction of ventricular muscle (arrows).

cavities of the heart, covers the valves and small muscles associated with opening and closing of the valves, and is continuous with the lining membrane of the large blood vessels. (S.W.J./Ed.)

*Blood supply to the heart.* Because of the watertight lining of the heart (the endocardium) and the thickness of the myocardium, the heart cannot depend on the blood contained in its own chambers for oxygen and nourishment. It possesses a vascular system of its own, called the coronary arterial system. In the most common distribution, this comprises two major coronary arteries, the right and the left; normally, the left coronary artery divides soon after its origin into two major branches, called the left anterior descending and the circumflex coronary arteries. The right, the left anterior descending, and the left circumflex coronary arteries have many branches and are of almost equal importance. Thus, there are commonly said to be three main functional coronary arteries rather than two.

The right and left coronary arteries originate from the right and left aortic sinuses (the sinuses of Valsalva), which are bulges at the origin of the ascending aorta immediately beyond, or distal to, the aortic valve. The ostium, or opening, of the right coronary artery is in the right aortic sinus and that of the left coronary artery is in the left aortic sinus, just above the aortic valve ring. There is also a non-coronary sinus of Valsalva, which lies to the left and posteriorly at the origin of the ascending aorta. The left coronary arterial system is more important than the right because it supplies blood to the larger left ventricle, and the dimension of the left coronary ostium is larger than that of the right.

The right coronary artery has a lumen diameter of about 2.5 millimetres or more. It supplies the right ventricular outflow tract, the sinoatrial node (the principal pacemaker of the heart), the atrioventricular node, and the bulk of the right ventricle, with branches extending into the interventricular septum and joining with arteriolar branches from the left coronary artery more or less where the two ventricles join.

The left coronary artery

The main stem of the left coronary artery has a luminal diameter often exceeding 4.5 millimetres and is one of the shortest and most important vessels of the body. Usually, it is between one and two centimetres in length, but it may have a length of only two millimetres before dividing. Sometimes the main left coronary artery may actually be missing, with the left coronary ostium having two separate openings for the left anterior descending and the left circumflex arteries. The main left coronary artery divides into its two branches, the anterior descending and the circumflex, while still in the space between the aorta and pulmonary artery. The left anterior descending coronary artery usually begins as a continuation of the left main coronary artery, and its size, length, and distribution are key factors in the balance of the supply of blood to the left ventricle and the interventricular septum. There are many branches of the left anterior descending artery; the first and usually the largest septal branch is important because of its prominent role in supplying blood to the septum.

The left circumflex artery leaves the left main coronary artery to run posteriorly along the atrioventricular groove. It divides soon after its origin into an atrial branch and an obtuse marginal branch. The former branch sometimes has a branch to the sinoatrial node (more usually supplied from the right coronary artery). The obtuse marginal vessel supplies the posterior left ventricular wall in the direction of the apex.

Venous circulation

Venous blood from the heart is carried through veins, which usually accompany the distribution of the distal arteries. These cardiac veins, however, proceed into the atrioventricular grooves anteriorly and posteriorly to form the coronary venous sinus, which opens into the right ventricle immediately below the tricuspid valve.

In addition to these identifiable anatomic arterial and venous channels, nutritional exchange almost certainly takes place between the endocardial ventricular muscle layers and the blood in the cavity of the ventricles. This is of minor importance and probably is an adaptive system in situations of cardiac muscle pathology. (M.F.O.)

**Heartbeat.** *Regulation of heartbeat.* Regular beating of the heart is achieved as a result of the inherent rhythmicity of cardiac muscle; no nerves are located within the heart itself, and no outside regulatory mechanisms are necessary to stimulate the muscle to contract rhythmically. That these rhythmic contractions originate in the cardiac muscle can be substantiated by observing cardiac development in the embryo; cardiac pulsations begin before adequate development of nerve fibres. In addition, it can be demonstrated in the laboratory that even fragments of cardiac muscle in tissue culture continue to contract rhythmically. Furthermore, there is no gradation in degree of contraction of the muscle fibres of the heart, as would be expected if they were primarily under nervous control.

The mere possession of this intrinsic ability is not sufficient, however, to enable the heart to function efficiently. Proper function requires coordination, which is maintained by an elaborate conducting system within the heart that consists primarily of two small, specialized masses of tissue, or nodes, from which impulses originate and of nervelike conduits for the transmission of impulses, with terminal branches extending to the inner surface of the ventricles.

A basic understanding of the method by which an impulse is transmitted is essential before the conduction system that exists within the heart can be properly understood. Electrical potentials exist across membranes of essentially all cells of the body—that is, there is an electrical potential gradient created, generally by an excess of negative ions immediately inside the cell membrane and an equal excess of positive ions on the outside of the membrane, also known as a resting potential. (Ions are atoms or groups of atoms in solution that carry positive or negative electrical charges.) Further, some cells, such as nerve and muscle cells, have the additional distinction of being "excitable"—*i.e.,* capable of conducting impulses along their membranes.

Any factor that suddenly increases the permeability of the cell membrane and allows positive ions to flow through the membrane to the inside, while negative ions flow to the outside, is likely to bring about a sequence of rapid changes in the membrane potential, lasting only a fraction of a second. This change is followed by the return of the membrane potential to its resting value. This sequence of changes in potential is called an action potential and is responsible for the initiation of impulses transmitted, in the case of the heart, along the muscle fibres and the special conducting tissue fibres. Electrical stimulation, application of chemicals, mechanical damage, heat, and cold are among the factors that can bring about a change in the state of a cell membrane, momentarily disturbing its normal resting state and creating an action potential.

Action potential

The action potential occurs in two separate stages called depolarization and repolarization. During the process of depolarization, the normal negative potential inside the muscle fibre is lost, and the membrane potential actually reverses, or becomes slightly positive inside and negative outside the membrane. This process proceeds as a wave along the length of the muscle fibres. It lasts only a fraction of a second before the positive ions begin to resume

their original position on the outside of the membrane, necessary before another impulse can pass. This recovery process is termed repolarization. An action potential is necessary to generate each depolarization wave.

The sinoatrial node (Figure 14) possesses the ability to generate an action potential spontaneously. This highly important structure is a small strip of specialized muscle located in the posterior wall of the right atrium, immediately beneath the point of entry of the superior vena cava. After each action potential is generated in the sinoatrial node, the impulse immediately spreads through the atrial muscle in the form of a ripple pattern similar to the pattern of waves generated when a stone is thrown into a pool of water.

From S. Jacob and C. Francone, *Structure and Function in Man* (1970); W.B. Saunders Co.

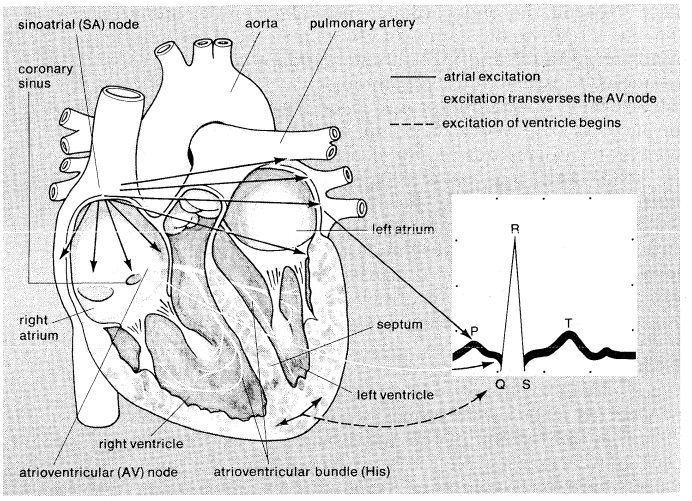



Figure 14: Conducting system of the heart showing source of electrical impulses produced on electrocardiogram.

A few specialized atrial fibres relay this potential, or impulse, to the atrioventricular node, located in the lower part of the right rear atrial wall. To permit sufficient time for complete contraction of the atria before subsequent simultaneous contraction of the ventricles, the impulse is delayed slightly in its passage through the atrioventricular node. The special conducting fibres that leave this node form the bundle of His, which terminates in the multiple branches of the Purkinje network in the left and right ventricles. (The bundle of His was named for the German cardiologist Wilhelm His, Jr., who described it in 1893.) The bundle of His passes between the atrioventricular valves and leads into the interventricular septum, where bundles of fibres in right and left branches project downward beneath the endocardium on either side of the septum. They then curve around the tip of the ventricles and back toward the atria along the side walls.

*Conducting system of the heart.* Both the atrioventricular node, with an intrinsic beat of 40–60 beats per minute, and the Purkinje fibres, with an intrinsic beat of 15–40 beats per minute, are self-stimulating and capable of rhythmic contraction, but at a rate slower than that of the sinoatrial node, which possesses an intrinsic capacity to beat 72 times per minute. Therefore, since recovery time is faster in the sinoatrial node, it controls the rate of the heartbeat, serving as the primary pacemaker. The atrioventricular node, with its 40–60 beats per minute, is termed the secondary pacemaker.

*Effect of ions on heart function.* Potassium, sodium, and calcium have a marked influence on transmission of action potentials within cardiac muscle. In addition, calcium-ion concentration is important in the contractile process. The concentration of ions in extracellular fluids (those outside the cells) also affects cardiac function to a degree.

Excess ions

Excess potassium ions in extracellular fluid can slow the heart rate and cause the heart to dilate and become flaccid (flabby). There is general weakness of cardiac muscle. This weakening of the strength of contraction is caused by a decreased resting membrane potential with resulting decrease in the intensity of the action potential. Calcium ions in excess produce just the opposite effect; the heart goes into spastic contraction. There is, however, little danger of the presence of excess calcium ions within the extracellular fluid in cardiac muscle, for excess calcium is precipitated as salts in the body tissues before a dangerous level is reached in the heart. A depressed cardiac function occurs when sodium ions are present in excess. These ions interfere with the effectiveness of calcium in bringing about normal muscular contraction.

Presumably because of an increased permeability of muscle membrane to ions, temperature may also affect heart function. An increased heart rate occurs as temperature increases, and, conversely, heart rate decreases with decreased temperature.

*Electrocardiogram.* As an impulse travels along the cardiac muscle fibres, an electric current is generated by the flowing ions. This current spreads into the fluids around the heart, and a minute portion actually flows to the surface of the body. An electrocardiogram is a record of this electrical activity (Figure 14) as measured by a device called a galvanometer. Leads (*i.e.*, wires to the galvanometer) are placed on the surface of the body at various points, depending on the type of information desired. An electrocardiogram thus has the prime function of assessing the ability of the heart to transmit the cardiac impulse. Each portion of the cardiac cycle produces a different electrical impulse, causing the characteristic deflections of an electrocardiographic recording needle. The deflections, or waves, on the recording apparatus are, in order, the P, QRS complex, and T waves.

As a wave of depolarization passes over the atria, the impulse is recorded as the P wave. As it continues on through the ventricles, it is registered as the QRS complex. The T wave is caused by currents generated as the ventricles recover from the state of depolarization. This repolarization process occurs in the muscle of the ventricles about 0.25 second after depolarization. There are, therefore, both depolarization and repolarization waves represented in the electrocardiogram. The atria repolarize at the same time that the ventricles depolarize. The atrial repolarization wave is, however, obscured by the larger QRS wave.

*Nervous control of the heart.* Nervous control of the heart is maintained by the parasympathetic fibres in the vagus nerve (parasympathetic) and by the sympathetic nerves. The vagus nerve is the cardiac inhibitor, and the sympathetic nerves are the cardiac excitors. Stimulation of the vagus nerve depresses the rate of impulse formation and atrial contractility and thereby reduces cardiac output and slows the rate of the heart. Parasympathetic stimulation can also produce varying degrees of impaired impulse formation or heart block in diseases of the heart. (In complete heart block the atria and the ventricles beat independently.) Stimulation of the sympathetic nerves increases contractility of both atria and ventricles.

Diastole and systole

The cardiac cycle is defined as that time from the end of one heart contraction to the end of the subsequent contraction and consists of a period of relaxation called diastole followed by a period of contraction called systole. During the entire cycle, pressure is maintained in the arteries; however, this pressure varies during the two periods, the normal diastolic pressure being 80 millimetres of mercury and the normal systolic pressure being 120 millimetres of mercury.

*Blood-pressure regulation and measurement.* Changes in blood pressure may depend on several regulating mechanisms. Arterioles present the main resistance to blood flow. Blood pressure, therefore, can be maintained only if resistance in these arterioles falls each time cardiac output increases. The nerves that control the action of the small muscle fibres of the vessels maintain resistance at a level sufficient for high arterial blood pressure by constricting the channel, or lumen, of the arteriole; during dilation of the vessel, pressure is decreased. Arterial pressure is also affected by the chemical composition of the blood. A decreased oxygen or increased carbon dioxide tension (partial pressure) causes a reflex elevation of blood pressure. Respiratory activity is, therefore, an important regulator of arterial pressure.

The renin–angiotensin system provides hormonal control

of blood pressure. Decreased blood flow to the kidney, changes in posture, or blockage of one or both renal (kidney) arteries may lead to increased production of the enzyme renin by the kidney. This substance causes development in the circulating blood of the substance angiotensin II, which causes blood vessels to contract, with resultant increase in blood pressure.

Receptors in great veins, in the aortic arch (the bend in the aorta above the heart), and the carotid sinus are sensitive to changes in blood pressure as blood is forced from the ventricles. These receptors, known as baroreceptors, help to modify shifts in pressure. When the receptors are stimulated by a rise in arterial pressure, which distends the arterial wall, reflexes are initiated that have an inhibiting effect on the heart, causing it to beat more slowly and with less force. At the same time there is a decrease in the constriction of the blood vessels. A fall in pressure, on the other hand, causes increased sympathetic and decreased parasympathetic nervous stimulation, with resultant increased heart rate and also a subsequent constriction of the blood vessels.

Force of cardiac contraction

The force of the heartbeat depends on the initial length of the heart muscle fibres, the length of the pause in diastole, the oxygen supply, and the integrity and mass of the heart muscle, or myocardium. The greater the initial length of the muscle fibres in the heart, the more forceful will be the contraction. Artificially increasing venous return of blood to the heart distends the heart and intensifies the force of the beat. The greater inflow is handled by an increased output of the heart, without a change in its rate. When the ventricle does not completely fill (for example, after loss of blood), the force of the heartbeat is reduced. When the venous inflow during diastole is increased, as in muscular exercise, the beats become more forceful. If, as a result of excessive filling, the fibres are overstretched, a weak contraction results, with diminished cardiac output; consequently, the heart does not adequately empty itself. The force of the heart is also diminished if the diastolic phase is too short and there is inadequate filling.

Blood pressure is measured with a device called a sphygmomanometer. The pressure of blood within the artery is balanced by an external pressure exerted by air contained in a cuff applied externally around the arm. Actually, it is the pressure within the cuff that is measured. The steps employed in determining blood pressure with a sphygmomanometer are:

1. The cuff is wrapped securely around the arm above the elbow.
2. Air is pumped into the cuff with a rubber bulb until pressure is sufficient to stop the flow of blood in the brachial artery (the principal artery of the upper arm). Pressure within the cuff is shown on the scale of the sphygmomanometer.
3. The observer places a stethoscope over the brachial artery just below the elbow and gradually releases the air from within the cuff. The decreased air pressure permits the blood to flow, filling the artery below the cuff. Faint tapping sounds corresponding to the heartbeat are heard. When the sound is first noted, the air pressure within the cuff is recorded on the scale. This pressure is equal to the systolic blood pressure.
4. As the air in the cuff is further released, the sounds become progressively louder, until the sounds change in quality from loud to soft and finally disappear. The point at which the sound completely disappears should be recorded as diastolic blood pressure.

## THE BLOOD VESSELS

Because of the need for the early development of a transport system within the embryo, the organs of the vascular system are among the first to appear and to assume their functional role. In fact, this system is established in its basic form by the fourth week of embryonic life. At approximately the 18th day of gestation, cells begin to group together between the outer skin (ectoderm) and the inner skin (endoderm) of the embryo. These cells soon become rearranged so that the more peripheral ones join to form a continuous flattened sheet enclosing more centrally placed cells; these cells remain suspended in a fluid medium as primitive blood cells. The tubes then expand and unite to form a network; the primitive blood vessels thus appear.

The blood vessels consist of a closed system of tubes that transport blood to all parts of the body and back to the heart. As in any biologic system, structure and function of the vessels are so closely related that one cannot be discussed without the other's being taken into account.

Arteries transport blood to body tissues under high pressure, which is exerted by the pumping action of the heart. The heart forces blood into these elastic tubes, which recoil, sending blood on in pulsating waves. It is, therefore, imperative that the vessels possess strong, elastic walls to ensure fast, efficient blood flow to the tissues.

Three layers of the artery wall

The wall of an artery consists of three layers (Figures 15 and 16), the innermost consisting of an inner surface of smooth endothelium covered by a surface of elastic tissues: the two form the tunica intima. The tunica media, or middle coat, is thicker in arteries, particularly in the large arteries, and consists of smooth muscle cells intermingled with elastic fibres. The muscle-cell and elastic fibres circle the vessel. In larger vessels the tunica media is composed primarily of elastic fibres. As arteries become smaller, the number of elastic fibres decreases while the number of smooth muscle fibres increases. The outer layer, the tunica adventitia, is the strongest of the three layers. It is composed of collagenous and elastic fibres. (Collagen is a connective-tissue protein.) The tunica adventitia provides a limiting barrier, protecting the vessel from overexpansion. Also characteristic of this layer is the presence of

From S. Jacob and C. Francone, *Structure and Function in Man* (1970); W.B. Saunders Co.

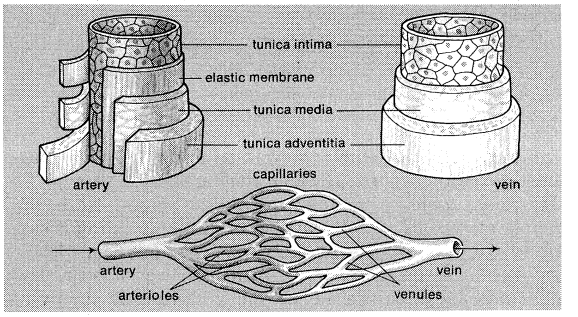



Figure 15: Component parts of arteries and veins.

small blood vessels called the vasa vasorum that supply the walls of larger arteries and veins; the inner and middle layers are nourished by diffusion from the blood as it is transported. The thicker, more elastic wall of arteries enables them to expand with the pulse and to regain their original size.

The transition from artery to arteriole is a gradual one, marked by a progressive thinning of the vessel wall and a decrease in the size of the lumen, or passageway. The tunica intima is still present as a lining covered by a layer of thin longitudinal fibres. A single layer of circular or spiral smooth muscle fibres now makes up the tunica media, and the tunica adventitia consists of connective tissue elements.

Being the last small branches of the arterial system, arterioles must act as control valves through which blood is released into the capillaries. The strong muscular wall of arterioles is capable of completely closing the passageway or permitting it to expand to several times its normal size, thereby vastly altering blood flow to the capillaries. Blood flow is by this device directed to tissues that require it most.

As the arterioles become smaller in size, the three coats become less and less definite, the smallest arterioles consisting of little more than endothelium, or lining, surrounded by a layer of smooth muscle. The microscopic capillary tubules consist of a single layer of endothelium, a continuation of the innermost lining cells of arteries and veins.

As the capillaries converge, small venules are formed whose function it is to collect blood from the capillary beds (*i.e.*, the networks of capillaries). The venules consist of an endothelial tube supported by a small amount of collagenous tissue and, in the larger venules, by a few smooth muscle fibres as well. As venules continue to increase in

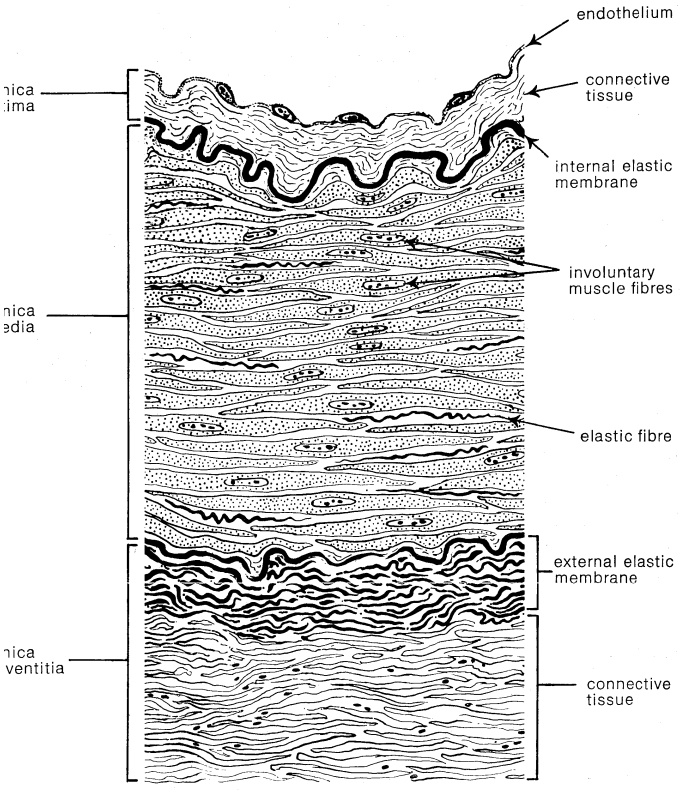



Figure 16: Transverse section of an artery.
From W. Bloom and D. Fawcett, *Textbook of Histology*, 9th ed.; W.B. Saunders Company

size, they begin to exhibit the wall structure that is a characteristic of the arteries, though they are much thinner.

In veins, which function to conduct blood from the peripheral tissues to the heart, an endothelial lining is surrounded by the tunica media, which contains less muscle and elastic tissue than is found in the arterial wall. The outermost layer, tunica adventitia, is composed chiefly of connective tissue. Blood pressure in these vessels is extremely low as compared with that in the arterial system, and blood must exit at an even lower pressure. This creates a need for a special mechanism to keep blood moving on its return to the heart.

**Venous valves**

To achieve this, many veins possess a unique system of valves. These valves, formed by semilunar folds in the tunica intima, are present in pairs and serve to direct the flow of blood to the heart, particularly in an upward direction. As blood flows toward the heart, the flaps of the valves flatten against the wall of the vein; they then billow out to block the opening as the pressure of the blood and surrounding tissues fills the valve pocket. These valves are more abundant in the veins of the extremities than in any other parts of the body.

The veins are more distensible than arteries, and their walls are so constructed as to enable them to expand or contract. A major function of their contractility appears to be to decrease the capacity of the cardiovascular system by constriction of the peripheral vessels in response to the heart's inability to pump sufficient blood.

Veins tend to follow a course parallel to that of arteries but are present in greater number. Their channels are larger than those of arteries, and their walls are thinner. About 60 percent of the blood volume is in the systemic circulation, and 40 percent is normally present in the veins.

**Pulmonary circulation**

The pulmonary circuit consists of the right ventricle; the exiting pulmonary artery and its branches; the arterioles, capillaries, and venules of the lungs; and the pulmonary veins that empty into the left atrium.

The pulmonary trunk, the common stem of the pulmonary arteries, arises from the upper surface of the right ventricle and extends four to five centimetres beyond this origin before dividing into the right and left pulmonary arteries, which supply the lungs. The pulmonary valve, which has two leaflets, or cusps, guards the opening between the right ventricle and the pulmonary trunk. The trunk is relatively thin walled for an artery, having walls approximately twice the thickness of the vena cava and one-third that of the aorta. The right and left pulmonary arteries are short but possess a relatively large diameter. The walls are distensible; the vessels are able, therefore, to accommodate the stroke volume of the right ventricle, which is a necessary function equal to that of the left ventricle.

The pulmonary artery and its branches operate under high pressure in order to accommodate the great force of deoxygenated blood ejected from the right ventricle into the lungs. The pulmonary veins operate under lower pressure as they return oxygenated blood to the left atrium.

The pulmonary trunk passes diagonally upward to the left across the route of the aorta. Between the fifth and sixth thoracic vertebrae (at about the level of the bottom of the breastbone), the trunk divides into two branches—the right and left pulmonary arteries—which enter the lungs. After entering the lungs, the branches go through a process of subdivision, the final branches being capillaries. Capillaries surrounding the air sacs (alveoli) of the lungs pick up oxygen and release carbon dioxide. The capillaries carrying oxygenated blood join larger and larger vessels until they reach the pulmonary veins, which carry oxygenated blood from the lungs to the left atrium of the heart.

**The arteries.** *The aorta and its principal branches.* The aorta is the largest vessel in the systemic circuit, arising from the left ventricle. It is commonly said to have three regions: the ascending aorta, the arch of the aorta, and the descending aorta; the latter may be further subdivided into the thoracic and the abdominal aorta (Figure 17).

**The three regions**

Originating from the ascending portion of the aorta are the right and left coronary arteries, which supply the heart with oxygenated blood. Branching from the arch of the aorta are three large arteries named, in order of origin from the heart, the innominate, the left common carotid, and the left subclavian. These three branches supply the head, neck, and arms with oxygenated blood.

As the innominate (sometimes referred to as the brachiocephalic) artery travels upward toward the clavicle, or collarbone, it divides into the right common carotid and right subclavian arteries. The two common carotid arteries, one branching from the innominate and the other directly from the aorta, then extend in a parallel fashion on either side of the neck to the top of the thyroid cartilage (the principal cartilage in the voice box, or larynx), where they divide, each to become an internal and an external carotid artery. The external carotid arteries give off branches that supply much of the head and neck, while the internal carotids are responsible for supplying the forward portion of the brain, the eye and its appendages, and the forehead and nose.

The two vertebral arteries, one arising as a branch of the innominate and the other as a branch of the left subclavian artery, unite at the base of the brain to form the basilar artery, which in turn divides into the posterior cerebral arteries. The blood supply to the brain is derived mainly from vessels that may be considered as branches of the circle of Willis, which is made up of the two vertebral and the two internal carotid arteries and connecting arteries between them.

The arms are supplied by the subclavian artery on the left and by the continuation of the innominate on the right. At approximately the border of the first rib, both of these vessels become known as the axillary artery; this, in turn, becomes the brachial artery as it passes down the upper arm. At about the level of the elbow, the brachial artery divides into two terminal branches, the radial and ulnar arteries, the radial passing downward on the distal (thumb) side of the forearm, the ulnar on the medial side. Interconnections (anastomoses) between the two, with branches at the level of the palm, supply the hand and wrist.

**Thoracic aorta**

The thoracic (chest) portion of the descending aorta gives off branches that supply the viscera (visceral branches) and the walls surrounding the thoracic cavity (parietal branches). The visceral branches provide blood for the pericardium, lungs, bronchi, lymph nodes, and esophagus.

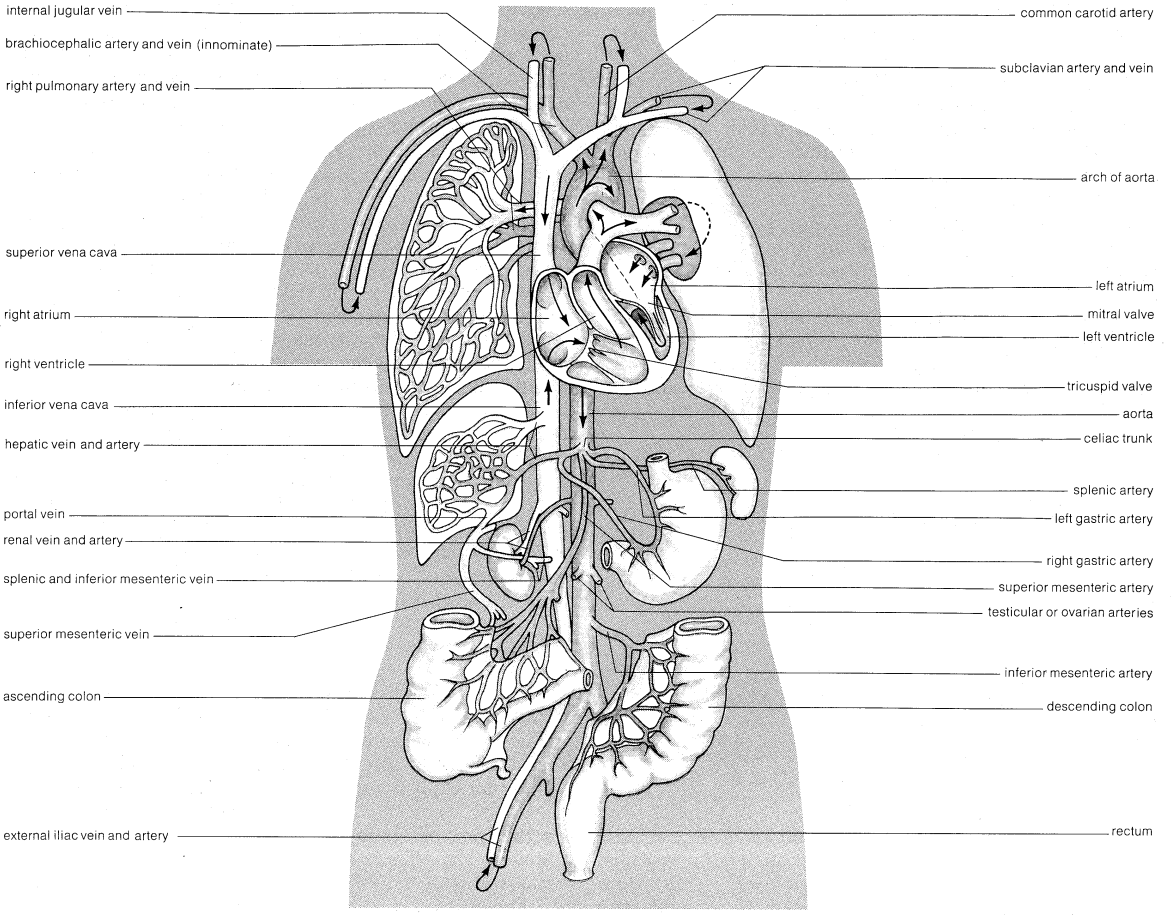



Figure 17: Human arterial supply and venous drainage of the organs.
From S. Jacob and C. Francone, *Structure and Function in Man* (1970); W.B. Saunders Co.

The parietal vessels supply the intercostal muscles (the muscles between the ribs) and the muscles of the thoracic wall; they supply blood to the membrane covering the lungs and lining the thoracic cavity, the spinal cord, the vertebral column, and a portion of the diaphragm.

As the aorta descends through the diaphragm, it becomes known as the abdominal aorta and again gives off both visceral and parietal branches. Visceral vessels include the celiac, superior mesenteric, and inferior mesenteric, which are unpaired, and the renal and testicular or ovarian, which are paired. The celiac artery arises from the aorta a short distance below the diaphragm and almost immediately divides into the left gastric artery, serving part of the stomach and esophagus; the hepatic artery, which primarily serves the liver; and the splenic artery, which supplies the stomach, pancreas, and spleen.

The superior mesenteric artery arises from the abdominal aorta just below the celiac artery. Its branches supply the small intestine and part of the large intestine. Arising several centimetres above the termination of the aorta is the inferior mesenteric artery, which branches to supply the lower part of the colon. The renal arteries pass to the kidneys. The testicular or ovarian arteries supply the testes in the male and the ovaries in the female, respectively.

Parietal branches of the abdominal aorta include the inferior phrenic, serving the suprarenal (adrenal) glands, the lumbar, and the middle sacral arteries. The lumbar arteries are arranged in four pairs and supply the muscles of the abdominal wall, the skin, the lumbar vertebrae, the spinal cord, and the meninges (or spinal-cord coverings).

The abdominal aorta divides into two common iliac arteries, each of which descends laterally and gives rise to external and internal branches. The right and left external iliac arteries are direct continuations of the common iliacs and become known as the femoral arteries after passing through the inguinal region, giving off branches that supply structures of the abdomen and lower extremities.

At a point just above the knee, the femoral artery continues as the popliteal artery; from this arise the posterior and anterior tibial arteries. The posterior tibial artery is a direct continuation of the popliteal, passing down the lower leg to supply structures of the posterior portion of the leg and foot.

Arising from the posterior tibial artery a short distance below the knee is the peroneal artery; this gives off branches that nourish the lower leg muscles and the fibula (the smaller of the two bones in the lower leg) and terminates in the foot. The anterior tibial artery passes down the lower leg to the ankle, where it becomes the dorsalis pedis artery, which supplies the foot.

*Pulse.* An impulse can be felt over an artery that lies near the surface of the skin. The impulse results from alternate expansion and contraction of the arterial wall because of the beating of the heart. When the heart pushes blood into the aorta, the blood's impact on the elastic walls creates a pressure wave that continues along the arteries. This impact is the pulse. All arteries have a pulse, but it is most easily felt at points where the vessel approaches the surface of the body.

Discerning the pulse

The pulse is readily distinguished at the following locations: (1) at the point in the wrist where the radial artery approaches the surface; (2) at the side of the lower jaw where the external maxillary (facial) artery crosses it; (3) at the temple above and to the outer side of the eye, where the temporal artery is near the surface; (4) on the side of the neck, from the carotid artery; (5) on the inner side of the biceps, from the brachial artery; (6) in the groin, from the femoral artery; (7) behind the knee, from the popliteal artery; (8) on the upper side of the foot, from the dorsalis pedis artery.

The radial artery is most commonly used to check the pulse. Several fingers are placed on the artery close to the wrist joint. More than one fingertip is preferable because of the large, sensitive surface available to feel the pulse wave. While the pulse is being checked, certain data are recorded, including the number and regularity of beats per

minute, the force and strength of the beat, and the tension offered by the artery to the finger. Normally, the interval between beats is of equal length.

**The veins.** Venules collect blood from the capillaries and the blood channels known as sinusoids and unite to form progressively larger veins that terminate as the great veins, or venae cavae. In the extremities there are superficial and deep veins; the superficial lie just under the skin and drain the skin and superficial fasciae (sheets of fibrous tissue), while the deep veins accompany the principal arteries of the extremities and are similarly named. Interconnections between the superficial and the deep veins are frequent.

Venous blood enters the right atrium from three sources: the heart muscle by way of the coronary sinus; the upper body by way of the superior vena cava; and the lower body by way of the inferior vena cava (Figures 18 and 19).

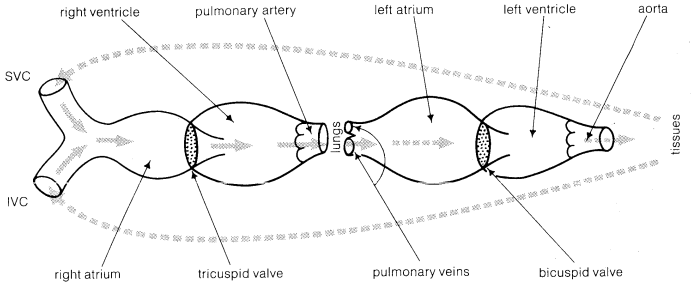



Figure 18: *Venous blood flow through the heart, lungs, and tissues of the body.*
Pulmonary circulation is indicated by solid arrows; systemic circulation by dashed arrows. SVC is superior vena cava; IVC is inferior vena cava.

*Superior vena cava and its tributaries.* Tributaries from the head and neck, the arms, and part of the chest unite to form the superior vena cava. Venous channels called venous sinuses lie between the two layers of the dura mater, the outer covering of the brain; they possess no valves. Venous drainage of the brain is effected by these sinuses and communicating vessels. The internal jugular vein is a continuation of this system downward through the neck; it receives blood from parts of the face, neck, and brain. At approximately the level of the collarbone, each unites with the subclavian vein of that side to form the innominate veins.

The external jugular vein is formed by the union of its tributaries near the angle of the lower jaw, or mandible. It drains some of the structures of the head and neck and pours its contents along with the subclavian into the innominate vein of the same side. All of the veins of the arm are tributaries of the subclavian vein of that side. They are found in both superficial and deep locations and possess valves. Most of the deep veins are arranged in pairs with cross connections between them.

Superficial and deep veins of the arm

Venous drainage of the hand is accomplished superficially by small anastomosing (interconnecting) veins that unite to form the cephalic vein, coursing up the radial (thumb) side of the forearm, and the basilic vein, running up the ulnar side of the forearm and receiving blood from the hand, forearm, and arm. The deep veins of the forearm include the radial veins, continuations of deep anastomosing veins of the hand and wrist, and the ulnar veins, both veins following the course of the associated artery. The radial and ulnar veins converge at the elbow to form the brachial vein; this, in turn, unites with the basilic vein at the level of the shoulder to produce the axillary vein. At the outer border of the first rib, the axillary vein becomes the subclavian vein, the terminal point of the venous system characteristic of the upper extremity.

The subclavian, external jugular, and internal jugular veins all converge to form the innominate vein. The right and left innominate veins terminate in the superior vena cava, which opens into the upper posterior portion of the right atrium.

In addition to the innominate veins, the superior vena cava receives blood from the azygous vein and small veins from the mediastinum (the region between the two lungs) and the pericardium. Most of the blood from the back and from the walls of the chest and abdomen drains into veins lying alongside the vertebral bodies (the weight-bearing portions of the vertebrae; Figure 19). These veins form what is termed the azygous system, which serves as a connecting link between the superior and inferior vena cava. The terminal veins of this system are the azygous, hemiazygous, and accessory hemiazygous veins. At the level of the diaphragm, the right ascending lumbar vein continues upward as the azygous vein, principal tributaries of which are the right intercostal veins, which drain the muscles of the intercostal spaces. It also receives tributaries from the esophagus, lymph nodes, pericardium, and right lung, and it enters into the superior vena cava at about the level of the fourth thoracic vertebra.

The left side of the azygous system varies greatly among individuals. Usually the hemiazygous vein arises just below the diaphragm as a continuation of the left ascending lumbar vein and terminates in the azygous vein. Tributaries of the hemiazygous drain the intercostal muscles, the esophagus, and a portion of the mediastinum. The accessory hemiazygous usually extends downward as a continuation of the vein of the fourth intercostal space, receiving tributaries from the left intercostal spaces and the left bronchus. It empties into the azygous vein slightly above the entrance of the hemiazygous.

*Inferior vena cava and its tributaries.* The inferior vena cava is a large, valveless, venous trunk that receives blood from the legs, the back, and the walls and contents of the abdomen and pelvis.

The foot is drained primarily by the dorsal venous arch, which crosses the top of the foot not far from the base of the toes. The arch is connected with veins that drain the sole. Superficially the lower leg is drained by the large and small saphenous veins, which are continuations of the dorsal venous arch. The small saphenous vein extends up the back of the lower leg to terminate usually in the popliteal vein. There is some interconnection with deep veins and with the great saphenous vein. The latter vein, the longest in the body, extends from the dorsal venous arch up the inside of the lower leg and thigh, receiving venous branches from the knee and thigh area and terminating in the femoral vein.

Superficial veins of the leg

Most blood from the lower extremity returns by way of

From S. Jacob and C. Francone, *Structure and Function in Man* (1970); W.B. Saunders Co.

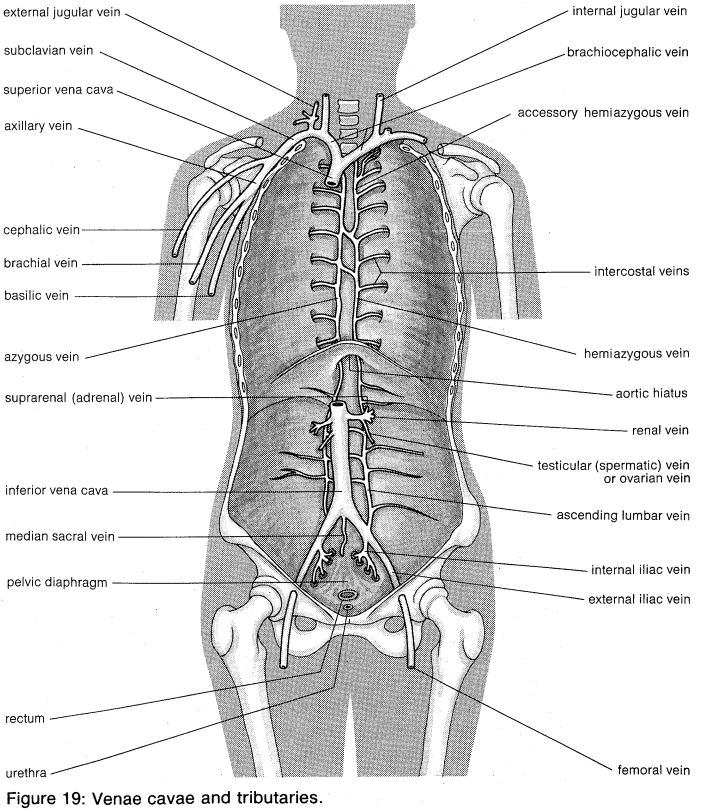



Figure 19: Venae cavae and tributaries.

the deep veins. These include the femoral and popliteal veins and the veins accompanying the anterior and posterior tibial and peroneal arteries. The anterior and posterior tibial veins originate in the foot and join at the level of the knee to form the popliteal vein; the latter becomes the femoral vein as it continues its extension through the thigh.

At the level of the inguinal ligament (which is at the anterior, diagonal border between the trunk and the thigh), the femoral vein becomes known as the external iliac vein; the latter unites with the internal iliac vein to form the common iliac vein. The internal iliac vein drains the pelvic walls, viscera, external genitalia, buttocks, and a portion of the thigh. Through the paired common iliac veins, the legs and most of the pelvis are drained. The two common iliacs then unite at a level above the coccyx (the lowest bone in the spine) to become the inferior vena cava. As it courses upward through the abdomen, the inferior vena cava receives blood from the common iliacs and from the lumbar, renal, suprarenal, and hepatic veins before emptying into the right atrium.

The pairs of lumbar veins (which drain blood from the loins and abdominal walls) are united on each side by a vertical connecting vein, the ascending lumbar vein; the right ascending lumbar vein continues as the azygous and the left as the hemiazygous. These veins usually enter separately into the inferior vena cava.

Renal veins lie in front of the corresponding renal artery; the right renal vein receives tributaries exclusively from the kidney, while the left receives blood from a number of other organs as well. The right suprarenal vein terminates directly in the inferior vena cava as does the right phrenic, above the gonadal vein. Two or three short hepatic trunks empty into the inferior vena cava as it passes through the diaphragm.

*Portal system.* The portal system may be described as a specialized portion of the systemic circulatory system. It is unique in that blood from the spleen, stomach, pancreas, and intestine first passes through the liver before it moves on to the heart. Blood flowing to the liver comes from the hepatic artery (20 percent) and the portal vein (80 percent); blood leaving the liver flows through the hepatic vein and then empties into the inferior vena cava. The hepatic arterial blood supplies oxygen requirements for the liver. Blood from the abdominal viscera, particularly the intestinal tract, passes into the portal vein and then into the liver. Substances in the portal blood are processed by the liver (see the article DIGESTION AND DIGESTIVE SYSTEMS).

Hepatic artery and portal vein

*Venous pulmonary system.* From the pulmonary capillaries, in which blood takes on oxygen and gives up carbon dioxide, the oxygenated blood in veins is collected first into venules and then into progressively larger veins; it finally flows through four pulmonary veins, two from the hilum of each lung. (The hilum is the point of entry on each lung for the bronchus, blood vessels, and nerves.) These veins then pass to the left atrium, where their contents are poured into the heart.

**The capillaries.** The vast network of some 10,000,000,000 microscopic capillaries functions to provide a method whereby fluids, nutrients, and wastes are exchanged between the blood and the tissues. Even though microscopic in size, the largest capillary being approximately 0.2 millimetre in diameter (about the width of the tip of a pin), the great network of capillaries serves as a reservoir normally containing about one-sixth of the total circulating blood volume. The number of capillaries in active tissue, such as muscle, liver, kidney, and lungs, is greater than the number in tendon or ligament, for example; the cornea of the eye, epidermis, and hyaline cartilage (semitransparent cartilage such as is found in joints) are devoid of capillaries.

The interconnecting network of capillaries into which the arterioles empty is characterized not only by microscopic size but also by extremely thin walls only one cell thick. The vessels are simply tubular continuations of the inner lining cells of the larger vessels, normally uniform in size, usually three to four endothelial cells in circumference, except toward the venous terminations, where they become slightly wider, four to six cells in circumference. A thin membrane, called a basement membrane, surrounds these cells and serves to maintain the integrity of the vessel.

A single capillary unit consists of a branching and interconnecting (anastomosing) network of vessels, each averaging 0.5 to 1 millimetre in length. The wall of the capillary is extremely thin and acts as a semipermeable membrane that allows substances containing small molecules, such as oxygen, carbon dioxide, water, fatty acids, glucose, and ketones, to pass through the membrane. Oxygen and nutritive material pass into the tissues through the wall at the arteriolar end of the capillary unit; carbon dioxide and waste products move through the membrane into the vessel at the venous end of the capillary bed. Constriction and dilation of the arterioles is primarily responsible for regulating the flow of blood into the capillaries. Muscular gatekeepers, or sphincters, in the capillary unit itself, however, serve to direct the flow to those areas in greatest need.

There are three modes of transport across the cellular membrane of the capillary wall. Substances soluble in the lipid (fatty) membrane of the capillary cells can pass directly through these membranes by a process of diffusion. Some substances needed by the tissues and soluble in water but completely insoluble in the lipid membrane pass through minute water-filled passageways, or pores, in the membranes by a process called ultrafiltration. Only $^{1}/_{1,000}$ of the surface area of capillaries is represented by these pores. Other substances, such as cholesterol, are transported by specific receptors in the endothelium.

In the fetus, oxygenated blood is carried from the placenta to the fetus by the umbilical vein. It then passes to the inferior vena cava of the fetus by way of a vessel called the ductus venosus. From the inferior vena cava, the blood enters the right atrium, then passes through the foramen ovale into the left atrium; from there it moves into the left ventricle and out the aorta, which pumps the oxygenated blood to the head and upper extremities. Blood from the upper extremities returns via the superior vena cava into the right atrium, where it is largely deflected into the right ventricle.

Fetal circulation

From the right ventricle, a portion of the blood flows into the pulmonary artery to the lungs. The largest fraction flows through an opening, the ductus arteriosus, into the aorta. It enters the aorta beyond the point at which the blood of the head leaves. Some of the blood supplies the lower portion of the body. The remainder returns to the placenta via the umbilical arteries, which branch off from the internal iliac arteries.

The changes that take place at birth and that permit routing of the blood through the pulmonary system instead of the umbilical vessels have been described above in the section on the origin and development of the heart.

## EVALUATING THE CARDIOVASCULAR SYSTEM

Certain diagnostic techniques with respect to the heart and blood vessels are important factors in determining the degree of disease and their appropriate medical and surgical treatment.

**Invasive techniques.** *Right-heart catheterization.* Right-heart catheterization is performed by insertion of a catheter (a long tube) into the cubital vein (at the bend of the elbow), the saphenous vein (in the inner thigh), or the femoral vein (at the groin). The catheter, which is opaque to X ray, is advanced into the right atrium, right ventricle, and pulmonary artery under fluoroscopy. This procedure makes it possible to measure pressure and oxygen saturation in the right heart chamber itself and thus to diagnose abnormalities in the valves.

*Left-heart catheterization.* Left-heart catheterization is accomplished by introducing a catheter into the brachial or femoral artery (in the upper arm and thigh, respectively) and advancing it through the aorta across the aortic valve and into the left ventricle. Mitral and aortic valvular defects and myocardial disease can be evaluated by this technique.

*Angiocardiography and arteriography.* Angiocardiography permits direct visualization of the chambers and great vessels of the heart from injections of dyes that are opaque to X rays. Anatomic defects, such as congenital and

acquired lesions, can be detected readily. Left ventriculography (X-ray pictures of the left ventricle) provides information about the synchrony and adequacy of the forces of contraction in areas of the left ventricle. Arteriography (X-ray pictures of an artery after the injection of dyes that are opaque to X rays) of the coronary arteries permits identification, localization, and assessment of the extent of obstructive lesions within these arteries. It is the most important means of defining the presence and severity of coronary atherosclerosis and, in conjunction with left ventriculography, the related state of myocardial function. Although invasive techniques involving left ventricular catheterization and radio contrast angiocardiography and arteriography provide reliable measurements of ejection fraction and regional formation, they have limited applications.

Echocardiography

**Noninvasive techniques.** The term echocardiography refers to a group of tests that use ultrasound (sound waves above frequencies audible to humans) to examine the heart and record information in the form of echoes, or reflected sonic waves. M-mode echocardiography records the amplitude and the rate of motion of moving objects, such as valves, along a single line with great accuracy. M-mode echocardiography, however, does not permit effective evaluation of the shape of cardiac structures, nor does it depict lateral motion (*i.e.,* motion perpendicular to the ultrasonic beam). Real-time (cross-sectional or two-dimensional) echocardiography depicts cardiac shape and lateral movement not available in M-mode echocardiography by moving the ultrasonic beam very rapidly, and such recording may be displayed on film or videotape. New techniques allow measurement by ultrasonography of rates of flow and pressures, for example, across heart valves.

Radionuclide imaging (radioactive nuclides) provides a safe, quantitative evaluation of cardiac function and a direct measurement of myocardial blood flow and myocardial metabolism. Radionuclide imaging is used to evaluate the temporal progress of cardiac disease, hemodynamics, and the extent of myocardial damage during and after infarction and to detect pulmonary infarction following emboli. The primary requirement of radionuclide imaging is that the bolus of radionuclide should remain within the blood vessels during its first passage through the right and left sides of the heart. The second requirement is that the physical properties of the radionuclide be satisfactory with respect to the instrumentation being used.

The radionuclide used in virtually all phases of radionuclide imaging is technetium-99. It has the disadvantage of a long half-life (six hours), however, and other radionuclides with shorter half-lives are also used. These radionuclides all emit gamma rays, and a scintillation camera is used to detect gamma-ray emission. The data are assessed with the R wave of the electrocardiogram as a time marker for the cardiac cycle. Radionuclide cineangiography is a further development of radionuclide imaging. These techniques are used to assess myocardial damage, left ventricular function, valve regurgitation, and, with the use of radionuclide potassium analogues, myocardial perfusion.

Measures of heart metabolism

There are techniques that measure metabolism in the myocardium using the radiotracer method (*i.e.,* a radioactive isotope replaces a stable element in a compound, which is then followed as it is distributed through the body). Positron emission tomography uses positron radionuclides that can be incorporated into true metabolic substrates and consequently can be used to chart the course of selected metabolic pathways, such as myocardial glucose uptake and fatty-acid metabolism. Magnetic resonance imaging (MRI; also called nuclear magnetic resonance [NMR]), also allows high resolution tomographic (one-plane) and three-dimensional imaging of tissues. Magnetic resonance imaging uses magnetic fields and radio frequencies to penetrate bone and obtain clear images of the underlying tissues. (M.F.O./S.W.J.)

## Cardiovascular system diseases and disorders

Diseases of the heart and blood vessels constitute one of the major human health problems of modern times. Life depends on the functioning of the heart; thus, the heart is involved in all death, but this does not account for its prominence in causing death. To some degree, as medical science advances, more people are saved from other illnesses, only to die from one of the unsolved and uncontrolled disorders of the cardiovascular system. Some forms of cardiovascular diseases are becoming less frequent causes of death, and continued research and preventive measures may provide even greater benefits.

Heart disease as such was not recognized in nontechnological cultures, but the beating heart and its relationship to death have always been appreciated. Sudden death, now usually attributed to heart disease, was recognized as early as the 5th century BC by the Greek physician Hippocrates and was noted to be more common in the obese. The role of disease in affecting the heart itself did not become apparent until the 17th century, when examination of the body after death became acceptable.

Gradually, the involvement of the heart valves, the blood vessels, and the heart muscle was observed and categorized in an orderly fashion. The circulation of the blood through the heart was described in 1628 by the British physician William Harvey. The recognition of the manifestations of heart failure came later, as did the ability to diagnose heart ailments by physical examination through the techniques of percussion (thumping), auscultation (listening) with the stethoscope, and other means. It was not until early in the 20th century that the determination of arterial blood pressure and the use of X rays for diagnosis became widespread.

In 1912 James Bryan Herrick, a Chicago physician, first described what he called coronary thrombosis (he was describing symptoms actually caused by myocardial infarction; see below). Angina pectoris (described and discussed in a later section) had been recorded centuries earlier. Cardiovascular surgery in the modern sense began in the 1930s, and open-heart surgery in the 1950s.

Incidence of heart disease

The exact incidence of heart disease in the world population is difficult to ascertain because complete and adequate public health figures for either prevalence or deaths are not available. In the more technologically developed countries of the world, such as the United States, the United Kingdom, and most other European countries, arteriosclerotic heart disease (heart disease resulting from thickening and hardening of the artery walls; see below) constitutes by far the most predominant form. In other areas, such as the countries of Central Africa, other forms of heart disease, often nutritional in nature, are a common cause of death. In Asia and the islands of the Pacific, hypertensive cardiovascular disease, disease involving high blood pressure, constitutes a major health hazard. (J.V.W./Ed.)

### CONGENITAL HEART DISEASE

The heart's complicated evolution during embryological development presents the opportunity for many different types of congenital defects to occur. Congenital heart disease is one of the important types of diseases affecting the cardiovascular system, with an incidence of about eight per 1,000 live births. In most patients the causes appear to fit in the middle of a continuum from primarily genetic to primarily environmental.

Of the few cases that have a genetic nature, the defect may be the result of a single mutant gene, while in others it may be associated with a chromosomal abnormality, the most common of which is Down's syndrome, in which about 50 percent of afflicted children have a congenital cardiac abnormality. In the even smaller number of cases of an obvious environmental cause, a variety of specific factors are evident. The occurrence of rubella (German measles) in a woman during the first three months of pregnancy is caused by a virus and is associated in the child with patent ductus arteriosus (nonclosure of the opening between the aorta and pulmonary artery). Other viruses may be responsible for specific heart lesions, and a number of drugs, including antiepileptic agents, are associated with an increased incidence in congenital heart disease.

In most cases, congenital heart disease is probably caused by a variety of factors, and any genetic factor is usually unmasked only if it occurs together with the appropriate environmental hazard. The risk of a sibling of a child

with congenital heart disease being similarly affected is between 2 and 4 percent. The precise recurrence can vary for individual congenital cardiovascular lesions.

Prenatal diagnosis

Prenatal diagnosis of congenital cardiovascular abnormalities is still at an early stage. The most promising technique is ultrasonography, used for many years to examine the fetus in utero. The increasing sophistication of equipment has made it possible to examine the heart and the great vessels from 16 to 18 weeks of gestation onward and to determine whether defects are present. Amniocentesis (removal and examination of a small quantity of fluid from around the developing fetus) provides a method by which the fetal chromosomes can be examined for chromosomal abnormalities associated with congenital heart disease. In many children and adults the presence of congenital heart disease is detected for the first time when a cardiac murmur is heard. A congenital cardiovascular lesion is rarely signaled by a disturbance of the heart rate or the heart rhythm.

Congenital cardiac disturbances are varied and may involve almost all components of the heart and great arteries. Some may cause death at the time of birth, others may not have an effect until early adulthood, and some may be associated with an essentially normal life span. Nonetheless, about 40 percent of all untreated infants born with congenital heart disease die before the end of their first year.

Congenital heart defects can be classified into cyanotic and noncyanotic varieties. In the cyanotic varieties, a shunt bypasses the lungs and delivers venous (deoxygenated) blood from the right side of the heart into the arterial circulation. The infant's nail beds and lips have a blue colour due to the excess deoxygenated blood in their systems. Some infants with severe noncyanotic varieties of congenital heart disease may fail to thrive and may have breathing difficulties.

**Abnormalities of individual heart chambers.** Abnormalities of the heart chambers may be serious and even life threatening. In hypoplastic left heart syndrome, the left-sided heart chambers, including the aorta, are underdeveloped. Infants born with this condition rarely survive more than two or three days. In other cases, only one chamber develops adequately. Survival often depends on the presence of associated compensatory abnormalities such as continued patency of the ductus arteriosus or the presence of a septal defect, which may allow either decompression of a chamber under elevated pressure or beneficial compensatory intracardiac shunting either from right to left or from left to right.

**Abnormalities of the atrial septum.** The presence of a septal defect allows blood to be shunted from the left side of the heart to the right, with an increase in blood flow and volume within the pulmonary circulation. Over many years the added burden on the right side of the heart and the elevation of the blood pressure in the lungs may cause the right side of the heart to fail.

Locations of defects

Defects in the atrial septum may be small or large and occur most commonly in the midportion in the area prenatally occupied by the aperture called the foramen ovale. Defects lower on the atrial septum may involve the atrioventricular valves and may be associated with incompetence of these valves. In its most extreme form, there may be virtually no septum between the two atrial chambers. Atrial septal defect is a noncyanotic type of congenital heart disease and usually is not associated with serious disability during childhood. A small defect may be associated with problems in young adults, although deterioration can occur in later life. Atrial septal defects, unless small, must usually be closed in childhood.

**Abnormalities of the ventricular septum.** Defects in the interventricular septum, the partition that separates the lower chambers of the heart, may be small or large, single or multiple, and may exist within any part of the ventricular septum. Small defects are among the most common congenital cardiovascular abnormalities and may be less life threatening, since many such defects close spontaneously. Small defects often create loud murmurs but, because they restrict the flow of blood from left to right, no significant change in the circulation occurs. On the other hand, when a defect is large a significant amount of blood is shunted from the left ventricle to the right, with a high flow and volume of blood into the pulmonary circulation.

The pulmonary circulation may be damaged by the stresses imposed by a high blood flow over a long period of time. If unchecked, this damage can become irreversible. A further hazard in both small and large ventricular septal defects is the increased risk of bacterial endocarditis (inflammation of the heart lining as a result of bacterial infection). This risk is likely to be high during procedures such as dental extractions, when infection may enter the bloodstream.

Ventricular septal defects are often combined with other congenital cardiac defects. The best known of these is tetralogy of Fallot, named for the French physician Étienne-Louis-Arthur Fallot, who first described it. In this condition there is a ventricular septal defect, pulmonary stenosis (narrowing of the opening to the pulmonary artery), deviation of the aorta to override the ventricular septum above the ventricular septal defect, and right ventricular hypertrophy (thickening of the muscle of the right ventricle). As a result of the obstruction imposed by the pulmonary stenosis, deoxygenated venous blood is shunted from the right to the left side of the heart into the arterial circulation. A child with this cyanotic form of congenital heart disease can survive beyond infancy, but few survive to adulthood without surgery.

Tetralogy of Fallot

**Abnormal origins of the great arteries.** In many complex forms of congenital heart disease, the aorta and pulmonary artery do not originate from their normal areas of the ventricles. In one of the most common of such cases—transposition of the great arteries—the aorta originates from the right ventricle and receives deoxygenated blood from the superior and inferior venae cavae, and the pulmonary artery arises from the left ventricle and receives fully oxygenated pulmonary venous blood. Survival in such cases depends on a naturally occurring communication between the two sides of the heart that allows oxygenated blood to enter the aorta; if such a communication is not present naturally, it may be created medically or surgically. Both the aorta and the pulmonary artery may originate from the right ventricle; this form of abnormal origin of the arteries usually is associated with a ventricular septal defect and, on occasion, pulmonary stenosis. This combination of defects is a severe form of cyanotic heart disease.

**Abnormalities of the valves.** The most common congenital abnormality of the cardiac valves affects the aortic valve. The normal aortic valve usually has three cusps, or leaflets, but the valve is bicuspid in 1 to 2 percent of the population. A bicuspid aortic valve is not necessarily life threatening, but in some persons it becomes thickened and obstructed (stenotic). With age the valve may also become incompetent or act as a nidus (focus of infection) for bacterial endocarditis. Congenital aortic valve stenosis, if severe, results in hypertrophy of the left ventricular myocardium and may rarely be responsible in asymptomatic individuals for sudden death. Even minor forms of aortic valve stenosis may grow progressively severe, and are likely, with the passage of time, to require surgical treatment.

In contrast to aortic valve stenosis, pulmonary valve stenosis, if mild, is usually well tolerated and does not require surgical treatment. More severe forms of the disease may require surgery or balloon dilation (see below *Surgical treatment of the heart*).

**Abnormalities of the myocardium and endocardium.** Congenital abnormalities in the myocardium, for example, tumours, may be present at birth, but they are rare. Abnormalities of the endocardium may be present at birth, but they are also rare. They include fibroelastosis, a disease in which the endocardium develops a thick fibrous coat that interferes with the normal contraction and relaxation of the heart. This condition cannot be treated surgically and is usually life threatening.

**Abnormalities of the coronary arteries.** The coronary arteries may arise abnormally from a pulmonary artery rather than from the aorta, with the result that deoxygenated blood instead of oxygenated blood flows through

the heart muscle. Abnormal openings, called coronary arterial venous fistulas, may be present between the coronary arteries and chambers of the heart. One or more of the three main coronary arteries may be absent. If necessary, most coronary arterial abnormalities can be corrected surgically.

**Abnormalities of the aorta.** One of the most common congenital cardiovascular abnormalities involves the aorta. In coarctation of the aorta there is a narrowing of the aortic wall usually at that portion of the aorta just beyond the site at which the main blood vessel to the left arm (the subclavian artery) originates. As a result of the narrowing or obstruction at this point, blood flow to the lower half of the body is diminished and hypertension develops in the upper half of the body. This defect may give rise to heart failure in early infancy or complications in later childhood and adulthood.

During fetal life and immediately after birth, the ductus arteriosus is a channel that connects the pulmonary artery and the first segment of the descending thoracic aorta. The function of this duct in utero is to shunt blood away from the lungs. If the ductus remains open after birth, excessive blood may flow into the lungs, resulting in pulmonary congestion and heart failure. Spontaneous closure of the ductus arteriosus may be delayed in premature newborn infants, exacerbating the respiratory problems common to them. If necessary, the ductus arteriosus can be induced to close with drugs in premature infants or surgically closed in older infants and children. Finally, there may be direct communication between the aorta and pulmonary arteries because the truncus arteriosus has either partially or completely failed to partition.

Patent ductus arteriosus

**Anomalous pulmonary venous return.** The pulmonary veins from the right and left lungs may connect either directly or indirectly to the right, instead of the left, atrium. In this condition the abnormal venous channel draining to the right side of the heart may become obstructed. Infants born with total anomalous (abnormal) pulmonary venous drainage usually develop problems within the first few weeks or months and thus require cardiac surgery. Partial forms of anomalous pulmonary venous return, in which only one or two pulmonary veins are connected abnormally, may have few symptoms, although surgical correction may be done if required.

**Anomalies of the vena cava.** The most common abnormalities of the venae cavae, the major veins returning venous blood to the right side of the heart, are a persistent left superior vena cava (normally there is only one superior vena cava opening to the right side of the heart) and an abnormal connection of the inferior vena cava to the heart. These abnormalities are frequently associated with intracardiac structural faults. (M.F.O./M.J.G.)

## ACQUIRED HEART DISEASE

**Atherosclerosis.** Atherosclerosis is a type of thickening and hardening of the medium- and large-sized arteries. When affecting the coronary arteries, it accounts for a large proportion of heart attacks and cases of ischemic (inadequate blood supply to a region due to a constriction or obstruction of a blood vessel) heart disease. Atherosclerosis also accounts for many strokes (those due to cerebral ischemia and infarction), numerous instances of peripheral vascular disease, and most aneurysms (a dilation in the wall of a blood vessel to form a blood-filled sac) of the aorta, which can rupture and cause fatal hemorrhage.

The basis of coronary artery disease is the slow development of areas of thickening in the coronary arteries. These thickenings are called atherosclerotic plaques or atheromatous lesions and they develop early in life, progressing over a period of many years with phases of quiescence or even regression interspersed with periods of progression. Such coronary atheromatous lesions are found in virtually all adults in the industrialized world although most persons never have signs or symptoms of heart disease. In others, however, atheromatous lesions intrude into the lumen of the coronary arteries and progressively impede blood flow to the myocardium, leading to the clinical syndromes of coronary heart disease (Figure 20).

Coronary atherosclerosis

Two major elements of the atheromatous lesion determine its progress. One is the accumulation of cholesterol at the points of thickening and the other is the incorporation of minute clots, or thrombi, into the endothelial (inner) surface of the artery. Accumulation of cholesterol in atherosclerotic lesions is related to the concentration of cholesterol-carrying lipoproteins in the blood that flows through the coronary arteries. Elevation of the concentration of these lipoproteins is primarily determined by genetic factors but is also influenced by environmental factors, such as a high-fat diet. Environmental and genetic factors may interact adversely. The inverse of this relationship is also true; human beings with low levels of cholesterol, as well as many species of mammals that do not have an excess of cholesterol-carrying lipoproteins, rarely develop atherosclerosis.

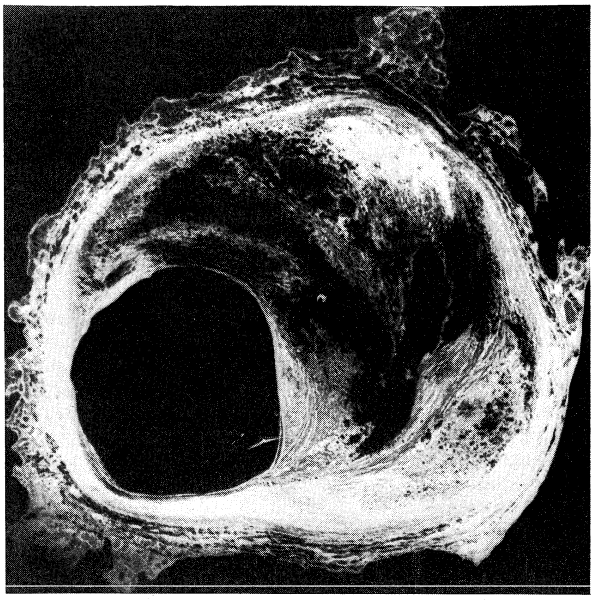


Figure 20: *A typical atheromatous plaque in a coronary artery.* The plaque has reduced the lumen (large dark circle at bottom left) to 30 percent of its normal size. The white areas are lipid and cholesterol deposits. The darker layers represent fibrous areas, probably resulting from earlier incorporation of thrombi from the lumen.

There are alternative views on what actually causes the irreversible accumulation of cholesterol-carrying lipoproteins to form atheromatous lesions. One theory is that the degree of elevation of plasma cholesterol, in association with defective receptors on the lining of the arteries and arterioles, determines whether cholesterol accumulates. The second is that there is a low-grade inflammatory process in this lining (the endothelium) that leads to the formation of microthrombi on the surface and the trapping of cholesterol-carrying lipoproteins on the arterial wall.

It is probable that these two previously divergent theories of the development of atherosclerosis—the accumulation of cholesterol-carrying lipoproteins and endothelial thrombosis (formation of a clot of blood products)—may be combined by accepting that both occur. Under most conditions incorporation of cholesterol-rich lipoproteins is the predominant factor in determining whether or not plaques progressively develop. The endothelial injury that results or may occur independently leads to involvement of two cell types circulating in the blood—platelets and monocytes. Platelets adhere to areas of endothelial injury and to themselves. They trap fibrinogen, a plasma protein, leading to the development of platelet-fibrinogen thrombi. Platelets, monocytes, and other elements of the blood release hormones, called growth factors, that stimulate proliferation of muscle cells in arteries.

Atherosclerotic lesions are focal in nature, and their distribution is determined by the interrelation of hemodynamic physical forces such as blood pressure, blood flow, and turbulence within the lumen. These lead to physical forces of parallel strain, or shear, on the endothelial lining, giving rise to areas of relatively positive and

Distribution of lesions

negative pressure. These hemodynamic forces are particularly important in the system of coronary arteries, where there are unique pressure relationships. The flow of blood through the coronary system into the myocardium takes place during the phase of ventricular relaxation (diastole) and virtually not at all during the phase of ventricular contraction (systole). During systole, the external pressure on coronary arterioles is such that blood cannot flow forward, and the system of coronary arteries is the most extreme example of phasic blood flow in the body. The external pressure exerted by the contracting myocardium on coronary arteries also influences the distribution of atheromatous obstructive lesions.

Arterial atheromatous lesions may undergo reversal, although much of the evidence comes from changes that have been observed in artificial experimental circumstances in primates and early lesions in human femoral arteries. The extent to which it is possible to induce regression of advanced obstructive lesions in human coronary arteries depends upon a better knowledge of the spontaneous changes that occur in these lesions and upon improvements in the methods of their measurement during life.

**Coronary artery disease.** Coronary artery disease is a term describing those diseases, which also occur in other arteries, that lead to obstruction to the flow of blood in the vessel. It is commonly used synonymously with the specific entity of atheromatous intrusion into the artery lumen. Coronary heart disease is a term used to describe the symptoms and features that can result from advanced coronary artery disease. The same symptoms are also called ischemic heart disease because the symptoms result from the development of myocardial ischemia. There is no one-to-one relationship between coronary atherosclerosis and the clinical symptoms of the disease or between coronary artery disease and coronary heart disease.

Coronary artery disease due to atherosclerosis is present to varying degrees in all adults in industrialized communities. Symptoms of disease, however, will occur only when the extent of the lesions or the speed of their development (acute thrombosis) reduces the flow of blood to the myocardium below a critical level. One or more major coronary arteries may progressively narrow without leading to any symptoms of coronary heart disease, provided the area of the heart muscle supplied by that artery is adequately perfused with blood from another coronary artery circuit. The small coronary arteries anastamose (interconnect) and are not, as previously thought, end arteries. Thus, they can open up and provide a collateral, or supportive, circulation that protects against progressive occlusion (obstruction). Exercise improves coronary collateral flow and for this reason may protect against coronary heart disease.

Although coronary artery disease is most frequently caused by atherosclerosis, inflammation of the blood vessels may, in rare cases, cause obstructive lesions of the coronary vessels. In persons with familial hypercholesterolemia, the disease process may involve the mouth of the coronary vessels as they leave the aorta and cause obstruction to blood flow. On rare occasions, clots arising from the left atrium or left ventricle may enter the coronary vessels and cause acute obstruction and symptoms of disease.

Triggers of coronary heart disease

There are influences, or "triggers," that convert coronary artery disease into coronary heart disease; these include coronary thrombosis, vasomotor influences such as coronary spasm, and the hemodynamic needs of the myocardium. Influences within the myocardium itself also may increase the demand for blood flow above the level available, making the myocardium vulnerable to alterations in function, contractility, and the maintenance of normal rhythm. These interrelationships are indicated in Figure 21.

**Coronary heart disease.** Coronary heart disease is a general term for a number of syndromes. Ischemic heart disease, an alternative term, is actually more correct because the syndromes described are all to some degree manifestations of myocardial ischemia (a lack of blood supply to the myocardium).

Coronary heart disease includes a number of interdependent syndromes: angina pectoris, acute myocardial infarction, and sudden cardiac death (due to lethal arrhythmias). There are also features of coronary occlusion (blockage of a coronary artery), including phasic abnormalities during rest and exercise electrocardiograms that indicate the presence of myocardial ischemia. Knowledge of the mechanisms that lead to a particular syndrome is inexact. Thus, a coronary thrombosis may lead to myocardial infarction in one person, sudden death in another, a minor episode of angina in a third, or no symptoms at all in a fourth. There is, however, no alternative to using the orthodox syndromes as the means of recognizing and recording the incidence of coronary heart disease.

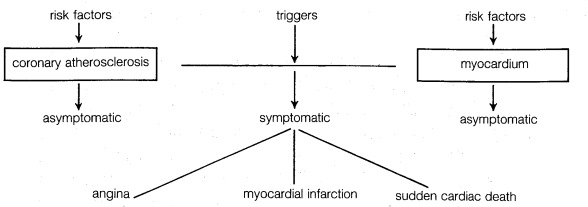



Figure 21: *Alternative interventions for the prevention of coronary heart disease (CHD).*
(Left) Health education to identify risk factors and to alter one's life-style in order to ultimately reduce CHD. (Middle) Identification and alteration of principle triggers, such as thrombosis, that convert atherosclerosis into CHD. (Right) Strengthening the resistance of the myocardium to impending sudden cardiac death by conditioning it to tolerate the effects of progressive arterial occlusion.
From M.F. Oliver, "Prevention of Coronary Heart Disease—Propaganda, Promises, Problems, and Prospects," *Circulation* (January 1986)

*Epidemiology.* There is an uneven geographic distribution of coronary heart disease. It is responsible for one-third to one-half of the deaths in most industrialized countries and is the most common single cause of death in North America, Europe, and the Soviet Union. The disease is relatively uncommon, however, in Asia (including China, Japan, India, and the Middle East), Central Africa, and Central and South America. Many studies have linked the geographic differences in coronary heart disease with various aspects of life-style, such as cigarette smoking, diet, physical inactivity, and obesity.

*Risk factors.* Three main risk factors have been identified: cigarette smoking, a high level of cholesterol in the blood (hypercholesterolemia), and high blood pressure (hypertension). Important as these risk factors are, they are found only in about one-half of those who experience heart attacks. The proportion of persons with any or all of these three risk factors is greater in young and middle-aged adults than in older adults. It is impossible to incriminate any one of these risk factors over another, since the manifestations of coronary heart disease are undoubtedly due to many independent and interdependent influences, but the coexistence of the three greatly increases the risk of developing coronary heart disease.

The familial predisposition to the disease is not well understood, although it is stronger in families with hypercholesterolemia and hypertension. It is most likely to develop prematurely in the presence of polygenic forms, or the rare monogenic form, of familial hypercholesterolemia. There is a progressive relationship between serum cholesterol concentrations and the incidence of coronary heart disease. This is also true for hypertension. Of the three major risk factors, however, excessive cigarette smoking is probably the most important. Other influences—such as a predisposition to develop thrombosis, diabetes mellitus, physical inactivity, obesity, and, rarely, oral contraceptives—may induce premature coronary heart disease in susceptible persons.

Cigarette smoking and heart disease

*Angina pectoris.* The term angina pectoris was first used in 1772 by the British physician William Heberden when he wrote:

> There is a disorder of the breast. . . . The seat of it, and sense of strangling and anxiety, with which it is attended, may make it not improperly be called angina pectoris. Those, who are afflicted with it, are ceased [*sic*] while they are walking and most particularly when they walk soon after eating, with a painful and most disagreeable sensation in the breast, which seems as it would take their breath away, if it were to increase

or to continue; the moment they stand still, all this uneasiness vanishes.

Heberden's initial description is still accurate; however, there are no truly characteristic symptoms of angina pectoris. While the chest discomfort may be variously described as "constricting," "suffocating," "crushing," "heavy," or "squeezing," there are many patients in whom the quality of the sensation is imprecise. The discomfort is usually, but not always, behind the breastbone, but pain radiating to the throat or jaw or down the inner sides of either arm is common. There may be no physical abnormalities, and an electrocardiogram may be normal or show only transient changes with exercise.

Coronary arteriography assesses the extent of coronary artery occlusion, which may vary from a small increase in coronary artery muscle tone at a partly occluded site in a branch of one of the three main coronary arteries, to a 90 percent or greater occlusion of the left main coronary artery with involvement of other major coronary arteries. But the extent of coronary artery disease revealed by coronary arteriography does not predicate action or treatment.

The myocardial ischemia that causes angina is due to a disturbance of the balance between myocardial demands and supply. If demands are reduced sufficiently, the temporarily endangered supply may be adequate. The disturbance of the equilibrium may be short-lived and may correct itself. Unstable angina has an appreciably worse prognosis than stable angina because of a higher risk of myocardial infarction and sudden cardiac death, and it requires daily observation and active intervention.

When coronary arteriography reveals relatively isolated, incompletely obstructive lesions, there are two alternative treatments—medication or coronary angioplasty (balloon dilation of the localized obstruction by a special catheter; see Figure 22). When coronary arteriography reveals a severe occlusion of the left main coronary artery or proximally in one or more of the major arteries, coronary artery bypass graft surgery may be necessary.

In unstable angina pectoris, coronary arteriography may help determine whether coronary angioplasty or coronary artery bypass surgery is needed. Drugs that cause coronary dilation and peripheral arterial vasodilation, and that reduce the load on the heart, are usually necessary. Drugs that reduce the work of the heart by blocking adrenoreceptors (receptors in the heart that respond to epinephrine) and drugs that reduce a thrombogenic tendency are given at this stage. For patients with stable angina, drugs that reduce the heart's work are administered.

*Myocardial infarction.* A syndrome of prolonged, severe chest pain was first described in the medical literature in 1912 by James Brian Herrick, who attributed the syndrome to coronary thrombosis, the development of a clot in a major blood vessel serving the heart. As a result, the disorder was termed coronary thrombosis, or coronary occlusion (blockage of a coronary artery). Later evidence indicated, however, that, though thrombotic occlusion of an atheromatous lesion in a coronary artery is the most common cause of the disorder, the manifestations are the result of the death of an area of heart muscle (infarction). The term myocardial infarction, therefore, is more appropriate. The more general and less specific term heart attack may be more desirable because of these difficulties in describing the causation of the disease entity.

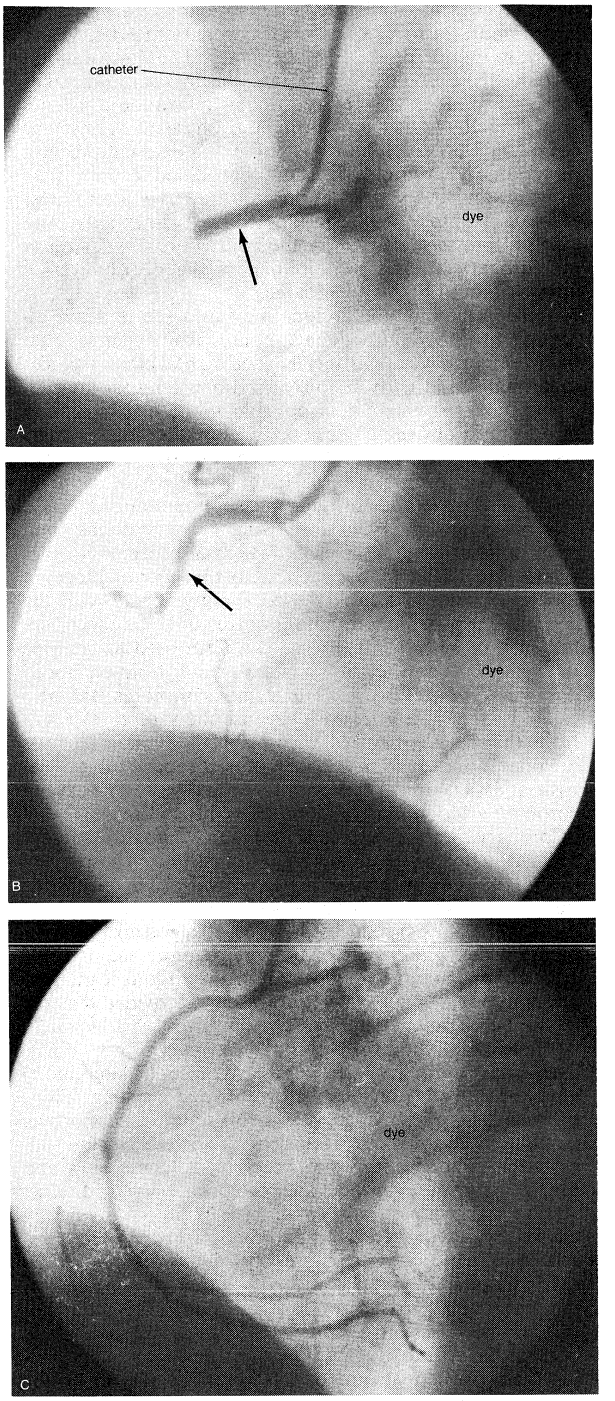



Figure 22: *Successful coronary angioplasty.*
(A) The right coronary artery (arrow) is being injected with radiopaque dye through a catheter in the aorta. The artery is completely blocked by a thrombus one centimetre from its origin. (B) The same right coronary artery (arrow) 30 minutes after the start of intravenous thrombolytic treatment; the clot is beginning to dissolve. (C) The same right coronary artery completely unblocked.

)harac-
eristics
if myo-
ardial
nfarction

Myocardial infarction is characterized by cellular death (necrosis) of a segment of the heart muscle. Generally, it involves an area in the forward wall of the heart related to the blood distribution of the anterior descending coronary artery, though in other instances the inferior wall or the septum (partition) of the ventricle is involved. Coronary thrombosis is present in a majority of the hearts examined at autopsy and undoubtedly plays an important role. In others, changes in metabolic demands of the heart muscle in the presence of a restricted blood flow may be enough to cause death of ischemic cells.

The outstanding clinical feature of myocardial infarction is pain, similar in many respects to that of angina pectoris. The important difference is that the pain lasts for a much longer period, at least half an hour, and usually for several hours and perhaps for days. The pain is described as "crushing," "compressing," "like a vise," and is often associated with some difficulty in breathing. As with angina pectoris, the pain may radiate to the left arm or up the neck into the jaws. There is often nausea, vomiting, and weakness. Fainting (syncope) may occur. The affected person frequently is pale, and he may perspire profusely. Infrequently, these symptoms may be absent, and the occurrence of infarction can then be detected only by laboratory tests. Laboratory studies may show an elevation of the number of white blood cells in the blood or a rise in the enzyme content of the blood, indicating leakage from damaged heart muscle cells. The electrocar-

diogram in most instances shows distinct and characteristic abnormalities at the onset, but the electrocardiographic abnormalities may be less characteristic or totally absent.

Physical signs of infarction

In most persons who experience an acute myocardial infarction, the circulation remains adequate, and only by subtle evidence such as rales (abnormal respiratory sounds) in the lungs or a gallop rhythm of the heartbeat may the evidence of some minor degree of heart failure be detected. In a small percentage of cases, the state of shock occurs, with pallor, coolness of the hands and feet, low blood pressure, and rapid heart action. In these cases, myocardial infarction is deadly, with survival rates of only 20–40 percent, in contrast to the current survival rate in many hospitals of 80–90 percent. Mortality is also related to age, for the process is more lethal in the elderly. In a small number of persons there may be thromboembolism (obstruction caused by a clot that has broken loose from its site of formation) into an artery elsewhere in the body.

In some individuals, the damage caused by the infarction may interfere with the functioning of the mitral valve, the valve between the left upper and lower chambers, and result in a form of valvular heart disease. It may cause a rupture of the interventricular septum, the partition between the left and right ventricles, with the development of a ventricular septal defect, such as is seen in some forms of congenital heart disease. Rupture of the ventricle also may occur.

Recovery

Drugs are used to control arrhythmias and to strengthen myocardial function. Convalescence from an acute myocardial infarction may last several weeks, allowing time for scar tissue to form in the area of an infarction and a gradual return to activity. Although some persons may have residual evidence of heart failure or other cardiac malfunction, most individuals may return to an active life after a period of weeks and are not in any way invalided by the process. These individuals do, however, have an increased potential for subsequent myocardial infarction.

*Sudden death.* The term sudden death is used imprecisely and includes death that is almost instantaneous as well as death in which rapidly deteriorating disease processes may occupy as much as two or three days. In heart disease both may occur, but the term characteristically refers to instantaneous death, which is frequent in coronary heart disease. Sudden death from coronary heart disease occurs so frequently that less than half of the persons who die from heart attacks each year in the United States survive long enough to reach the hospital.

Instantaneous cardiac death is usually due to ventricular fibrillation (an uncontrolled and uncoordinated twitching of the ventricle muscle), with total mechanical inadequacy of the heart and erratic and ineffective electrical activity. Sudden death may occur without any previous manifestations of coronary heart disease. It may occur in the course of angina pectoris and causes about one-half of the deaths due to acute myocardial infarction in hospitalized patients, though this number is decreasing with the more widespread use of coronary-care units. Although a reduced supply of blood to the heart undoubtedly is the precipitating factor, acute myocardial infarction does not always occur. In most persons who have died almost instantaneously, no infarction was present, but there was widespread coronary artery disease. In rare instances, sudden death occurs without a major degree of coronary artery disease.

The use of closed-chest cardiopulmonary massage (massage of heart and lungs without making an opening in the chest wall), coupled with electrical defibrillation (the use of electrical shocks), if applied within a few minutes of the sudden death episode, may successfully resuscitate the majority of patients. In coronary-care units where the facilities and trained personnel are immediately available, the percentage of successful resuscitations is high. In general hospitals where resuscitation teams have been established, the percentage is less satisfactory. Sudden death outside the hospital is, of course, a more difficult problem, but mobile coronary-care units responding as emergency ambulances to the site of a sudden death are being used widely and successfully. Effective resuscitation depends upon the prompt arrival of the unit. The use of drugs and other means to prevent the onset of sudden death has been relatively successful in the coronary-care unit, except in situations in which the disease has been present for a long period of time.

*Survival during and after a heart attack.* The risk of death from an arrhythmia is greatest within the first few minutes of the onset of an occlusion in a coronary artery or of acute ischemia occurring in the region of the myocardium. Thus, of those likely to die during the first two weeks after a major heart attack, 40 percent will die within one hour of the onset and another 20 percent within the next three hours.

Ventricular fibrillation

During the first few hours most persons have some disturbances of rhythm and conduction. Ventricular fibrillation is particularly common in the first two hours, and its incidence decreases rapidly during the next 10 to 12 hours. If undetected, ventricular fibrillation is lethal. It can be reversed, however, in 80 to 90 percent of patients with the use of appropriate electronic devices for monitoring heart rhythm, for giving a direct-current shock to stop it, and for resuscitation.

Both the immediate and long-term outlook of persons after myocardial infarction depends on the extent of myocardial damage and the influence of this damage on cardiac function. Efforts to limit or reduce the size of the infarct have been unsuccessful in improving the short- or long-term outlook. Procedures that cause thrombi to dissolve (thrombolysis), however, have led to the dramatic and immediate opening of apparently occluded coronary arteries. When such measures are implemented within four hours (and preferably one hour) of the onset of the heart attack, the chances of survival are greater, although it has not been shown that long-term prognosis is improved. Naturally occurring lytic enzymes (such as streptokinase) or genetically engineered products are used.

*Coronary artery bypass surgery.* Coronary artery bypass surgery is widely used to restore adequate blood flow to the heart muscle beyond severe atheromatous obstruction in the main coronary arteries. The most common operation is one in which lengths of superficial veins are taken from the legs and inserted between the aorta, usually above the sinuses of Valsalva, and joined to a part of a coronary artery below the obstructive atheromatous lesion. Multiple grafts are often used for multiple atheromatous occlusions. The internal mammary arteries are also used to provide a new blood supply beyond the point of arterial obstruction; however, since there are only two internal mammary arteries, their use is limited.

There are two principal uses for coronary artery bypass surgery. One is to relieve chest angina that is resistant to medication. The other is to prolong a person's life; however, this is only achieved when all three main coronary arteries are severely obstructed and when the contractility of the left ventricle has been impaired somewhat. Coronary artery bypass surgery does not prolong life when it is used to overcome an obstruction in only one or even two arteries.

*Prevention of coronary heart disease.* Most physicians recommend the elimination of cigarette smoking, a reduction of dietary saturated fat, a proportional increase in fat calories from polyunsaturated sources, a reduction in high blood pressure, an increase in physical activity, and the maintenance of a weight within normal limits. While the circumstantial evidence from many kinds of studies supporting these measures is impressive, they have not yet been shown to be as effective as expected or predicted. While claims are made that a decrease in the rate of death from heart disease in the United States, Australia, Finland, and other countries is due to such changes in life style, similar changes in life style elsewhere have not been associated with a decrease in the death rate. In some countries (such as Sweden) coronary mortality has actually risen. Those persons, however, who come from high-risk families, such as those with familial hypercholesterolemia, have benefited from reduction of high levels of serum cholesterol. Reduction of high blood pressure, on the other hand, has not been shown to lower coronary mortality.

**Rheumatic heart disease.** Rheumatic heart disease results from inflammation of the heart lining, heart mus-

cle, and pericardium occurring in the course of acute rheumatic fever, an infection with *Streptococcus pyogenes* organisms. The disease includes those later developments that persist after the acute process has subsided and may result in damage to a valve, which may in turn lead to heart failure.

Rheumatic fever is poorly understood. The disease process occurs days or weeks following the initial streptococcal infection. Later infections may bring about recurrences of the rheumatic fever that damage the heart. Immunologic processes (reactions to a foreign protein) are thought to be responsible for the response that damages the heart and particularly the heart valves. Rapid and effective treatment or prevention of streptococcal infections stops the acute process.

Many other factors of a geographic, economic, and climatic nature influence the incidence of rheumatic fever but are not the primary causes. Rheumatic fever is becoming less common in the second half of the 20th century, and with better control of streptococcal infections, there is an indication of a sharp decline in rheumatic heart disease.

'athologic hanges

It is thought that the basic pathologic lesion involves inflammatory changes in the collagen, the main supportive protein of the connective tissue. There is also inflammation of the endocardium and the pericardium. Only a relatively small percentage of deaths occur in the acute phase, with evidence of overwhelming inflammation associated with acute heart failure. There may be a disturbance of the conduction system of the heart and involvement of other tissues of the body, particularly the joints. About one-half of the persons found to have late rheumatic valvular disease give some indication that they have had acute rheumatic fever.

Major effect of heumatic ever

The major toll of rheumatic fever is in the deformity of the heart valves created by the initial attack or by frequently repeated attacks of the acute illness. Although there may be valve involvement in the acute stages, it usually requires several years before valve defects become manifest as the cause of heart malfunction. The valve most frequently affected is the mitral valve, less commonly the aortic valve, and least common of all, the tricuspid valve. The lesion may cause either insufficiency of the valve, preventing it from operating in a normal fashion and leading to regurgitation, or stenosis (narrowing) of the valve, preventing a normal flow of blood and adding to the burden of the heart.

Mitral valve involvement is usually symptomless initially but may lead to left ventricular failure, with shortness of breath. Heart murmurs are reasonably accurate signposts for specific valvular diagnoses. A murmur during the diastolic, or resting, phase of the heart, when blood normally flows through the mitral valve to fill the ventricle, generally indicates the presence of mitral stenosis. On the other hand, a murmur during the systole, or contraction, of the left ventricle, indicates an abnormal flow of blood back through the mitral valve into the left atrium (mitral regurgitation). When this latter condition is present, each beat of the heart must pump enough blood to supply the body as well as the wasted reflux into the pulmonary vascular system. This additional work load causes dilation and enlargement of the ventricle and leads to the development of congestive heart failure.

Involvement of the aortic valve is common, and again there may be evidence of stenosis or insufficiency. The presence of aortic stenosis may lead to a marked hypertrophy (enlargement) of the left ventricle of the heart. Involvement of either the tricuspid or pulmonic valve occurs in a similar fashion. In many persons with rheumatic valvular disease, more than one valve is involved. The specific type of valve involved influences the clinical picture of congestive failure (see below *Heart failure*).

**The heart, the pulmonary artery, and the aorta.** *Pulmonary heart disease (cor pulmonale).* An obstruction to blood flow through the network of vessels in the lungs develops in various types of lung disease. Impingement on blood flow in this integral part of the total path of blood flow through the body places a burden on the right side of the heart, which normally pumps against a low-pressure load with little resistance to blood flow. Pulmonary-artery pressures are normally low compared with those in the aorta.

Acute and chronic

Pulmonary heart disease may be divided into acute and chronic forms. The classic form of acute pulmonary heart disease (acute cor pulmonale) occurs when there is sudden obstruction to the pulmonary blood-flow pattern, as occurs with a massive embolus—a blood clot that has broken loose from its point of formation. This impairs blood flow through the lungs, causes additional reflex changes that add to the heart's burden, and creates an acute form of high blood pressure in the pulmonary arteries, with dilation and failure of the right ventricle. The right ventricle's pumping ability is acutely depressed, and, therefore, the amount of blood available for the left side of the heart is also restricted, so that systemic circulatory failure occurs.

Respiratory symptoms are not prominent, and the disorder in its early stages is not accompanied by edema in the lung. The clinical picture in the more severe form is one of shock, with cold, pale, and clammy skin, low arterial pressure, and high pulse rate. Oxygen transfer in the lungs is severely impaired, and the heart may be acutely dilated. Treatment is with anticoagulant or thrombolytic drugs (such as streptokinase) and oxygen, which relieve the hypoxia (low serum oxygen levels), or in some instances, surgical removal of the obstruction.

Chronic cor pulmonale may be caused by either a form of pulmonary disease, such as chronic bronchitis and emphysema—in which lung tissue is destroyed and replaced with air spaces, causing a loss of pulmonary blood vessels—multiple blood clots in the vessels of the lung, or a primary disorder of the pulmonary blood vessels. The result is a form of heart failure partly based on an obstruction to flow through the pulmonary blood vessels, producing pulmonary arterial hypertension. Cyanosis may be evident, indicating that the arterial blood is not saturated with oxygen. In patients with chronic bronchitis and emphysema, the lack of oxygen or desaturation contributes to pulmonary hypertension. The manifestations of heart failure are present, particularly where there is edema, except that shortness of breath is often due to the underlying lung disease. The right side of the heart is enlarged, the valve sounds from the pulmonic valve may be loud, and there may be electrocardiographic evidence of chronic strain on the right side of the heart. Drugs that dilate the pulmonary blood vessels or relieve the edema and drugs with anticoagulant effects can be useful in the treatment of chronic pulmonary heart disease. However, the course that affords the best chance of remission of symptoms in patients with cor pulmonale due to chronic bronchitis and emphysema includes prompt treatment of infection, termination of smoking, and correction of the lack of oxygen.

*Hypertensive heart disease.* Arterial hypertension is a disease in which the regulation of blood pressure is abnormal, resulting in arterial pressure that is chronically higher than normal. Hypertension results from several causes, but the cause of the most common form (essential hypertension) is not understood. A family tendency to hypertension has been found in persons with the disease, and there may be a basic genetic abnormality involving altered cell membrane permeability in the blood vessels. This defect might make such persons less able to handle salt and in turn more responsive to hormonal or nervous stimulation.

Causes of hypertension

Excessive dietary intake of salt has long been held to be responsible for hypertension in certain groups, while other groups with a low sodium intake do not show an increase in blood pressure with age. Stress has also been shown to cause hypertension, and fear and anxiety can induce a rise in blood pressure owing to increased activity in the sympathetic nervous system. Other hormones and vasoactive substances have a direct effect on blood pressure, but the interaction of these factors remains unclear. Hypertension also results from a number of types of chronic renal (kidney) diseases and some tumours of the adrenal gland. In certain structural abnormalities of the aorta, such as coarctation, in which the artery's middle coat is deformed with resulting narrowing of the channel, arterial pressure in the upper half of the body is abnormally high.

Regardless of the cause, but in some ways coloured by it, the effects on the cardiovascular system are similar. The impact on the vascular system varies from person to person. In some persons, for unknown reasons, the body withstands the abnormal elevation of blood pressure with minimum change in the heart and blood vessels. In other persons, blood vessel damage is early and severe, coupled with serious deterioration of heart function. In general, the rule is that the higher the blood pressure, the higher the degree of cardiovascular damage, though there are many exceptions. Rarely, a vicious and damaging form of hypertension occurs, often called malignant hypertension, which results in small blood vessel damage throughout the body, but particularly affecting the heart, brain, and kidneys.

Persons with hypertensive disease have an increased susceptibility to atherosclerosis of the coronary arteries, thus making it difficult to separate the cardiac manifestations from those actually caused by hypertension. Hypertensive persons, therefore, may eventually have congestive heart failure following enlargement of the heart caused by chronic increase in arterial pressure. In addition, they may suffer the effects of a decline in blood supply to the heart because of coronary artery disease and the classic manifestations of coronary arteriosclerosis, such as angina pectoris or death of a portion of heart muscle (myocardial infarction). Hypertensive cardiovascular disease may also become manifest through defects in the vessels supplying the brain, leading to stroke or other abnormalities of the brain. Furthermore, hypertensive cardiovascular manifestations may be complicated by the development of kidney failure and resultant abnormal retention of fluid in the tissues, adding to the problems of congestive heart failure.

Before the use of antihypertensive drugs, high blood pressure was associated with a greatly increased mortality, with survival measured in months in the most severe cases. Antihypertensive drugs have dramatically increased the life expectancy of patients with severe hypertension, and heart failure, stroke, and kidney failure are now relatively uncommon in treated hypertensive patients. A similar reduction in coronary heart disease among this group of patients, however, has not occurred. Other factors, such as smoking and diet, may be important in this aspect of therapy. The treatment of mild hypertension is more controversial; although some patients benefit when actively treated as compared with those given a placebo.

*Other diseases of aorta and pulmonary arteries.* Arteriosclerosis may involve the aorta and its major branches. Indeed, it seems to be an almost inevitable process with increasing age, but the rate of development and the extent of involvement vary greatly. The process may merely limit the elasticity of the aorta and allow for some dilation and tortuosity as age advances. In more severe instances, there may be a major degree of dilation or localized formation of aneurysms (bulging of the vessel wall at a point of weakness), generally in the abdominal portion of the aorta. These aneurysms may result in pain and may occasionally rupture, causing sudden death. The arteriosclerotic process may impair the flow of blood to the tributaries of the aorta and lead to a variety of ischemic states—*i.e.,* result in various types of damage that come from an insufficient supply of blood. This condition is particularly notable when the renal vessels are involved, creating a state of renal ischemia, occasionally creating hypertension, and possibly terminating in renal failure.

Aneurysms of the aorta

Medial necrosis is a lesion of the aorta in which the media, the middle coat of the artery, deteriorates and, in association with arteriosclerosis and often hypertension, may lead to a dissecting aneurysm. In a dissecting aneurysm a rupture in the intima, the innermost coat of the artery, permits blood to enter the wall of the aorta, causing separation of the layers of the wall. Obstruction to tributaries may occur, which is usually associated with severe chest pain. In many instances there is a secondary rupture of the exterior wall, which may lead to fatal internal bleeding. Inflammation of the aortic wall may occur as an isolated process.

Deposition of calcium salts in the aorta wall may occur as a part of the arteriosclerotic process or of other disease involvement. In certain conditions, such as congenital heart disease, blood clots (thrombi) may form in the pulmonary artery, and these may break loose. Blood clots in the lungs (pulmonary emboli) may arise from this and other sources in the systemic venous circulation. These fragments of clot may be small, causing no detectable manifestations, or large, causing obstruction of either the total pulmonary arterial flow or of flow to an area of lung.

*Syphilis of the heart and aorta.* Syphilis, a disease caused by infection with a microorganism, *Treponema pallidum,* in its early phases becomes widespread in the human body. At this time there may be transient inflammation of the heart muscle, but usually with little or no impairment of the circulation. In the late stages of the disease there may be syphilitic involvement of the heart, confined almost purely to the aorta and aortic valve. A particularly severe form of aortic insufficiency may develop, with subsequent dilation and enlargement of the heart and, eventually, heart failure. The disease process often involves the base of the aorta and the blood flow through the openings into the coronary vessels from the aorta, causing impairment of the coronary circulation, with resultant angina pectoris and, on rare occasions, myocardial infarction, the death of portions of heart muscle.

Areas of involvement

The syphilitic process may also involve the wall of the aorta, resulting in the loss of the elastic properties, in dilation, and, at times, in the formation of aneurysms of the aorta. The aneurysms may become large and interfere with blood flow through the tributaries of the aorta in the involved area. They may be the source of pain and eventually may rupture, causing sudden death from loss of blood into the thoracic cavity. Syphilis of the aorta was common in the past, but with the advent of more modern control mechanisms, plus effective early treatment with the use of penicillin, the disorder has become much less common. Late complications can be effectively avoided by early antisyphilitic treatment.

**Diseases of the endocardium and valves.** Bacterial endocarditis, a disease in which bacterial or fungus infection becomes established on the surface of a heart valve or, less commonly, in a blood vessel wall or in the endocardium of the heart, usually occurs where there has been some previous lesion, either congenital or acquired. Most frequently, the location is at the line of closure of the valve. The disease may be acute and severe, or it may be a more chronic situation, often referred to as subacute bacterial endocarditis. It may erode the valve structure, or it may be of an inflammatory nature, producing nodules with the ulcerative surface of active infection. Because the bacteria are embedded in the lesion, the normal body defenses contained in the blood have difficulty entering into play; for this reason, certain types of bacterial endocarditis become more chronic and more slowly progressive. The effects of the lesion are complex, being related to the presence of a bacterial infection in the body, local damage to the valve, and systemic damage caused by fragments of a blood clot that breaks off and travels through the bloodstream to distant organs. These clots cause infarctions or abscesses, a type of kidney disease, or other small areas of bleeding and necrosis in the skin, eyes, and other parts of the body.

Endocarditis

Before the advent of antibiotic therapy, bacterial endocarditis was almost always a fatal disease. Many affected persons can now be successfully treated, given the best conditions, though the mortality rate still remains relatively high. Inflammation of the heart lining, endocarditis not caused by infection, may occur in some illnesses, but it does not result in formation and breaking loose of blood clots.

In the course of rheumatoid arthritis, a chronic inflammation of the joints of unknown cause, a type of valvular damage has been recognized. It is different from that caused by rheumatic fever but leads to valvular insufficiency and stenosis (narrowing) in much the same fashion and is particularly likely to attack the aortic valve. The tendencies toward heart failure and toward impairment of heart function are the same as in rheumatic valvular disease.

**Diseases of the myocardium.** There has been increasing recognition of a type of heart disease characterized as primary myocardial disease. The cardiomyopathies are

diseases involving the heart muscle itself. They are unique in that they are not the result of hypertensive, congenital, valvular, or pericardial diseases and are rarely the result of ischemic heart disease. This form of heart disease is often sufficiently distinctive, both in general symptoms and in patterns of blood flow, to allow a diagnosis to be made. Increasing awareness of the condition, along with improved diagnostic techniques, has shown that cardiomyopathy is a major cause of morbidity and mortality. In some areas of the world it may account for as many as 30 percent of all deaths due to heart disease.

Some cardiomyopathies are primary because the basic disease process involves the myocardium rather than other cardiac structures and because the cause of the disease is not known and not part of a disorder of other organs. Other cardiomyopathies are conditions in which the cause of the myocardial abnormality is known and the cardiomyopathy is a manifestation of a systemic disease process. Clinically, the cardiomyopathies fall into three categories: dilated cardiomyopathy, characterized by ventricular dilation and often symptoms of congestive heart failure; hypertrophic cardiomyopathy, characterized by hypertrophy of the ventricle, particularly the left ventricle; and restrictive cardiomyopathy, marked by scarring of the ventricle and impairment of filling in diastole.

A large number of cardiomyopathies are apparently not related to an infectious process but are not well understood. A number of these are congenital and many cause enlargement of the heart. About one-third of these diseases are familial, and some of these are transmitted as a non-sex-linked autosomal dominant trait (*i.e.,* a person may be affected if he inherits the tendency from one parent). They are particularly common among black populations. A number of metabolic diseases associated with endocrine disorders may also cause cardiomyopathies. Other metabolic disorders that may contribute to cardiomyopathy include beriberi, caused by a nutritional deficiency; cobalt poisoning in heavy beer drinkers; or a form of cardiomyopathy in chronic alcoholics. There are also rare cardiomyopathies caused by drugs. Infections, such as acute rheumatic fever and several viral infections, may cause any of a number of types of myocarditis. Myocarditis may also occur as a manifestation of a generalized hypersensitivity (allergic or immunologic) reaction throughout the body.

The cardiomyopathies may cause no symptoms and may be detected only by evidence of an enlarged heart and disturbances in cardiac conduction mechanisms detected on electrocardiography. In other instances, extensive involvement may lead to heart failure. Some cases may be chronic, with exacerbations and remissions over a period of years.

The collagen diseases

The heart may be affected by any of a considerable number of collagen diseases. Collagen is the principal connective-tissue protein, and collagen diseases are diseases of the connective tissues. They include diseases primarily of the joints (*e.g.,* rheumatoid arthritis); primarily of the skin (*e.g.,* scleroderma); and systemic disease (*e.g.,* systemic lupus erythematosus).

**Diseases of the pericardium.** Pericardial disease may occur as an isolated process or as a subordinate and unsuspected manifestation of a disease elsewhere in the body. Acute pericarditis—inflammation of the pericardium—may result from invasion of the pericardium by one of a number of agents (viral, fungal, protozoal); as a manifestation of certain connective tissue and allergic diseases; or as a result of chemical or metabolic disturbances. Cancer and specific injury to the pericardium are also potential causes of pericardial disease.

Pain is the most common symptom in acute pericarditis, though pericarditis may occur without pain. A characteristic sound, called friction rub, and characteristic electrocardiographic findings are factors in diagnosis. Acute pericarditis may be accompanied by an outpouring of fluid into the pericardial sac. The presence of pericardial fluid in excessive amounts may enlarge the silhouette of the heart in X rays but not impair its function. If the pericardial fluid accumulates rapidly or in great amounts, if there is a hemorrhage into the sac, or if the pericardium is diseased so that it does not expand, the heart is compressed, a state called cardiac tamponade. There is interference with the heart's ability to fill with blood and reduction of cardiac output. In its more severe form, cardiac tamponade causes a shocklike state that may be lethal. Removal of the fluid is lifesaving in an emergency and aids in the identification of the cause.

Cardiac tamponade

Chronic constrictive pericarditis, caused by scar tissue in the pericardium, restricts the activity of the ventricles. In many instances, the cause is not known, but in some it is the result of tuberculosis and other specific infections. It is treated most effectively by surgery. Tumours that either arise directly from the pericardium or are secondary growths from other sources may impair cardiac function and cause pericardial effusion (escape of fluid into the pericardium).

### DISTURBANCES IN RHYTHM AND CONDUCTION

The heart's rhythmical beat is initiated and regulated from centres within the organ. The primary pacemaker, the sinoatrial node, is a small mass of specialized muscle cells located at the juncture of the upper vena cava and the right atrium. Electrical impulses are emitted by this group of cells. The excitation spreads through the two atria and, by way of a band of fibres called the bundle of His, into the ventricles. The bundle of His has its beginning in a small mass of cells, the atrioventricular node, located beneath the lining of the right atrium.

Normally initiated heart rhythm, originating in the sinoatrial node, is called sinus rhythm. Under stimulation from the central nervous system and other metabolic factors, heart rate may normally rise and fall, with a slight variation, in part related to respiratory activity. In young individuals in excellent physical condition, the resting heart rate may fall as low as 40 to 50 beats per minute, and, under stressful psychological stimulation, the heart rate may rise to as high as 200 beats per minute. These situations are to be differentiated from pathological variations in heart rate. Abnormal slowing of heart rate, or sinus bradycardia (a slow sinus rhythm with a rate below 60, caused by disturbance of the sinoatrial node) or acceleration of heart rate, or tachycardia (excessive rapidity in the action of the heart with a pulse rate of above 100 beats per minute) may occur in a wide variety of disease states and be symptomatic of the underlying disease.

Extra beats arising from the atrium, the nodal tissues, or the ventricle are not in themselves abnormal, though beats arising from the ventricle are more often associated with organic heart disease. Occasional extra systoles (contractions) occur in many normal individuals. In cardiovascular disease they are much more common. They do not interfere with normal cardiovascular function if infrequent. It has been noted that continued psychological stress, excessive smoking, and drinking of large amounts of tea and coffee enhance the tendencies for premature contractions of the atria. Premature contractions of the ventricles are more ominous, especially those that occur after exercise. They have been found to be associated with coronary artery disease in a large percentage of instances. If frequent enough, they may be the harbinger of more serious ventricular arrhythmias.

**Atrial arrhythmias.** Abnormalities in the rhythm of atrial contractions include atrial tachycardia, atrial flutter, and atrial fibrillation. As noted earlier, atrial tachycardia may be a manifestation of underlying disease such as thyrotoxicosis or may be merely a matter of stressful stimulation in a normal individual. Some individuals have characteristic episodes of paroxysmal atrial tachycardia of varying frequency and duration, with a rapid onset and termination. Ordinarily the episodes occur in the absence of any other heart abnormality. Occasionally, especially when the episodes are prolonged or when they occur in the presence of organic cardiovascular disease, they may be accompanied by evidence of heart failure or, in rare instances, of the shock state.

Atrial flutter and fibrillation

Atrial flutter represents another form of arrhythmia associated with very rapid atrial activity. The atrial activity is regular but so rapid that the conduction of the impulses to the ventricle may be delayed and impaired so that

only one of two, three, or four impulses excite ventricular activity. This disorder is most often, but not always, seen in persons with organic heart disease.

Atrial fibrillation is another form of arrhythmia, in which there is wildly erratic and ineffective atrial contraction. The ventricular response is also erratic, so that the pulse is irregular without a basic underlying rhythm. This disorder is seen most frequently in persons with organic heart disease, such as rheumatic heart disease with mitral involvement and thyrotoxicosis, but also occurs in normal individuals, often on a paroxysmal basis. It renders the atrium ineffective, and therefore the contribution of this chamber to the normal pumping of blood is negated. This condition may be of no real functional importance in the normal heart but may have a significant detrimental effect in the failing heart. The circulation functions reasonably well, though in the case of severe mitral stenosis a rapid heart action is not well tolerated because of the limited speed of ventricular filling through a small mitral opening.

A complication of atrial fibrillation is the development of blood clots within the walls of the fibrillating atria. Because these clots may eventually fragment and pass into the circulation (arterial embolism), atrial fibrillation is a condition that should be treated if possible. The abnormal rhythm may be slowed by digitalis or terminated by the use of electrical shocks (electrical defibrillation).

**Atrioventricular node mechanisms.** Normally, the impulse for ventricular contraction arises from a focus in the atria, passes through the atrioventricular node and bundle of His, which delay progression, and then excites ventricular activity. This mechanism may not operate properly in various pathological states. In some instances, more rapidly firing (*i.e.,* impulse-emitting) tissue in the atrioventricular node may initiate the ventricular activity, supplanting the normal atrial focus and resulting in tachycardia, called nodal tachycardia because of the part played by the atrioventricular node. More frequently, the abnormal mechanism is one of delay or obstruction in progression of the impulse, creating "heart block," a lack of synchronization between atria and ventricles; there may be varying degrees of severity. In first-degree heart block, there is merely a short delay over the normal conduction time from the atrium to the ventricle. This condition does not result in clinical manifestation but is detected electrocardiographically by a longer than normal period of time between the P wave, which is associated with atrial activity, and the QRS complex, which is associated with ventricular excitation.

Three degrees of heart block

In a more severe (second-degree) form, the impulse may travel relatively normally through the conduction system, but some beats are blocked. Alternate beats may be blocked, resulting in what is known as a two-to-one heart block, or the ratio may be three-to-one, four-to-one, etc. The degree of blockage of conduction may be influenced by stimulation from outside the heart, such as from the carotid sinus or the central nervous system, and is dependent in part on drug action. (The carotid arteries are the principal arteries to the head. The carotid sinus, located at the fork where the common carotid divides into the external and internal carotid arteries, is an important structure for monitoring and regulating blood pressure.)

The conduction system below the atrioventricular node divides into two major branches, to the two ventricles, and, therefore, blockage of impulse transmission in either of these is termed right or left bundle branch block. This condition is identified by electrocardiographic changes and is most frequently associated with organic heart disease, though it may occur rarely in normal individuals.

The most severe form of heart block, complete heart block, is characterized by total dissociation of atrial activity and ventricular activity. In this situation, atrial activity usually continues at a normal or higher than normal rate. Independently the ventricle establishes its own rate, usually slower than normal, but in some instances within the normal range. If the latter is the case, the difficulty is without major consequence on the circulation. More frequently, however, it is at a slower rate, and there may be some resultant inadequacy of the circulation. This condition is usually associated with forms of organic heart disease, and the slowness of the rate may precipitate congestive heart failure or other manifestations. It may be treated by the use of drugs or electrical pacemakers to increase the rate of ventricular activity.

At times, the degree of heart block or its occurrence may be variable and erratic. This condition may result in considerable pauses in left-ventricular depolarization and contraction, the basic mechanism that precipitates the Adams-Stokes syndrome, which leads to occasional episodes of unconsciousness. (The syndrome is named for the Irish physicians Robert Adams and William Stokes, whose descriptions of the disease are celebrated.)

**Ventricular arrhythmias.** The ventricle may respond regularly or erratically to atrial or nodal (atrioventricular) disturbances of rhythm. In addition, it is subject to other intrinsic forms of abnormal rhythm. In the course of severe heart disease, such as coronary artery disease, ventricular tachycardia (fast beating of the ventricles) may occur. The beat is regular but may be so rapid that it interferes with normal cardiac filling and ejection and, therefore, results in either congestive failure, if prolonged, or in the development of a shock state, if severe and acute. Ventricular tachycardia is perhaps most important because it may be the forerunner of ventricular fibrillation, in which, as in atrial fibrillation, the contractions are widely erratic and ineffective, so that ventricular fibrillation interferes so much with ventricular function that circulation of the blood effectively ceases. Unless reversed within seconds or minutes, it is lethal. The recognition of ventricular tachycardia and fibrillation has been the basis for the success of modern coronary-care units. It is treatable either by drugs or, more frequently, by the use of external electrical defibrillation, the application of electrical shocks.

Ventricular tachycardia and fibrillation

Some of the major recognized forms of arrhythmia have been presented. There are many special variants on this pattern, but even with the most complex electrocardiographic studies, some arrhythmias defy explanation.

## HEART FAILURE

Congestive heart failure, a syndrome resulting from disease that has caused the heart to be inadequate as a pump, is characterized by manifestations distant from the heart, predominantly related to salt and water retention in the tissues. It may vary from the most minimal symptoms to sudden pulmonary edema (abnormal accumulation of fluid in the lungs) or to a rapidly lethal shocklike state. Chronic states of varying severity may last years. The symptoms are related predominantly to secondary manifestations resulting from retention of fluid and vascular congestion throughout the body, rather than to the direct effect of lessened blood flow. In most instances, failure results from a diseased heart, though on occasion the burden of other systemic diseases may exceed the capacity of the previously adequate heart and produce the condition.

A physiological characteristic of the normal heart is its ability to meet the fluctuating demands of the body for blood flow. The diseased heart may no longer be able to respond in this way. This failure may be the result of severe and acute damage, such as an acute myocardial infarction, or of chronic and lesser impairment, such as scarring of a valve in the course of a long-standing rheumatic heart disease.

The possible causes of the underlying heart disease are numerous. One is coronary heart disease that has resulted in profound myocardial damage. Other common causes are rheumatic valvular disease, hypertensive vascular disease with involvement of the heart, or one of the multitudinous but less common types of primary disease of the myocardium. Regardless of the cause, the common denominator leading to heart failure is altered function of the heart muscle, with lessened ability to pump blood. In many sophisticated studies, clear-cut disturbances in the mechanics of heart muscle contraction have been demonstrated.

The presence of cardiac failure becomes apparent largely by signs and symptoms not directly related to the heart. With early and untreated heart failure, the person usually has a normal salt and water intake, but his ability to regulate and promptly excrete excess sodium and water is

impaired. The nature of this impairment is not entirely clear. Salt and water accumulate in the body as extracellular fluid that manifests itself as clinically detectable edema. If the person is active and in the upright position, the fluid may gather particularly about the ankles and legs. If he is in bed, it may accumulate in the back, overlying the sacrum. In heart failure, there may also be low blood sodium levels.

**Left ventricular failure.** The pulmonary circuit (the blood vessels in the lungs) usually becomes congested in heart failure, because heart diseases most frequently affect the left ventricle in one way or another. If disease impedes the left ventricle's pumping of blood into the systemic circulation, the left side of the heart is unable to receive the normal flow of oxygenated blood from the lungs; consequently there is back pressure, and blood accumulates in the lung's blood vessels. If this congestion of the pulmonary vessels occurs, it lessens the amount of space available in the lungs for air and tends to stiffen the lungs. Finally, pulmonary capillary pressure may reach the point at which fluid flows into the tissues outside the vessels, a condition called pulmonary edema. These phenomena account for the frequency of difficulty in breathing, inability to breathe except in an upright position, attacks of respiratory distress without apparent cause during sleep at night (paroxysmal nocturnal dyspnea), and other respiratory symptoms in congestive heart failure.

Effect on the lungs

**Pulmonary edema.** Acute pulmonary edema developing in the course of heart failure may be severe, occasionally even fatal. It occurs particularly in such situations as arterial hypertension and aortic valve disease. Peripheral edema may or may not be present. This combination of events is often called left heart failure.

**Right heart failure.** In heart disease affecting primarily the right side of the heart, such as cor pulmonale or disease of the tricuspid valve, between the right atrium and right ventricle, the effect on respiration may be less, and there is greater evidence of edema, congestion of the liver, and high pressure in the veins. Although actually the circulation as a whole is failing, those factors that produce manifestations in the lungs are not prominent.

**Incipient heart failure.** A person affected rarely speaks of symptoms that can be directly related to the inadequacy of cardiac output. In severe and chronic heart failure, there may be a general deterioration of the body with a form of general ill health and malnutrition that simulates that seen in cancer. There is rarely evidence that blood flow is insufficient to provide adequate tissue function except in the shock state that occurs in acute and severe heart failure, in which there is acute inadequacy of flow to some critical tissues. Although heart failure may occur in certain circumstances, such as severe anemia, in which there is a high cardiac output, in general, heart failure occurs when the cardiac output is reduced from the normal value for the particular person. The inadequacy of the heart may not be constant but may occur, especially in early heart failure, only under the stress of exercise.

Persons in the early stages of heart failure may have no circulatory impairment, though the heart may be dilated or enlarged. Such persons may eventually be unable to respond to stressful situations that demand a high cardiac output. Difficulty in breathing may occur during exertion, but usually edema is not present.

**Mild to moderate congestive heart failure.** The person with mild to moderate congestive heart failure may have a heart that has an adequate pumping function for the demands at rest but that is unable to meet circulatory needs under stress. Such a person may experience difficulty in breathing during exertion and other early manifestations of heart failure, and these symptoms may even be present to a limited degree at rest. In those situations affecting primarily the left ventricle, episodes of paroxysmal nocturnal dyspnea indicating temporary inadequacy of the left ventricle may be present. There may be mild edema of the ankles at the end of the day. The manifestations are not severe or grossly disabling. As the condition becomes more severe, increasing frequency of difficulty in breathing (dyspnea) and greater amounts of edema may be present.

The degrees of heart failure

**Severe congestive heart failure.** Heart failure may become totally disabling, with severe respiratory distress and inability to lie flat and to exercise. Gross edema including abnormal amounts of fluid in the abdomen (ascites) and in the chest cavity (hydrothorax) is present. When such circumstances are prolonged, general body deterioration may develop, with various secondary effects, such as appetite loss or diarrhea.

**Cardiogenic shock.** A degree of heart failure from mild to severe may occur after acute and severe cardiac damage, such as an acute myocardial infarction. In some instances, this condition may progress to a shocklike state, with such classic manifestations as low blood pressure, a rapid, thready (thin) pulse, and pallor, a condition known as cardiogenic shock. It may or may not be accompanied by respiratory symptoms or evidence of edema. No edema from congestion in the peripheral vessels develops if the condition is of short duration, but acute pulmonary edema may be experienced. (J.V.W./M.F.O.)

## SURGICAL TREATMENT OF THE HEART

**Cardiopulmonary bypass.** Cardiopulmonary bypass serves as a temporary substitute for a patient's heart and lungs during the course of open-heart surgery. The patient's blood is pumped through a heart-lung machine for artificial introduction of oxygen and removal of carbon dioxide. Before its first successful application to operations on the human heart in the early 1950s, all heart operations had to be done either by the sense of touch or with the heart open to view but with the patient's whole body held to a subnormal temperature (hypothermia). The latter procedure was feasible only for very brief periods (less than five minutes).

The oxygenator

The first heart-lung machine (pump oxygenator) resembled only slightly the complicated apparatus currently used for correction of cardiac defects. With this machine the blood bypasses the heart and lungs so that the surgeon has an unobstructed view of the operative field. Cardiopulmonary bypass is accomplished by use of large drainage tubes (catheters) inserted in the superior and inferior venae cavae, the large veins that return the blood from the systemic circulation to the right upper chamber of the heart. The deoxygenated blood returning to the heart from the upper and lower portions of the body enters these tubes and by gravity drainage flows into a collecting reservoir on the heart-lung machine. Blood then flows into an oxygenator, the lung component of the machine, where it is exposed to an oxygen-containing gas mixture or oxygen alone. In this manner, oxygen is introduced into the blood and carbon dioxide is removed in sufficient quantities to make the blood leaving the oxygenator similar to that normally returning to the heart from the lungs.

From the oxygenator, blood is pumped back to the body and returned to the arterial tree through a cannula (small tube) introduced in a major systemic artery, such as the femoral (groin) artery. Oxygenated blood then flows to the vital organs, such as the brain, kidneys, and liver. Meanwhile, the heart may be opened and the corrective operation performed. This procedure permits a surgeon to operate on the heart for many hours, if necessary.

The assemblage and sterilization of the components of the heart-lung machine are essential considerations because the blood comes in contact with the apparatus outside of the body. Heart-lung machines have totally disposable tubing and plastic bubble oxygenators. Cardiopulmonary bypass is now more often carried out by using cardioplegic solutions designed to provide the heart with the necessary minimal nutrient and electrolyte requirements. Blood is also needed, and administration of an anticoagulant (heparin) prevents clotting of the blood while it is circulating in the heart-lung machine.

**Congenital cardiac defects.** Most congenital cardiac defects can be repaired surgically. Operations are of two general types: those that can be performed without a heart-lung machine, such as surgeries for patent ductus arteriosus and coarctation of the aorta; and those, such as intracardiac abnormalities, that require a heart-lung machine.

*Persistent (patent) ductus arteriosus.* The ductus arteriosus is the channel in utero between the pulmonary artery and the first segment of the descending thoracic

aorta. Before birth, blood flows from the right ventricle into the pulmonary artery and across the ductus arteriosus to the descending aorta. The ductus shunts blood away from the lungs because oxygen–carbon dioxide exchange begins only at the time of birth. Normally functional closure of the ductus arteriosus is completed within the first few days after birth, although complete anatomic closure may not occur for several months. If it remains open, excessive levels of blood may flow through the lungs.

First surgery for congenital heart disease

Ligation of the ductus arteriosus performed by Robert E. Gross in Boston in 1938 was the first successful operation for congenital heart disease and initiated the modern era of cardiac surgery for congenital cardiovascular lesions.

*Coarctation of the aorta.* Coarctation of the aorta, which is a constriction of the aorta, usually in the same region as the ductus arteriosus, is one of the most common congenital cardiac defects. It was first successfully repaired by Clarence Crafoord in Sweden in 1944. In older children and adolescents the narrowed area is repaired by cutting out the constriction and stitching the two normal ends together. In infants, a modified operation is used in which the left subclavian artery (the artery that supplies the left arm) is tied, divided, and used as a flap to repair the narrowed aortic area. With this procedure the stricture has less of a tendency to redevelop at that site. In adults it often may be necessary to bridge the narrowed area with a graft tube, which is attached to the aorta above and below the narrowed segment; the blood is thus able to bypass the constricted area to reach the organs below the defect.

*Pulmonary valve stenosis.* The most common congenital defect of the valves in children is a narrowing of the pulmonary valve (the valve opening to the pulmonary artery). The valve cusps in this condition are not well formed and, as a result, the valve cannot open normally. The valve cusps are thickened, and the size of the orifice varies in diameter from one millimetre to about two-thirds of that of the circumference of the pulmonary artery. As a result blood flow from the right ventricle into the lungs is obstructed. Mild stenosis is usually compatible with normal activities and normal life, but moderate and severe stenosis may result in clear symptoms.

The surgical procedure used to correct this condition is usually performed on cardiopulmonary bypass, with the valve approached through the pulmonary artery and cut in three places to create a valve with three cusps. An alternative approach to surgery is the use of a special balloon catheter, which is passed from the femoral vein (the vein in the groin) into the right side of the heart and positioned across the pulmonary valve. A balloon at the tip of the catheter is then inflated to enlarge the valve orifice.

*Aortic valve stenosis.* Although mild aortic valve stenosis is manageable in children, deterioration may occur with growth. Severe aortic stenosis in infancy and childhood may be associated with either sudden death or heart failure. The usual basis for the stenosis is fusion of the valve, which is usually bicuspid rather than tricuspid. The valve is often both obstructed and incompetent (allowing blood to leak back from the aorta into the left ventricle). Patients with more than a trivial degree of aortic stenosis usually should not take part in competitive sports such as swimming or football. In moderate to severe degrees of aortic stenosis, surgery usually is necessary and is performed using cardiopulmonary bypass. The aorta is opened just above the valve, and the surgeon incises the valve sufficiently to convert severe stenosis to a mild or moderate degree of obstruction. In older patients the valve is often thickened and calcified, and it may need to be replaced.

*Atrial and ventricular septal defects.* If atrial and ventricular septal defects require surgical closure, the patient's circulation must be supported by the heart-lung machine.

Septal repairs

Atrial septal defects are usually repaired by sewing the tissue on either side of the defect together, although very large defects may require a patch of material to close the opening. Because of the frequency of spontaneous natural closure, small ventricular septal defects are observed for a period of time before the decision is made to perform surgery. Large ventricular septal defects are usually closed by a patch.

*Cyanotic cardiovascular abnormalities.* The first attempt to treat "blue babies" affected with cyanotic abnormalities was performed by the American physicians Alfred Blalock and Helen B. Taussig in 1944. This procedure transformed the outlook for cyanotic children and for the first time made survival possible. In the early 1950s, heart-lung cardiac surgery and procedures definitive for repair were developed. Surgical treatment of the tetralogy of Fallot has been an important model for developments in more complex forms of cardiac surgery, and long-term results have been excellent. Most, but not all, forms of cyanotic congenital heart disease can now be repaired, and palliative surgery may produce considerable benefits for those in whom definitive treatment is not possible.

**Acquired heart defects.** *Valvular disease.* Destroyed heart valves can be replaced with artificial valves (prostheses) made of stainless steel, Dacron, or other special materials. The heart-lung machine is used during these operations, in which one, two, or even three cardiac valves may be removed and replaced with the appropriate artificial valve. The use of both homograft valves (obtained from human beings after death) and heterograft valves (secured from animals) is widespread. One of the advantages of both types is the absence of clotting, which occurs occasionally with the use of artificial valves. Most homograft and heterograft valves have a durability of 10–15 years. There is a risk of endocarditis with all types of values.

*Chronic constrictive pericarditis.* Chronic constrictive pericarditis can affect the surface of the heart and the sac (pericardium) surrounding it. The pericardium becomes thickened and fibrotic, and over a period of time constricts the heart so that the normal filling of the ventricles during the resting phase of the cardiac cycle is limited. This condition in turn reduces the output of the heart and eventually affects all the organ systems, including the brain, liver, and kidneys. Treatment is surgical removal of the thickened pericardium around the heart, which permits normal filling and expansion of the ventricles and restores adequate cardiac output to the vital organs.

The pericardium

*Cardiac pacemakers.* The normal rhythm of the heart is generated by spontaneous electrical activity in cells in an area of the heart called the sinoatrial node. The electrical activity is usually at a rate of about 70 beats per minute at rest and is transmitted to the pumping chambers of the heart, the atria, and the ventricles through a specialized conducting system. The electrical activity causes contraction of the heart muscle, which results in a detectable pulse at the wrist and elsewhere. Disease of the sinus node (sick sinus syndrome) or the conducting system (heart block) can cause an abnormally slow rhythm of the heart; because blood supply to the brain is inadequate, severe disease can cause loss of consciousness. This occurs if there is no heartbeat for about six seconds.

Sick sinus syndrome and heart block

A pacemaker is a device that artificially stimulates the heart when the abnormal electrical activity is absent. A pacemaker comprises a pulse generator connected to the heart by wire or electrode. The pulse generator has a battery power source and electronic circuitry that can generate an artificial stimulus at a predetermined rate. It can also detect normal activity of the heart so that the artificial stimulus is only discharged when the natural activity is absent. In this way the pacemaker functions on demand, inserting an artificial beat as required.

The pulse generator is usually placed under the skin over the right or left chest and has enough power to last several years. The electrode is passed from the pulse generator along a vein and is connected to either the atrium or ventricle, depending on whether the underlying problem is sick sinus syndrome or heart block. In many models the performance of the pacemaker can be altered by using radio frequency signals to alter its programmed settings. Some pacemakers may last up to 15 years and can be reused; the most common lifetime is seven years.

*Heart arrest.* Cessation of the heartbeat, in addition to occurring in heart block, may occur in numerous other circumstances and require emergency treatment. Until the early 1960s, when the heart stopped (cardiac arrest) the chest was opened, frequently with a penknife or any other available instrument, and the heart was massaged with

the hands until beating resumed. Closed cardiac massage now is commonly used, whereby pressure is applied to the breastbone (sternum) directly over the heart. The sternum is compressed intermittently about 60 times a minute while the lungs are expanded with oxygen. Persons may be kept alive with normal blood pressure and circulation for hours by this method. With an adequate airway or when mouth-to-mouth breathing is used, closed cardiac massage is so effective that many laymen are trained in its use for emergency support of circulation.

*Heart wounds.* Heart wounds are caused by blunt or penetrating instruments. Rapid deceleration, often experienced in automobile accidents, is a common cause of injury to the heart muscle, resulting in bruising and even disruption of a valve or the ventricular septum. Both bullet and stab wounds account for many patients treated in the emergency clinics of major hospitals. Prompt diagnosis and effective surgical treatment, usually consisting of control of bleeding by sewing the heart muscle at the point of entry of the foreign object, have resulted in a high rate of successful treatment.

Provision of new blood supply to the heart

*Coronary arterial disease.* Operations have been devised to bring a new blood supply into the heart when the coronary arteries become narrowed by atherosclerosis. A commonly used technique is to use a vein removed from the leg as a bypass around the diseased portion (Figure 23). The vein is attached to the aorta above as it leaves the left ventricle. The other end of the vein is then sutured directly to any one of the coronary arteries. Large quantities of blood can be delivered to the heart muscle by this direct form of myocardial revascularization. Implantation of an artery below the breastbone (internal mammary artery) into a coronary artery beyond the block is increasingly used.

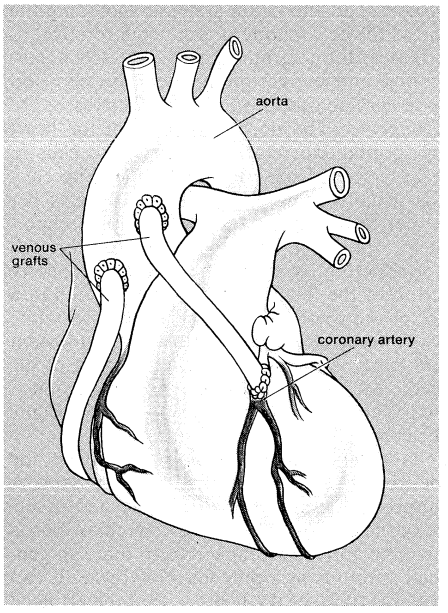



Figure 23: Restoration of blood supply to heart muscle. The vessels shown in white are venous grafts leading from the aorta to coronary vessels.

Carbon dioxide endarectomy is also performed. In this operation, carbon dioxide, which in the bloodstream is readily absorbed and does not cause blockage of the small vessels, is injected directly into the obstructed coronary artery. The gas tends to loosen the inner atheromatous core from the outer adventitial wall of the artery. The atheromatous core is then totally removed, and blood flow is restored through the artery. The heart-lung machine is used to support the circulation during most operations on the coronary arteries.

*Angioplasty.* The development of catheters with strong inflatable balloons constructed toward their end and along the line of the catheter has greatly changed cardiac surgery. The balloons can be inflated by compressed air at different controlled pressures. They are used for dilation of a partly obstructed coronary artery (percutaneous transluminal coronary angioplasty, or PTCA), with restoration of blood flow to the heart muscle, and of a severely obstructed heart valve, particularly the aortic valve, relieving the pressure on the left ventricle.

The procedure generally requires no anesthetic and, using specialized radiological imaging techniques, is sometimes done on an outpatient basis. Several coronary arteries may be dilated in this way, with flattening of the atheromatous material against and into the arterial wall. Although there are operative risks, such as emboli and tearing, the results are excellent and the technique may be repeated, if necessary.

*Heart transplantation.* If the heart muscle has been damaged beyond surgical repair, heart transplantation may be performed. The diseased heart is removed and the donor's heart is sewn in position. This procedure is particularly useful in advanced cardiomyopathy. Half of all heart transplant patients survive five years after the surgery. Heart-lung transplants are used for some intractable cardiopulmonary diseases, such as cystic fibrosis. The success rates of both of these operations has grown.

*Artificial heart.* Many scientists are directing their efforts toward development of an artificial heart. The heart-lung machine effectively supports the circulation for a few hours; if it is used for longer periods, the blood is damaged by the pumping and oxygenation. Current research is directed toward development of pumps and oxygenators that can be used for partial or total support of the circulation for long periods without damage to the blood.

Partial cardiac assistors have been used successfully to aid the heart after operation until it has had sufficient time to recover. One of these is the left ventricular bypass pump. It reduces the work load of the heart because blood is removed from the left atrium of the heart and pumped back to the body through the axillary artery. No oxygenator is required because the lungs are not bypassed.

(M.E.DeB./E.B.D./M.F.O.)

## DISEASES OF THE ARTERIES

There are many types of arterial disease. Some are generalized and affect arteries through the body, though often there is variation in the degree they are affected. Others are localized. These diseases are frequently divided into those that result in arterial occlusion (blockage) and those that are nonocclusive in their manifestations.

**Occlusive disease.** *Arteriosclerosis.* The various types of arteriosclerosis are by far the most common occlusive diseases of the arterial system and a major cause of death in one form or another. Atherosclerosis of the coronary arteries has already been discussed. Arteriosclerotic lesions of the cerebral vessels may lead to formation of blood clots and stroke.

Medial (Mönckeberg's) arteriosclerosis is the condition affecting the media (middle coat) of the main arteries. There are deposits of calcium salts in the media, but in general the channel (lumen) is not blocked. This disease is quite common and may contribute to the other major form of widespread arteriosclerosis, atherosclerosis obliterans, in which there is complete obliteration of the vessel channel.

Intermittent claudication

The classic symptom of arteriosclerosis affecting the legs is called intermittent claudication (intermittent lameness), which results from inadequate blood flow to the muscles involved in walking. In such individuals a cramplike discomfort associated with a limp occurs in the calf of the leg on exercise and is relieved by rest. The vascular disease may lead to changes in the extremities caused by reduced blood supply. These may be manifested as ulcerated areas and eventually gangrene, and most frequently involve the lower extremities. The lack of adequate blood flow renders the extremities susceptible to infection, and, therefore, the manifestations of reduced blood supply (ischemia) are frequently combined with those of infection. There is no specific therapy available.

*Thromboangiitis obliterans (Buerger's disease).* Thromboangiitis obliterans is a disease of an inflammatory nature, poorly understood, that involves the arteries of the extremities, though other vessels, including those of the heart, may be involved. It is thought to be a disease of

hypersensitivity, and the persons who are afflicted with it are often heavy cigarette smokers. It almost exclusively affects men of any race between the ages of 17 and 45 years. The symptoms are similar to those of arteriosclerosis of the peripheral vessels, with intermittent claudication, later pain during periods of rest, and eventually gangrene. An inflammation of the veins, phlebitis, occurs in almost 50 percent of those affected.

Effects of polyarteritis nodosa

*Polyarteritis nodosa.* Polyarteritis nodosa, also called periarteritis nodosa, is an uncommon disease of unknown cause; hypersensitivity may play an important role. It is more common in males and may occur at any age. Small arteries and veins in various parts of the body are affected, producing effects as a result of occlusion or bleeding or a combination of the two. The course may be rapid, involving only weeks or months, or it may be highly prolonged. The involvement of the blood vessels may affect blood flow to the skin, the gastrointestinal tract, the kidneys, and the heart. There may be associated symptoms of arthritis, and involvement of almost all organs has been noted. There is an associated fever in most instances, an increase in the number of leukocytes in the blood, and evidence of inflammation. No recognized specific mode of therapy is available.

*Arteritis.* Arteritis is an inflammation in localized segments of arteries. One particularly notable type is cranial arteritis (temporal arteritis), a disease of variable duration and unknown cause that is accompanied by fever and involves the temporal and occasionally other arteries of the skull. In general, older persons are affected. Excision of the involved artery may be carried out, but the general symptoms may remain.

*Ergotism.* Ergotism results from the excessive ingestion of ergot (a fungus that grows on rye and other cereals) and its derivatives. There is evidence of lack of blood flow and eventual gangrene in the small peripheral arteries, particularly in the legs, the nose, and the ears. Treatment includes use of measures to improve the circulation.

*Arterial thrombosis.* Arterial thrombosis, the formation of blood clots in arteries, occurs commonly in diseased vessels.

*Frostbite.* Frostbite may occur after exposure to subfreezing temperature momentarily or to less severe temperature for a longer period. It occurs more readily if blood vessels are diseased. Several degrees of frostbite produce thrombosis of the arteries and arterioles and also may involve veins. Symptoms may vary from a mild stage of reddening to gangrene and eventual loss of the extremity.

Types of embolus

*Arterial embolism.* An embolus, a foreign or abnormal particle circulating in the blood, may block a vessel too small to permit further passage. The sources of emboli include blood clots from the chambers of the diseased or abnormally functioning heart. Mural thrombosis on the infarcted ventricular wall or clots in the atrium in atrial fibrillation are common sources. Fat emboli may occur after fracture of bones and discharge of fatty marrow. Air emboli may be suspected after major injury, especially when large veins are opened during accidents or during vascular surgery of the neck or chest cavity. Bacterial emboli occur in bacterial endocarditis and occasionally in other infections. Cancers may produce minute emboli of tumour cells. Fungus growth or foreign materials, such as fragments of bullets, may become emboli. These emboli may cause transient local symptoms from diminished blood flow and may result in death of tissue. Treatment may include anticoagulant therapy and surgical removal of the clot.

**Nonocclusive disease.** *Arteriovenous fistula.* A penetrating injury such as that caused by a bullet or a sharp instrument may result in an opening between an artery and its immediately adjacent vein (an arteriovenous fistula). Large amounts of blood may be shunted from the artery to the vein. Arteriovenous fistulas are particularly common in wartime as a result of shell fragments and other types of injury involving the arms and legs. They may also occur as a complication of surgery. Others are congenital in origin.

The physician may hear a loud murmur caused by the turbulent flow of blood from the artery to the vein. Enlargement of the heart and all of the manifestations of congestive heart failure may occur if the amount of blood shunted is large. In the area around the site of the arteriovenous fistula, the blood vessels become dilated and bacterial infection of the artery lining may develop. A cure can usually be achieved by surgery, though in some situations the remaining arterial flow may be impaired.

A special kind of arteriovenous fistula occurs from the pulmonary artery to the pulmonary vein. Here the situation is complicated by the fact that unoxygenated venous blood is being shunted into a vessel normally containing oxygenated blood. Cyanosis results and produces a stimulation for formation of red blood cells, leading to a form of secondary polycythemia, abnormally high red-blood-cell level. Arteriovenous fistulas are treated by surgery.

*Physical injuries.* Physical injuries to arteries may lead to damage of the vascular wall, with consequent formation of blood clots and blockage. On other occasions, a form of inflammation may develop that may lead to rupture and may be the source of emboli in the peripheral arteries. Sudden disastrous external stress such as in severe automobile accidents, airplane crashes, and underwater explosions may cause death through rupture of the major arteries, such as the aorta, rupture of the heart valves, or rupture of the heart itself.

*Radiation injuries.* X ray, radium, and other radioactive substances in large dosages have marked effects on the vascular system. Initial reactions are inflammatory, and secondary changes caused by scarring and retractions may occur, which in turn lead to vascular occlusion (obstruction). The effects may be progressive for a period of years and are, at times, complicated by the development of cancer.

Raynaud's phenomenon and Raynaud's disease

**Functional disease.** *Vasoconstriction.* Raynaud's phenomenon is said to occur when the extremities, including occasionally even the ears, nose, or cheeks, become pale, cyanotic, and numb under the influence of cold or emotion. Pain is also present at times. On cessation of the stimulus redness develops and there is a tingling or burning sensation lasting some minutes. This sequence of events is apparently caused by the excessive constriction of the small arteries and arterioles of the fingers upon stimuli that ordinarily cause only a minor degree of vasoconstriction (constriction of blood vessels). Raynaud's disease, which is initially manifested by this phenomenon, is a disease in which there is spasmodic contraction of the blood vessels, usually beginning in early adulthood and affecting women about three times as often as men. The limb involvement is usually symmetrical (on both sides) and may lead to gangrene. Attacks may subside after return to a warm environment or release from tension.

The symptoms associated with Raynaud's disease may occur in people without other evidence of organic disease, especially in cold and moist climates. It may result from the operation of pneumatic hammers or may occur in individuals with various disorders such as a cervical rib, a supernumerary (extra) rib arising from a neck vertebra. It may appear as a complication of arteriosclerosis and thromboangiitis obliterans. Various substances, such as nicotine, arsenic, ergot, and lead, have occasionally been blamed. Therapy includes treatment of the primary condition and avoidance of the precipitating cause.

Acrocyanosis is a similar condition, characterized by episodes of coldness and cyanosis of the hands and feet. It is often associated with profuse sweating and, at times, with local edema. It is a form of local sensitivity to cold and is frequently seen in mentally or emotionally disturbed people or in those with neurocirculatory asthenia (a symptom-complex in which there is breathlessness, giddiness, and a sense of fatigue, pain in the chest over the heart, and palpitation, a fast and forcible heartbeat of which the affected person is conscious). Reassurance and avoidance of cold help to eliminate attacks.

Cause and relief of erythermalgia

*Vasodilation.* Erythermalgia (erythromelalgia) is an uncommon condition in which the extremities, especially the palms of the hands and the soles of the feet, are red, hot, painful, and often somewhat swollen. Dilation of the blood vessels (vasodilation) is the underlying factor. The condition is relieved by elevation of the extremity and

cooling. Usually it occurs in middle and later life and is chronic in the primary form; it may occur as a secondary manifestation of underlying vascular disease. It may also occur as a manifestation of an abnormally high red-blood-cell level and, occasionally, as the result of injury or a variety of other disorders.

### DISEASES OF THE VEINS

**Organic disease.** In thrombophlebitis there is thrombosis (clot formation) in the veins and a variable amount of inflammatory reaction in the vessel wall. In some instances, the inflammatory reaction is predominant and thrombosis is secondary. In other instances, thrombosis appears before reaction in the vein wall. Embolization—breaking loose of a blood clot—is most likely to occur during this period, though it may occur at any stage of the disease. A form of the disease in which little or no inflammatory reaction or pain develops is called phlebothrombosis.

Thrombophlebitis most frequently involves the veins of the legs. It may occur without apparent cause and tends to recur. At times, it occurs as a result of local injury, either from a penetrating wound or from an external blow without a break in the skin. It may occur as a result of severe muscular effort or strain and in the course of infectious diseases, thromboangiitis obliterans, and a wide range of underlying diseases. Thrombophlebitis may develop in various parts of the body if there is cancer, especially cancer of the pancreas. The presence of varicose veins in the legs causes a tendency to the development of thrombophlebitis. Treatment includes bed rest and anticoagulant therapy.

Pulmonary embolism may occur in bedridden persons as a result of a clot from a thrombophlebitic lesion, or it may occur in an apparently healthy individual. If the embolus is small, it may not have any effect on the systemic circulation. With larger pulmonary emboli, there may be massive bleeding from the lungs and the development of a large area of pulmonary infarction, resulting in sudden death. Getting up and walking soon after an operation or after congestive heart failure is the best method for avoiding pulmonary embolism. Anticoagulant therapy is useful both as prevention and as therapy after the condition has developed. Surgical removal of a massive pulmonary clot has, on rare occasions, been spectacularly successful.

Varicose veins

Varicose veins are permanently tortuous (twisted) and enlarged. The medium and large veins, especially in the legs, are most likely to be affected. The condition may occur without obvious cause or as a result of postural changes, occupation, congenital anomaly, or localized causes of increased venous pressure. The veins may be near the surface and easily seen or they may be hidden and unrecognized. Without complication, they rarely cause symptoms, but they may become the site of thrombophlebitis with inflammatory changes and the production of emboli in the peripheral circulation. The veins may rupture on occasion, with bleeding into the surrounding tissues. Varicose veins may occur around the rectum and anus, producing hemorrhoids. If they occur within the scrotal sac in the region of the testes, they are called varicocele. In all forms of varicose veins, the walls of the veins become hardened, and a certain amount of inflammation develops through the years. Dilated veins in the legs may be supported by appropriate elastic-type stockings or bandages, or they may be treated by surgery.

**Functional disease.** Direct mechanical injury or an infection or other disease process in the neighbouring tissues may produce spasms in the veins (venospasms). Local venospasm is usually of relatively minor significance because of the adequacy of alternate pathways for the blood. If venospasm is widespread, however, involving an entire extremity, or the veins in the lungs, it may impair blood flow and therefore be of greater significance.

### DISEASES OF THE CAPILLARIES

The capillaries are the smallest blood vessels. Through their thin walls oxygen and nutrients pass to the tissue cells, in exchange for carbon dioxide and other products of cellular activity. Despite the small size and thin walls of the capillaries, the blood pressures may be quite high as, for instance, in the legs of a person in a motionless, upright position. In certain disease states, there is increased fragility of the capillary wall, with resultant hemorrhages into the tissues. These hemorrhages are referred to as petechiae when small or, if large, may become a large area of discolouration of the skin. Vitamin C deficiency and a variety of blood disorders may be associated with increased capillary fragility. Small petechial hemorrhages occur in bacterial endocarditis and certain other infectious processes. In some instances, petechiae are caused by minute emboli; in others, they appear to be directly related to capillary fragility itself. Treatment is of the underlying disorder.

Capillary fragility

The capillaries are freely permeable to water and small molecules but ordinarily are not highly permeable to proteins and other materials. In some pathological situations, such as in certain allergic states (*e.g.,* hives) or because of local injury as in burns, there may be local areas of permeability, with escape of fluid high in protein into the surrounding tissues. If the disease affects the entire body, a significant amount of plasma (the blood minus its cells) leaks into the nonvascular spaces, with resultant loss in blood volume. Again, treatment is of the underlying disorder.

### HEMODYNAMIC DISORDERS

**Hypertension.** Hypertensive heart disease has been discussed above.

**Hypotension.** Moderate hypotension (low blood pressure) may occur in persons who are weak and enfeebled but more often does not represent a diseased state. Indeed, life insurance figures demonstrate that the life expectancy of people with such a condition is greater than average. Hypotension of a severe degree may develop in heart failure, after hemorrhage, in overwhelming infections, and in a variety of circumstances that lead to the development of the clinical picture of shock. In shock, the circulation is inadequate, blood pressure is low, heart rate is rapid, and irreversible tissue damage from insufficient blood supply may occur if the condition is not terminated (see below *Physiological shock*). Transient hypotension may occur as a normal reaction in certain forms of syncope but is not necessarily associated with organic disease.

**Syncope.** Syncope is the sudden loss of consciousness associated with a transient disorganization of circulatory function, as differentiated from other brief losses of consciousness associated with abnormal central nervous system activities, as in certain forms of epilepsy.

Fainting

The most common kind of syncope is ordinary fainting. Some individuals are more susceptible than others. Blood loss, exhaustion, the presence of other illness, and psychological factors may contribute to a tendency to faint. An affected person is usually in the upright position, becomes weak, pale, and sweaty, and may have nausea. The heart rate at this time is usually relatively rapid, but with the abrupt onset of syncope the heart rate falls often to below the normal level and the person collapses as if dead. There is usually a rapid recovery without complications.

Syncope may be related to a transient cessation of circulatory activity due to heart block. Other forms of syncope occur as a result of lowered blood pressure upon assumption of an upright position, a condition often called orthostatic hypotension. In some individuals, disease of the autonomic nervous system prevents appropriate postural adjustments for the upright stance. The disorder may be caused by vascular or central nervous system involvement of the autonomic system. In other instances, postural hypotension may occur as a result of inadequate blood volume, of taking various drugs that affect the nervous control of the circulation, and from a wide variety of other causes. Transient hypotension also may result from hypersensitivity of the carotid sinus. (J.V.W./M.F.O.)

### PHYSIOLOGICAL SHOCK

Physiological shock may be defined as acute progressive circulatory failure in which the tissues receive an inadequate supply of blood and its components (such as nutrients and oxygen) and an inadequate removal of wastes. The result is cell damage and, eventually, cell death. This

definition is derived from the one constant feature of physiological shock, the failure of adequate blood flow through the capillaries, the smallest of the blood vessels. Shock may be so severe as to impair organ function or create a state of blood flow deficiency that grows progressively more dangerous.

All, or some, of the following features may be present: a raised pulse rate, each pulse being diminished in strength, or "thready"; diminished arterial blood pressure; a cold, sweaty skin; rapid, deep respirations (breathing called air hunger in this context); mental confusion; dilated pupils; a dry mouth; and diminished flow of urine.

Many classifications of various types of shock have been developed, the most popular being one that relates the state to its cause; *e.g.,* hemorrhage (bleeding) or sepsis (invasion by pus-forming organisms). This classification has the advantage of possessing diagnostic and therapeutic implications, but unfortunately the cause of shock is often uncertain. Alternatively, the state may be called hypovolemic if it is associated with a reduction in blood volume or normovolemic if no blood-volume diminution has occurred. Shock can also be subdivided into "warm hypotension," if the sufferer has lowered blood pressure and warm dry skin, and "cold hypotension," if the skin is cold.

**Shock due to inadequate blood volume.** Hemorrhage is the most common cause of shock. In the "average man" (weighing 70 kilograms, or about 154 pounds) the blood volume is about five litres (about 5.3 quarts), and the loss of any part of this will initiate certain cardiovascular reflexes. Hemorrhage results in a diminished return of venous blood to the heart, the output of which therefore falls, causing a lowering of the arterial blood pressure. When this occurs pressure receptors (baroreceptors) in the aorta and carotid arteries will initiate remedial reflexes either through the autonomic (nonvoluntary) nervous system by direct neural transmission or by epinephrine (adrenaline) secretion into the blood from the adrenal gland.

Baroreceptors

The reflexes consist of an increase in the rate and power of the heartbeat, increasing its output; a constriction of arterioles leading to nonessential capillary beds (notably the skin and some viscera); and a constriction of the veins, diminishing the large proportion of the blood volume normally contained therein. By these means arterial blood pressure will tend to be maintained, thus preserving blood flow to the vital areas such as the brain and the myocardium. After continued acute blood loss of 20–30 percent of the blood volume, the compensatory mechanism will begin to fail, the blood pressure will begin to fall, and shock will ensue.

Increased sympathetic (autonomic) nervous activity thus accounts for the fast pulse rate, pallor, and coldness of the skin in shock and, in addition, is the cause of increased sweating and dilation of the pupils of the eyes. Air hunger and mental confusion are caused by the inadequate carriage of oxygen, and decreased urine flow stems from a decrease in the renal (kidney) blood flow, which, if severe, can lead to kidney failure. If acute blood loss continues beyond about 50 percent, the inadequacy of flow through vital circulations will lead to death. Loss of whole blood is not necessary for the blood volume to be low, for plasma loss through burnt areas of the skin, dehydration following inadequate intake of fluid, or exceptional fluid loss can lead to contraction of the blood volume to levels capable of causing shock.

When it is possible to anticipate blood loss, and to measure it accurately—*e.g.,* during an operation—losses may be immediately replaced before significant volume depletion can occur. More often, however, hemorrhage is unexpected; it may not be possible to measure the amount of blood lost. If shock occurs in an otherwise healthy person it generally is replaced by transfusion into a vein. But since a preliminary matching between recipient serum and donor cells must be carried out, and cannot be done in less than 20 minutes, other fluid is usually given intravenously during the delay. This fluid, such as plasma or a solution of the carbohydrate dextran, must contain molecules large enough that they do not diffuse through the vessel walls. Since the main loss from burns is plasma and electrolytes, these require replacement in proportion to the area of the burn and the size of the patient.

**Shock due to inadequate cardiac output.** Sudden interference with the blood supply to the heart muscle, as by a thrombosis in a coronary artery, causes damage to the muscle with resultant diminution in its contractile force. The output of the heart falls; if the decline is severe, a fall in blood pressure stimulates the baroreceptors and, in the way just described, cardiogenic shock results. This occurs uncommonly after myocardial infarction. But low heart output alone may not account for the shock, for in chronic heart failure the cardiac output may be low without such a response in the peripheral circulation.

Cardiogenic shock

**Shock due to increased circulatory capacity.** If widespread dilation of the veins or of the capillary beds occurs, the blood volume is no longer sufficient to fill the larger space and shock ensues.

*Bacteremic shock.* Infection anywhere in the body may spread to the circulation, and the presence of organisms in the bloodstream—bacteremia—may lead to shock. Bacteria are conveniently divided into "gram-positive" and "gram-negative" groups according to their reaction to a special staining method called Gram's stain.

Gram-negative bacteremia is the more common and more lethal type of bacteremic shock. It is frequently caused by *Escherichia coli, Proteus, Pseudomonas,* or *Klebsiella* organisms; the first of these normally inhabits the intestine. The clinical picture of gram-negative bacteremia is much like that of hemorrhage, although no blood has been lost. This type of shock typically causes a rapid, thready pulse; a cold, sweaty skin; and low blood pressure. A fever may occur, in addition to the local signs of the associated infection. The cause of the type of reaction is uncertain. The response to bacteremia from gram-positive organisms such as *Staphylococcus* and *Streptococcus* is different. Widespread dilation of the blood vessels results in a warm, dry skin and a full volume pulse in spite of lowered blood pressure.

In both types of bacteremia the condition may be exacerbated by contraction in blood volume. This follows fluid loss, *e.g.,* that lost in the peritoneal cavity in peritonitis—inflammation of the peritoneum, the membrane that lines the abdominal cavity; in the tissues in streptococcal infection; or through the intestine in enteritis (inflammation of the intestine).

Exceeded in frequency only by cardiogenic and hemorrhagic shock, bacteremic shock is most often caused by gram-negative organisms. There are three aspects of treatment: collections of pus are drained as soon as possible; the circulatory volume is increased to compensate for enlargement of the vascular bed; and appropriate antibiotics are administered.

Treatment

*Anaphylactic shock.* Anaphylactic shock is discussed in detail in the article IMMUNITY. An anaphylactic reaction is the direct result of the entrance of a specific foreign material into the bloodstream of a person whose body has become sensitized against it as a result of previous exposure and subsequent formation of antibodies. During an anaphylactic reaction, lung bronchi constrict intensely, narrowing the airways and interfering seriously with respiration; blood pressure may fall precipitously because of the release of substances (serotonin, histamine, and bradykinin) that cause dilation of the arterioles and venules and an increase in the capillary-wall permeability. Thus the circulatory capacity is increased, and fluid is lost into the tissues.

The essence of treatment of anaphylactic shock is the injection of epinephrine, a powerful stimulatory drug also found naturally in the body, whose effects include an increase in the heart rate and constriction of the blood vessels, followed by an antihistamine, to counteract the reaction to the foreign substance, and a bronchodilator, to ease breathing.

*Psychogenic shock.* Psychogenic shock causes fainting, probably by initiating dilation of the blood vessels that perfuse the muscles. In this type of shock, blood pressure falls, the skin becomes cold and sweaty, and the pulse rate increases. A decrease in the amount of blood that is supplied to the brain leads to light-headedness and loss of

consciousness. A person who is suffering from psychogenic shock should be placed flat or even with the head slightly lower than the rest of the body in order to restore a good flow of blood to the brain and to bring about recovery from the fainting.

*Drugs and shock.* Most anesthetic drugs—nitrous oxide is a notable exception—have a profound effect on the circulation. They are able to decrease the contractility of the heart muscle as well as increase the circulatory capacity by dilating the blood vessels. In addition, the normal postural circulatory reflexes are lost, so that pooling of blood in the legs is liable to occur if the affected person is tilted to a head-up position. This is of particular importance after surgery; if a person is made to sit up too soon, it can lead to low blood pressure and an insufficient flow of blood to the brain. Overdosage of certain drugs, notably barbiturates, narcotics, and tranquilizers, blocks normal circulatory reflexes and causes dilation of the blood vessels, leading to a fall in blood pressure that often is accompanied by a slow, full-volume pulse.

A blood pressure that is dangerously low may be raised to safer levels by affecting the activity of the offending drug in one of many different ways. A therapeutic approach might entail, for instance, decreasing the dosage of the drug (such as an anesthetic agent), speeding up its elimination from the body, or administering a substance that is able to constrict the blood vessels. The choice of approach depends on the individual circumstance.

*Neurogenic shock.* The maintenance of the tone of the blood vessels by the autonomic nervous system may be affected by severance of one of these nerves or by its interruption of the flow of nervous impulses. Thus, spinal anesthesia—injection of an anesthetic into the space surrounding the spinal cord—or severance of the spinal cord results in a fall in blood pressure because of dilation of the blood vessels in the lower portion of the body and a resultant diminution of venous return to the heart.

Neurogenic shock does not usually require specific therapy; indeed, spinal anesthetics may be given with a view to producing a low blood pressure so as to diminish bleeding during an operation. If blood pressure becomes critically low, the legs are sometimes elevated and a vasoconstrictor administered.

*Endocrine causes of shock.* The endocrines play a vital role in the regulation of normal metabolic processes through the actions of their hormones. It is not surprising therefore that a malfunction in an endocrine gland or in its hormones has an effect on circulation. Inadequate secretion by the adrenal cortex, the outer substance of the adrenal gland, leads to shock both by the diminution of myocardial efficiency and by a decrease in the blood volume. Functional disorders of the pituitary, the adrenal medulla (the inner substance of the adrenal gland), the thyroid, and the parathyroids can all lead to circulatory upset and shock.

Definitions

**Refractory and irreversible shock.** The terms refractory shock and irreversible shock are widely used by physicians and other medical workers to refer to types of shock that present particularly difficult problems. The term refractory shock is applied when in spite of apparently adequate therapy the shock state continues. Commonly the treatment proves later to have been inadequate, in which case the shock was not true refractory shock. This often occurs following a major injury in which there is internal bleeding, leading to underestimation of true blood loss and therefore to inadequate transfusion. In certain cases, however, even if the therapy actually is appropriate, the shock state persists; if patients in such cases respond to further special treatment, then this is true physiological refractory shock.

In severe or prolonged shock states the myocardial blood supply is diminished sufficiently to damage the heart's pumping action temporarily or permanently. Also, noxious products of inadequately perfused tissues may circulate and affect the heart muscle.

While the flow of blood through major vessels is under the control of the nerves, circulation through the capillary beds is of a more primitive type and is under the influence of local metabolic products. In shock, arteriolar constriction causes inadequate flow through the tissues and local waste products increase. These cause dilation of the capillary sphincters and opening of the whole capillary bed, which thus contains an increased proportion of the blood volume. The capillaries become further engorged with slowly flowing blood, and fluid leaks through the vessel walls into the tissues. Thus the body is further deprived of circulating blood volume.

Possibility of clotting in capillaries

Widespread clotting of the blood can occur during capillary stagnation. This leads to severe damage to the cells unsupplied by flowing blood. Later, when the clot is dissolved under the influence of enzymes that dissolve the fibrin of the clots, the flow through these areas carries toxic metabolic products to vital organs such as the heart, kidneys, or liver, and the ensuing damage leads to irreversibility of shock.

**Identification of the causes of shock.** Several causes of shock may be present in any given instance. For example, in peritonitis—inflammation of the lining membrane of the abdominal cavity—bacteria enter the bloodstream as a result of the infection, and fluid loss into the peritoneal cavity leads to a low blood volume. In pancreatitis, inflammation of the pancreas, the cause of shock may be threefold: bleeding and escape of fluid into the tissues may lead to diminished blood volume; there may be increased release of the vasodilator bradykinin, leading to increased circulatory capacity; and bacteremia, the presence of bacteria in the bloodstream, may exist. In severe infection with *Streptococcus* organisms, there may be bacteremia, a low blood volume from loss of fluids into the tissues, and toxic damage to the heart muscle.

Hemorrhage may present no difficulties in recognition, since losses of blood from the intestine or from wounds and fractures usually have particular features. The bleeding, however, may not be evident if it occurs in the chest or abdomen after trauma or if the blood fails to drain from deep-seated operative wounds. When bacteremia occurs there is often an obvious source of infection, such as peritonitis, pneumonia, or urinary tract infection, but such loci of infection may be small or deep-seated, so that final proof of cause has to be determined from culture of the blood.

Cardiogenic shock usually occurs only along with typical features of coronary thrombosis—severe central chest pain perhaps radiating to the arm or neck and showing typical changes on an electrocardiogram (ECG). But myocardial infarction, the death of a section of heart muscle, can occur without causing any signs; it may happen while a person is under anesthesia, for example, and it does not always produce typical changes in the ECG at first. Furthermore, low blood pressure from any cause can result in ECG changes, further obscuring the cause of shock.

**Prognosis.** During the course of treatment, some indication of the outlook may be had from measurement of the lactic acid in the blood. The higher its level, the greater is the chance of irreversibility. The outlook may also depend on the cause of shock and the speed and vigour with which it is treated. At one end of the scale the mortality of immediately treated shock following hemorrhage is small, while at the other severe cardiogenic shock carries a mortality rate of about 80 percent. In septic shock the mortality is about 50 percent. (W.G.Pr./Ed.)

BIBLIOGRAPHY

*Circulation and circulatory systems:* General accounts and elementary descriptions of circulatory systems are found in many biology textbooks, including the following: RAYMOND F. ORAM, *Biology: Living Systems,* 5th ed. (1986); KAREN ARMS and PAMELA S. CAMP, *Biology,* 3rd ed. (1986); and PAUL B. WEISZ and RICHARD N. KEOGH, *The Science of Biology,* 5th ed. (1982). Textbooks dealing with animal structure at a more advanced level include the following: RALPH M. BUCHSBAUM, *Animals Without Backbones,* 3rd ed. (1987); ROBERT D. BARNES, *Invertebrate Zoology,* 5th ed. (1987); ALFRED SHERWOOD ROMER and THOMAS S. PARSONS, *The Vertebrate Body,* 6th ed. (1986); and CHARLES K. WEICHERT, *Anatomy of the Chordates,* 4th ed. (1970); KNUT SCHMIDT-NIELSEN, *Animal Physiology: Adaptation and Environment,* 3rd ed. (1983); and MILTON HILDEBRAND, *Analysis of Vertebrate Structure,* 2nd ed. (1982).

For the history of circulation studies, see HELEN RAPSON, *The Circulation of Blood* (1982); DAVID J. FURLEY and J.S. WILKIE

(eds.), *Galen on Respiration and the Arteries* (1984); *The Selected Writings of William Gilbert, Galileo Galilei, William Harvey* (1952), in "The Great Books of the Western World" series; FREDRICK A. WILLIUS and THOMAS J. DRY, *A History of the Heart and the Circulation* (1948); and ALFRED P. FISHMAN and DICKINSON W. RICHARDS, *Circulation of the Blood: Men and Ideas* (1964, reprinted 1982). Special studies of circulation include DONALD A. MCDONALD, *Blood Flow in Arteries,* 2nd ed. (1974); DAVID I. ABRAMSON and PHILIP B. DOBRIN (eds.), *Blood Vessels and Lymphatics in Organ Systems* (1984); COLIN L. SCHWARTZ, NICHOLAS T. WERTHESSEN, and STEWART WOLF, *Structure and Function of the Circulation,* 3 vol. (1980–81); and JERRY FRANKLIN GREEN, *Fundamental Cardiovascular and Pulmonary Physiology,* 2nd ed. (1987).

*The human cardiovascular system:* STANLEY W. JACOB, CLARICE ASHWORTH FRANCONE, and WALTER J. LOSSOW, *Structure and Function in Man,* 5th ed. (1982); and GARY A. THIBODEAU, *Anatomy and Physiology* (1987), are basic texts. ARTHUR C. GUYTON, *Human Physiology and Mechanisms of Disease,* 4th ed. (1987), is a technical description of the physiology of cardiac muscle, heart function, and hemodynamics. Also see PETER F. COHN, *Clinical Cardiovascular Physiology* (1985); JAMES J. SMITH and JOHN P. KAMPINE, *Circulatory Physiology: The Essentials* (1984); HARVEY V. SPARKS, JR., and THOM W. ROOKE, *Essentials of Cardiovascular Physiology* (1987); and PETER HARRIS and DONALD HEATH, *The Human Pulmonary Circulation,* 3rd ed. (1986).

*Cardiovascular system diseases and disorders:* J. WILLIS HURST *et al.* (eds.), *The Heart, Arteries, and Veins,* 6th ed. (1986); HARRY A. FOZZARD *et al.* (eds.), *The Heart and Cardiovascular System: Scientific Foundations,* 2 vol. (1986); WRYNN SMITH, *Cardiovascular Disease* (1987); ARTHUR J. MOSS, *Moss' Heart Disease in Infants, Children, and Adolescents,* 3rd ed., edited by FORREST H. ADAMS and GEORGE C. EMMANOUILIDES (1983); GAIL G. AHUMADA (ed.), *Cardiovascular Pathophysiology* (1987); ROBERT H. ANDERSON *et al.* (eds.), *Paediatric Cardiology,* 2 vol. (1987); EDWARD K. CHUNG (ed.), *Quick Reference to Cardiovascular Diseases,* 3rd ed. (1987); and ANDERS G. OLSSON (ed.), *Atherosclerosis: Biology and Clinical Science* (1987).

Treatment and prevention of vascular problems are the subject of JEANETTE KERNICKI, BARBARA L. BULLOCK, and JOHN MATTHEWS, *Cardiovascular Nursing: Rationale for Therapy and Nursing Approach* (1970); GEORGE E. BURCH and TRAVIS WINSOR, *A Primer of Electrocardiography,* 6th ed. (1972); JOSEPH K. PERLOFF, *Physical Examination of the Heart and Circulation* (1982); JEFFREY W. ELIAS and PHILLIP HOWARD MARSHALL (eds.), *Cardiovascular Disease and Behavior* (1987); and EUGENE BRAUNWALD (ed.), *Heart Disease: A Textbook of Cardiovascular Medicine,* 2nd ed. (1984).

(M.F.O.)

# Circus

In its modern sense the word circus refers to an entertainment composed primarily of trained animal acts and exhibitions of human skill and daring. The word has the same root as *circle* and *circumference* and therefore also recalls the distinctive environment in which such entertainment is presented—the ring, a circular performance area usually bounded by a short fence (or "curb") and surrounded by tiers of seats for spectators, which may itself be enclosed in a circular building or tent. But variations exist, and any attempt at a strict, all-inclusive definition would reduce the term's application to a particular nationality, generation, or proprietorship. Some circuses dispense with trained animals, for example; others, particularly in the United States, exhibit simultaneously in three or more rings, with the building or tent then taking on a rectangular or elliptical shape. At various times circuses have offered supplementary attractions such as street parades, menageries, sideshows or "museum" departments, and pantomimes and theatrical presentations. A number of circuses, especially in Europe, have been stationary, occupying permanent, often elegant buildings in the larger cities. Others have traveled extensively—originally by horse and wagon, then by railroad, boat, motor vehicle, or even airplane—exhibiting in tents or, occasionally, theatres and, more recently, in huge enclosed sports arenas. A few organizations, such as Ringling Bros. and Barnum & Bailey, can point to a history extending back a century or more; other circuses, such as those sponsored by fraternal organizations (*e.g.*, the Shriners), may exist for less than a single season. Some shows have traditionally emphasized spectacular elements such as gorgeous costumes, floats and pageant wagons, numbers set on water or ice, or perhaps a theme running through part or all of the program. Others have been built around the talents of outstanding individual artists such as the 19th-century American clown Dan Rice or the sensational German animal trainer Gunther Gebel-Williams. Through all the above, however, there runs a common thread: the ring, by which spectators readily recognize the entertainment known as "circus."

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, section 622.

This article is divided into the following sections:


Origins of the circus 418
Circus attractions 419
  Equestrian acts 419
  Wild animal acts 420
  Acts of skill 420
  Clowns 421
  Supporting attractions 421
Commercial and technical aspects 421
  The circus train 422
  The parade 422
  The big top 423
  Winter quarters 423
Variations on the circus theme 423
  Wild West shows 423
  Carnivals 423
The state of the art 423
  Early 20th-century changes 423
  Preservation of the past 424
  Contemporary developments 424
Bibliography 424


## ORIGINS OF THE CIRCUS

Contrary to popular opinion, the circus is of comparatively recent origin, having little in common with ancient Roman circuses and amphitheatres. The former were in fact racecourses, whereas the great oval amphitheatres were most often devoted to gladiatorial combats, the slaughter of animals, mock naval battles, and spectacular pantomimes. Certainly many numbers associated with the circus were known to ancient peoples. Acrobatics and juggling are probably as old as mankind itself; the Greeks saw ropedancers perform; the Romans made a specialty of training animals, including elephants, which were also supposedly taught to walk ropes; and clowns are found in nearly every period and civilization, both as characters in farces and as individual performers. Until the mid-18th century, however, there were no attempts to organize such acts into a distinct entertainment, although in the meantime they had continued to be exhibited by individuals and small troupes of performers who wandered through Europe, Asia, and Africa. They appeared wherever groups of people gathered: in nobles' halls, at community celebrations, and at marketplaces. King Alfred the Great of England (reigned 871–899) was entertained by a wild beast show; William the Conqueror (king of England 1066–87) brought performing troupes of ropedancers, tumblers, and contortionists from France to England. Itinerant players trained bears, monkeys, horses, dogs, and other animals and brought them to fairs. Fairs played an important role in developing trade throughout Europe from the 7th century onward. By the late medieval period, however, as more regular channels of marketing were standardized, they tended to become less a place for trading than for entertainment, featuring many of the performers, animals, and other characteristics later associated with the circus.

Early entertainers

The modern circus came into being in England in 1768 when Philip Astley (1742–1814), a former sergeant major turned trick rider, found that if he galloped in a circle while standing on his horse's back, centrifugal force helped him to keep his balance. In doing so he traced the first ring. Astley soon engaged a clown ("Mr. Merryman"), musicians, and other performers for his establishment, roofed over his ring, and added a stage for dramatic performances. Astley himself always referred to his building, which was located on the Lambeth side of London's Westminster Bridge, as an "amphitheatre" or "riding school." The term "circus" was first employed in 1782 when a rival horseman and former employee named Charles Hughes opened the Royal Circus a few hundred yards south of Astley's Amphitheatre.

Circumstances were favourable for the development of the circus. During the second half of the 18th century the fairs were going into decline, and some of their main attractions were banned. Many showmen, looking for a new outlet, found it in the circus, as did ropedancers, acrobats, jugglers, and others whose performances were based on dexterity, agility, and strength and who discovered that their performances were better appreciated in the ring.

Spread of the circus through Europe

In 1772 Astley went to France to present his "daring feats of horsemanship" before the king and the French court, and he found that there, too, many showmen were ready to forsake the fairgrounds. Ten years later he returned to Paris and opened an amphitheatre. When hostilities between Britain and France developed after the outbreak of the French Revolution, he leased it to Antonio Franconi (1737–1836), a member of a noble Venetian family, who had been forced into exile after a fatal duel. He became first a showman and later a trick rider, but it was as a director that he excelled. Franconi joined forces with Astley and, in Astley's absence, continued on his own. His sons, Laurent and Henri, together with their wives

and children, continued in his footsteps, and the Franconi family is generally credited with being the founders of the French circus. They are reputed to have standardized the diameter of the ring at 13 metres (approximately 42 feet), a size that is still recognized in most circuses.

In 1782 Astley traveled as far as Belgrade, visiting Brussels and Vienna on the way, and during his life he built at least 19 permanent circuses. It was Hughes, however, who introduced the circus to Russia. He added a company of trick riders to the stud of horses he had been commissioned to deliver to Catherine the Great in 1793, and he was rewarded with a private circus in the royal palace in St. Petersburg. The Russian circus was further developed by a Frenchman, Jacques Tourniaire (1772–1829).

In 1793 John Bill Ricketts, a rider who had previously performed in Britain, opened circuses in Philadelphia and New York City, the first seen in the New World. At the same time Benito Guerre was presenting his feats of horsemanship in Spain, and a cut on a contemporary handbill shows a rider leaping through a paper hoop—a scene that still epitomizes this form of entertainment. By the turn of the century the circus had spread throughout Europe and was firmly established in America. Performances were still given mostly in permanent or semipermanent buildings of flimsy construction. The greatest hazard was fire, from which Astley and Ricketts suffered particularly: Astley's Amphitheatre burned down three times in the first 62 years of its history, and Ricketts lost his circuses in both New York City and Philadelphia as a result of fires.

From the time of its origin in England the circus was often coupled with the theatre. At Astley's, the Royal Circus, and elsewhere a proscenium arch and large scenic stage were set behind the ring. Equestrian drama, with plots based on famous battles and sieges and with the horses and other artists participating at full gallop, became all the rage. In France this genre of entertainment was eventually exiled to regular theatres, but in England it continued to flourish in permanent and tenting circuses throughout the 19th century. Shakespeare's *Richard III* and *Macbeth,* and even Verdi's opera *Il trovatore,* were performed on horseback at Astley's during this period. Astley's, however, never became as fashionable as the permanent circuses on the Continent. The most exclusive club in Paris kept its own private box at the Cirque d'Été, and in St. Petersburg the stables were regularly scented for the benefit of aristocratic visitors.

Equestrian drama

Circus families became prominent during the 19th century. From one generation to another, members of a family would be trained from earliest childhood in the skills and discipline necessary to achieve perfection either in one specialty or in a group of related specialties. For example, the Cristiani family of Italy was known chiefly for its expert riders, but some members excelled in the skills of tumbling, ballet, and acrobatics. Circus families often intermarried. The Cooke family, which traveled from Scotland to New York City in the early 1800s, was an equestrian group that intermarried with the Coles and the Ortons, both well-known American circus families. As a circus family expanded, branches were established in numerous areas and members often went from one branch to another. The Cristiani family established branches elsewhere in Europe and the United States. Toward the end of the 19th century the Russian circus was dominated by an Italian circus family, the Cinisellis, which, like many others, made its name outside its own country.

Circus families

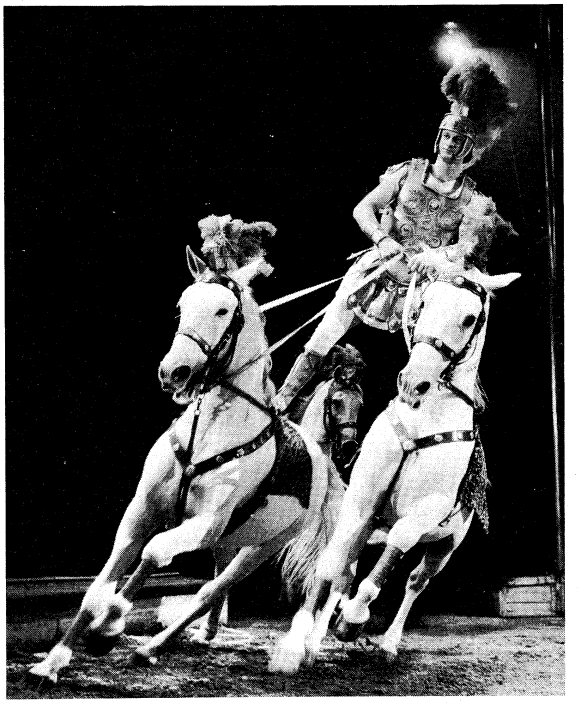
At the beginning of the 20th century the circus was still spreading to many other countries. The British circus family of Harmston settled in East Asia, and for years their only rival was the Russian circus, Isako. The Boswells left England for South Africa, where they met competition from the German Pagel. Frank Brown, whose father had been a clown at Astley's, toured South America for many seasons. In Australia the circus prospered under the Wirths. The Lobes, from Budapest, made Persia their tenting ground, and the Sidolis settled in Romania.

## CIRCUS ATTRACTIONS

**Equestrian acts.** From its origin under Astley, the circus has retained performing horses and riders as a fundamental feature. Most great riders have been champions of the art of bareback riding, performing acrobatic and gymnastic feats on the bare backs of loping horses. James Robinson, a mid-19th-century American, was one such rider. He was billed as "the One Great and Only Hero and Bareback Horseman and Gold Champion-Belted Emperor of All Equestrians." Horses that perform free of rider, reins, or harness, directed solely by visual or oral command, are called liberty horses. The Barnum & Bailey Circus, the most famous name in the American circus, in 1897 presented the largest troupe of liberty horses, 70 performing simultaneously in one ring. The traditional finale of the larger tent shows was the Great Roman Hippodrome Races, composed of novelty races, steeplechase, and the ancient arts of chariot racing and Roman standing riding.

Another type of riding, extremely popular in the 19th century before the purely acrobatic style supplanted it, was scenic riding, in which the equestrian, appropriately costumed, acted out a pantomime on horseback. The greatest exponent of this artistic mode of riding was the Englishman Andrew Ducrow (1793–1842), who also managed Astley's during the last two decades of his life. One of his creations, "The Courier of St. Petersburg," is still seen in the circus. In this act a rider straddles two cantering horses while other horses, bearing the flags of those countries that a courier would traverse on his journey to Russia, pass between his legs. Besides other solo acts, which were copied by equestrians throughout the world, Ducrow invented several duets and ensemble numbers. In "The Tyrolean Shepherd and Swiss Milkmaid," for example, he was joined by his wife, Louisa Woolford, and, while standing on the backs of their circling horses, the two performed the pursuit and wooing of the fair peasant, complete with a lovers' quarrel and reconciliation scene, followed by an exquisite pas de deux.

Keystone



D. Kossmeyer, in Bertram Mills Circus, London, performing an equestrian act reminiscent of feats originated by Andrew Ducrow.

Equestrian acts of the 20th century can be divided into three main groups: voltige, in which a rider vaults onto and off a horse's back; trick riding, in which the standing rider performs somersaults and pirouettes or forms human pyramids with other riders on one or more horses; and high school, a spectacular form of dressage in which a horse executes complex maneuvers in response to imperceptible commands communicated through slight shiftings in the rider's weight, pressure exerted by the knees and legs, or the handling of the reins. The Schumann fam-

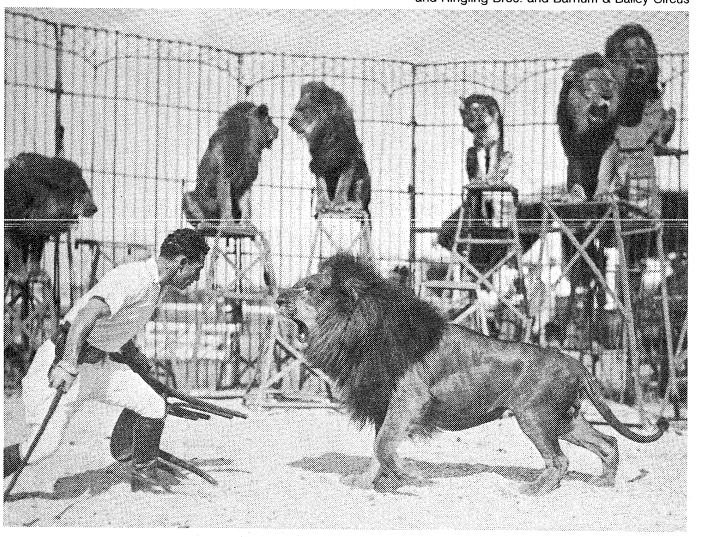
Voltige, trick riding, and high school

ily, for many years directors of the permanent circus in Copenhagen, excelled in high school and also exhibited many fine liberty-horse acts.

**Wild animal acts.** The introduction of wild animals to the circus dates from about 1831, when the French trainer Henri Martin (1793–1882) performed with his lions, boa constrictors, elephant, and other animals at the Cirque Olympique in Paris. He was soon followed by the American trainer Isaac A. Van Amburgh (1801–65), reputedly the first man to stick his head into a lion's mouth, who in 1838 took his act to England and so fascinated the young Queen Victoria that she commissioned the artist Edwin Landseer to paint a portrait of the brawny American with his "big cats." In addition to exhibiting at circuses, both Martin and Van Amburgh frequently appeared on the boards of regular theatres, where they and their animals were featured in melodramas with such titles as *The Lions of Mysore* and *The Brute Tamer of Pompeii.* Other dramas performed in theatres and circuses about this time featured elephants and bears in starring roles.

Until the late 20th century there was a marked difference between European and American styles of presenting wild animal acts. Van Amburgh customarily beat his animals into submission, and the belief that the trainer must demonstrate physical superiority over his "pupils" was held by many of his successors. The master of this style in the 20th century was the American Clyde Beatty (1903–65), who subjugated as many as 40 "black-maned African lions and Royal Bengal tigers" at one time.

By courtesy of Circus World Museum, Baraboo, Wis., and Ringling Bros. and Barnum & Bailey Circus



Clyde Beatty training lions, *c.* 1932.

In the European style of presentation, as developed by the Hagenbecks in Germany about the end of the 19th century and now followed by most American trainers as well, the trainer endeavours to prove his mastery and skill by presenting his jungle charges in the role of obedient, even playful, pets. The wild character of the animals, however, is revealed just often enough to remind the spectator that what he sees is indeed the result of skillful training.

The hallmark of circuses

Elephants are considered by many to be the very hallmark of the circus. Their awesome size makes their adaptability to humans a curious paradox. From single specimens exhibited at Astley's in the early 19th century, the number of performing elephants, especially in American three-ring circuses, has sometimes run as high as 50 or more. Despite their pleasant-looking appearance and ability to learn a surprising variety of feats, these seemingly docile beasts are among the most dangerous of circus animals.

Countless other species of wild animals have been trained to perform in the circus ring, including polar bears, giraffes, hippopotamuses, and rhinoceroses. The members of the Knie family of Switzerland are celebrated for their expertise in the handling of such exotic animals.

**Acts of skill.** In 1859 the invention of the flying trapeze by the French acrobat J. Léotard (1838–70), and Charles



The Bubnovs, aerial gymnasts from the Moscow Circus.
Novosti Press Agency

Blondin's crossings of Niagara Falls on a tightrope in the same year, rekindled the public interest in the work of the aerial gymnast and acrobat. Although the trapeze had never been seen before, ropedancing can be traced to ancient Greece. By the turn of the century, acrobats had had an extensive influence, although they never usurped the supreme position of the horse.

The human performer offers not the mystery of the exotic circus animal but the realization that a fellow human being, of the same substance as the spectator, can achieve astonishing discipline and mastery. The circus audience identifies itself with the performer as he or she attempts a feat of daring or skill. Completion of such feats as a double somersault to a four-high human pyramid fills the crowd with satisfaction and relief. A misstep on the tightwire seems to elicit greater distress in the spectators than in the performer. And the circus performance is not on film or tape but live: there can be no retake to erase an error. With evidence of human weaknesses and failures so prevalent everywhere, the circus is unusual in showing people at their best in physical achievement and skilled coordination. This realization comes to mind, for example, at the sight of the Wallendas, a family of high-wire artists originally from Germany, balancing three-high on bicycles on the high wire. For Americans it may have come from seeing petite Lillian Leitzel, born in Bohemia of a German circus family, who could pivot a hundred times on her shoulder socket in a maneuver called the "plange"; she spun from a rope like a pinwheel. Again, its source may have been the skill of Australian-born Con Colleano, "the Toreador of the Tight Wire," whose dance on the wire to a Spanish cadence thrilled American audiences for decades; or perhaps it was the grace and perfect timing of the Mexican trapeze artist Tito Gaona, who—even blindfolded—flawlessly performed the triple somersault from bar to catcher.

National traditions

Traditionally, certain nationalities tend to dominate specific areas of circus performance. Just as bareback riding long remained an English specialty and wild animal training a specialty of Germans and Americans, acrobatics and tumbling remain the realm of eastern Europeans. Asians excel in juggling and balancing acts and Mexicans in the flying trapeze. Although there are frequent exceptions to these tendencies, especially in recent years, these and other traits are identifiable. Such traditions may be related to the existence of circus families, whose specialties are passed on for several generations. To many people, such a life would appear to be a nomadic existence, without either the comforts of a permanent residence or the security of

permanent employment. On the other hand, to experience circus life is to see apparent disadvantages become sources of strength. The appeal of a traveling life is strong to many persons, as is the bond that is built between show people in overcoming common hardships.

**Clowns.** When the human capacity for amazement, thrill, and suspense approaches its limits, the circus unleashes its clowns to freshen the atmosphere, lighten the emotional load, and recondition the spectator's mind for the next turn. By tradition, there are several varieties of clowns, from the elegantly costumed whiteface clown, favoured in many European circuses, who appears rather severe and domineering; to the happy-go-lucky grotesque variety, whose exaggerated makeup and costumes are more outrageous and less predictable; to the woebegone, down-and-out "tramp" character, as popularized above all by the American Emmett Kelly. In 19th-century one-ring circuses it was usual for clowns to entertain audiences with songs and long monologues, in which they sometimes offered words of wisdom on politics and current events or quoted Shakespeare. More recently, especially in Soviet circuses, a number of clowns have attempted to strike out in new directions, abandoning traditional costumes and makeup and developing more natural characters. The great Soviet clown Oleg Popov became well known not only in the Soviet Union but also in Europe and America through his tours with the Moscow Circus. Wearing a minimum of makeup, he appeared in the ring with little to set him apart from the others except a wardrobe that was slightly unorthodox. Like other great comedians of the world, his mere appearance brought anticipatory laughter from the audience. Popov impersonated a rube character who is forever trying to mimic the legitimate performers. Frequently he almost succeeded, but only after sufficient bungling to make his performance a comedy. Actually, in areas such as balancing on the slack wire and juggling, he demonstrated professional abilities.

**Supporting attractions.** *Band music.* Behind it all is the circus band, setting the pace and mood and cuing the acts. The composition of circus music has, over the years, become an art of its own. Playing such characteristic forms as the march, the gallop, the fox-trot, the tango, or various national airs, the circus band maintains the tempo of each event. It signals the change of emphasis among the simultaneous presentations as one event after another is highlighted, maintaining the seemingly perpetual flow of music.

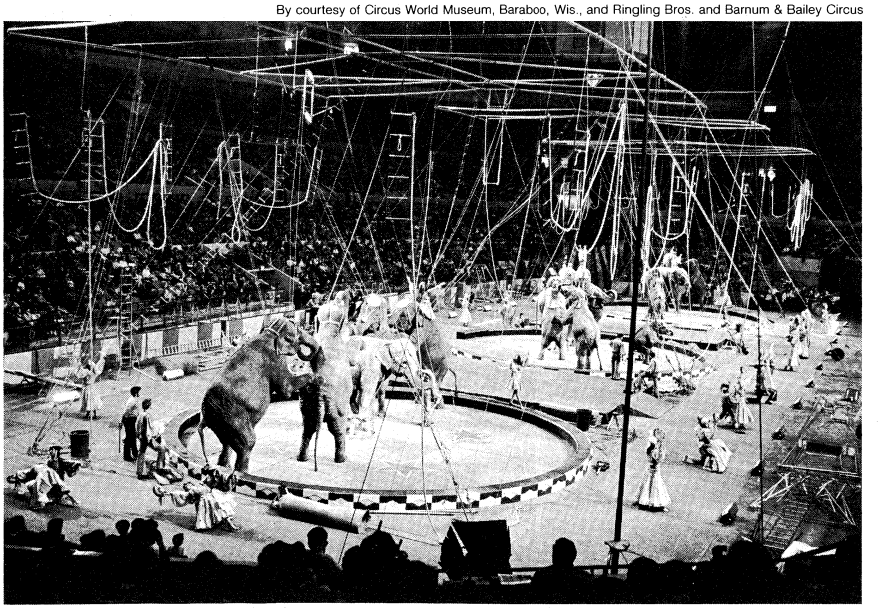
*The menagerie.* Exotic animals such as camels and lions, brought to America by enterprising ship captains and sold to showmen, had been shown in the colonies in the 18th century. About 1815 Hackaliah Bailey, of Somers, N.Y., toured New England with his elephant, "Old Bet," whose success inspired Bailey's farmer-neighbours to set out with menageries of their own. In a very short time, there were noteworthy traveling collections of wild animals, such as the Zoological Institute of June, Titus & Angevine and Van Amburgh's Menagerie, which also set up permanent establishments in the larger cities. Similar developments occurred in England, leading to such famous traveling shows as Wombwell's Menagerie. Subsequently, circuses and menageries were combined. By the time American circuses achieved their massive character in the 1870s, the menagerie was a major feature, and it remained so through the 1940s. The circus menagerie was exhibited in a separate tent through which the public passed on its way to the main performance. The beautifully carved wagons that held the wild-animal cages lined the perimeter of the tent. The elephants were chained in position, and the "lead stock," including uncaged animals such as camels, llamas, bison, and zebras, were picketed together. The larger circuses had extensive collections, including rhinoceroses and giraffes in their own portable corrals.

*The sideshow.* In 1871, after he had already achieved success through his famous New York museum, P.T. Barnum entered into a partnership with William C. Coup and the former clown Dan Castello to found "P.T. Barnum's Museum, Menagerie, and Circus." It offered several attractions borrowed from his earlier museum, from which evolved another major feature of American circuses—the sideshow, or annex. Typically this included human "abnormalities," such as fat ladies, giants and midgets, "armless wonders," and four-legged girls; illusions and magicians; automatons and curious inventions; and various works of art, among them Hiram Powers' titillating nude statue "The Greek Slave." Housed in its own tent, the sideshow typically was fronted by giant banners or panels illustrating the marvels inside. A unique and vital element of the sideshow was the barker, whose foghorn voice and unceasing patter attracted the public to the show.

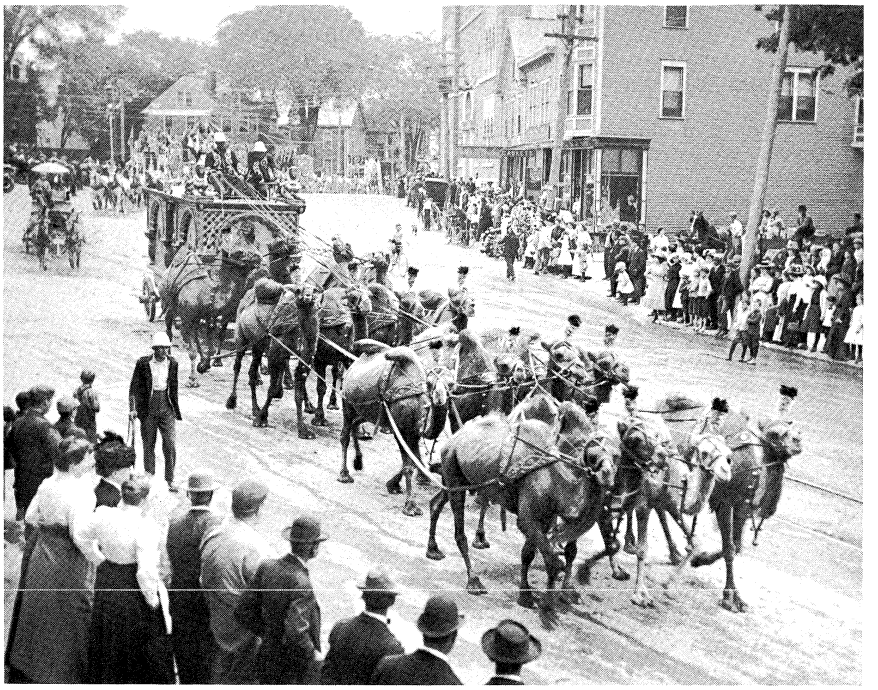
## COMMERCIAL AND TECHNICAL ASPECTS

From 1840 onward, circus combines and amalgamations became widespread in the United States, and partnerships were made and broken with bewildering frequency. In 1907, following the death of Barnum's final partner, James A. Bailey, the Ringling brothers, who had started out in the 1880s, bought Barnum & Bailey Circus and continued to run it as a separate show, finally combining it with their own circus in 1919 to form the concern that still flourishes as Ringling Bros. and Barnum & Bailey. In 1929 the Ringlings bought the American Circus Corporation, a syndicate comprising five circuses, with a view

By courtesy of Circus World Museum, Baraboo, Wis., and Ringling Bros. and Barnum & Bailey Circus



Elephants in three-ring performance from Ringling Bros. and Barnum & Bailey Circus, 1970.



A 16-camel team in a Ringling Bros. Circus parade, 1911.
By courtesy of Circus World Museum, Baraboo, Wis., and Ringling Bros. and Barnum & Bailey Circus

to ensuring their supremacy in the field. Such attempts at monopolies were never the case in Europe, however, where families tended to split up rather than combine, so that more than one circus may bear the name Pinder, Fossett, Ginnett, or Sanger.

The mid-19th century was a period of great technical development in the United States. Spencer Q. Stokes invented the derricklike apparatus for training trick riders, which still goes by the name of the American Riding Machine. At one time his circus traveled by riverboat, though he did not give performances on a showboat as did "Dr." Gilbert Spaulding (1811–80). This former chemist from Albany, N.Y., is said to have invented quarter poles (which support the canvas roof of the big top between the central king poles and the side poles) and was also one of the first circus proprietors to experiment with railroad transport.

**The circus train.** Before 1872 most itinerating circuses moved from town to town by horse and wagon, a form of transport that necessarily limited their size and the distances they could cover in a given season. In the spring of that year, Barnum and his partners loaded their show onto 65 railroad cars, thereby giving birth to the age of the giant railroad circuses. Barnum's own "Greatest Show on Earth" eventually traveled on three separate trains, going distances of 100 miles or more in a single night; and in the 20th century Ringling Bros. and Barnum & Bailey once crisscrossed the country aboard four trains comprising 107 70-foot cars.

Railroads were more frequently used in the United States than in any other country. In Europe rail transport for circuses was never very popular, and, although a few attempts at traveling by rail were made, it was not until the second quarter of the 20th century that rail travel came into regular use by any European show. In Latin America and East Asia it was used, when available, after 1900. William C. Coup introduced the end loading of circus trains, in which the gaps between flatcars were bridged by iron plates and each wagon, fully loaded, was pushed down the length of the train to its assigned place. A model of logistic efficiency, circus methods led the way to the creation of the modern system of rail-truck freight handling.

**The parade.** The free circus street parade evolved in the United States as a triumphal entry into town by each overland circus caravan. By the 1840s it was recognizable as a parade with bandwagons and trappings. It was the English, however, who created the finest and most ornately carved circus parade wagons. In 1865 Seth Howes imported to the United States several English wagons, which prompted an upsurge of interest in circus parades. These processions, which wound their way through the town back to the circus field ("lot" in the United States, "tober" in Britain), were a great feature of British tenting circuses, particularly that of "Lord" George Sanger (1825–1911), who also owned Astley's Amphitheatre from 1871 until it was demolished in 1893. (Once Sanger even tacked his parade onto the end of a military escort accompanying Queen Victoria across London.)

In the United States the circus parade was the climax of a highly systematized publicity campaign to arouse interest in the circus during its brief appearance at any one place.

The first steps were taken by the general agent of the circus, who made arrangements for a city license, showgrounds, railroad routing, and other details. The impact of organized publicity hit the community when the first advance car arrived in town two or three weeks before show day. In larger shows as many as four advance advertising cars involving 60 to 80 men were used. A concentrated mass publicity campaign then unfolded. Billposters, lithographers, and banner men plastered the town and its environs with tens of thousands of square feet of "paper." The programmer recruited youngsters to scatter thousands of printed heralds, multi-paged couriers, handbills, and "rat sheets" (used to disparage other circuses when they happened to be in the same vicinity at the same time). Press agents turned the power of the press to the purposes of the circus. Other agents attended to such matters as water, food, fuel, feed, and preparations for showgrounds and railroad crossing.

Heralds, handbills, and "rat sheets"

On circus day itself the train arrived with its stockcars, perhaps with animals probing outside openings, and a long line of flatcars loaded with red baggage wagons, pole wagons, bandwagons, tableaus, chariots, the steam boiler wagon, and canvas-covered wild-animal cages. In a large circus there would be several trains. The initial train would be the "flying squadron" bearing the cook house, horse tents, menagerie, and steam calliope, having departed while the crowds were still watching the last performance in the previous town. At intervals would follow the "canvas train" and "lumber train" bearing construction materials in the order they would be needed. Wagons were pulled by baggage horse teams to the showgrounds, directed by arrows chalked on posts and trees by an advance man.

The showgrounds became a scene of organized confusion: acres of canvas and a forest of poles were assembled before swarms of curious people—affectionately called

"lot lice" by showmen. Then "parade call" was trumpeted, and performers, musicians, animal attendants, wardrobe crews, drivers, and brakemen assembled for the grand free street parade that was usually scheduled for 11:00 AM. Following the bugle brigade heralding the grand event, there was a long procession of horses, flag bearers, bands on magnificent wagons, allegorical tableaus, clowns, knights in armour, beautiful ladies on steeds, Roman chariots, chimes, bells, a band organ, cage after cage of wild animals (some open to view and others closed to prompt curiosity), cowboys, Indians, and a long line of highly caparisoned elephants shuffling along trunk-to-tail. The traditional finale to the circus parade was the Pied Piper of the circus, the steam calliope: 32 steam whistles, keyboard operated, powered by coal fire and boiler, hissing steam and smoke, and blasting out such favourites as "Go Tell Aunt Rhodie" and "The Sidewalks of New York." Under steam pressure of from 80 to 100 pounds per square inch, its tones carried four miles against a breeze. After two shows daily and the teardown, which took place at night, the wagons and teams followed flares to the train, where they rolled back onto the flatcars to disappear into the night.

The calliope

**The big top.** The origin of the circus tent is believed to have occurred in 1825 on the itinerating show of the American J. Purdy Brown. At first tents were very small, housing one ring and a few hundred seats. In their heyday, however, after circuses had begun to exhibit simultaneously in two (1872) and then three (1881) rings, the great American tented circuses played under what could be called canvas coliseums. The big top of the Ringling Bros. and Barnum & Bailey Circus covered the equivalent of a hectare, or more than two acres; supported by several centre poles that rose 65 feet (nearly 20 metres), it sheltered from 10,000 to 12,000 people. From portable bleachers, crowds could watch as many as seven rings and "stages" (platforms set between and at the ends of the three rings) at the same time, in addition to the aerial acts above. A typical big top of standard American circuses such as Hagenbeck–Wallace or Cole Bros. would be somewhat smaller, housing about 5,000 seats. European and British circuses generally retained the one-ring format, although their programs were often of the highest calibre and their tents might also seat as many as 5,000 spectators. In order to maintain the one-ring design while expanding the area beneath the tent, the European tent was designed with the four centre poles forming a square instead of a single-file line as with American big tops.

One or three rings

Just as some European circuses, such as Italy's Circo Americano, attempted to emulate the American pattern and exhibited simultaneously in three rings, in the second half of the 20th century there was increasing interest among American showmen in adopting, or reverting to, the more intimate, less confusing one-ring format. The Big Apple Circus of New York City, one of the most elegant shows ever seen in the United States, performed in a single ring inside a tent that seated only about 1,600 spectators.

**Winter quarters.** In the United States and to some extent in Europe, circuses annually retired to winter quarters to refurbish for another season. Among the cities that became identified as winter-quarters towns were Peru, Ind., which sheltered Hagenbeck–Wallace and other shows; Baraboo, Wis., the winter home for the circus of the Ringling Bros. and their cousins the Gollmar Bros.; and Bridgeport, Conn., which for nearly 50 years served as headquarters for Barnum's "Greatest Show on Earth," until the Ringlings moved the operations of the combined show to Sarasota, Fla., in 1927.

### VARIATIONS ON THE CIRCUS THEME

**Wild West shows.** Wild West shows, which were sometimes attached to circuses, emphasized displays and events of America's old West. A Wild West show usually presented its exhibition in a large open field surrounded by bleachers protected by a canvas canopy. Typically, such shows featured American Indian ceremonies; cowboys who engaged in broncobusting, roughriding, roping, and sharpshooting; and dramatic representations of life on the frontier and Indian attacks on wagon trains or stagecoaches, whose occupants were saved by the timely arrival of the U.S. cavalry. Although Barnum had added such an exhibition to his circus as early as 1876, the credit for establishing the Wild West show as a separate entertainment goes to the former cavalry scout William F. "Buffalo Bill" Cody (1846–1917) and his partner W.F. "Doc" Carver, who launched their own Wild West & Congress of Rough Riders of the World in 1883. Pawnee Bill's Wild West and Miller Bros. 101 Ranch Real Wild West were prominent competitors of Buffalo Bill over the years. The famous riflewoman Annie Oakley, "Little Miss Sure-Shot," gained her fame as a star of Wild West shows. Many film stars were associated with them, too, including Tom Mix and Will Rogers. The last Wild West show was Colonel Tim McCoy's Wild West of 1938.

**Carnivals.** The public sometimes confuses circuses with carnivals, but they are quite different, even though both are traveling amusements. Instead of presenting a unified performance in one large enclosure, a carnival presents a midway lined with independent tented sideshows, concessions, special attractions, and rides or thrill events in which the patrons are participants, not spectators. Carnivals usually stay a full week or longer in any one location. In their modern American form, carnivals originated around the end of the 19th century when such rides as the Ferris wheel and steam-operated merry-go-round (carousel) became popular. Carnivals became standard features of state and county fairs in the United States.

Ferris wheels and carousels

### THE STATE OF THE ART

**Early 20th-century changes.** Circuses in Europe began to encounter economic difficulties in the period after World War I, when foreign travel was inhibited by passport formalities, customs duties, quarantine restrictions, and currency regulations. For large companies with much equipment the difficulties were particularly acute. In order to evade inflation and crisis at home, one German circus toured South America in 1923 and 1934.

In the 1920s the circus in Britain declined. It was spectacularly revived by Bertram Mills, a coach builder, who introduced the greatest international circus stars to the British public at Olympia, London, and ran the only tenting circus to travel by rail in England.

Meanwhile, in the United States the circus was also encountering difficulties. During the Great Depression the Ringling empire collapsed; and the federal government, in order to give employment to out-of-work performers, organized a Works Progress Administration Circus—the only example of a state-run circus ever seen in the United States. In 1944 a disastrous fire in the Ringling big top, which took 168 lives and left hundreds of other spectators burned and injured, added to the woes of American circus proprietors. By the 1950s, faced with stiff competition from motion pictures and especially television, American circuses were generally declining; and in 1956 John Ringling North, citing economic and labour problems, announced that the "Greatest Show on Earth" would abandon its big top and in the future perform only in permanent exhibition halls and sports arenas. For many this announcement signaled the imminent demise of the circus.

By then, too, the spectacular street parade had long since disappeared, and most other American circuses were smaller and again traveling by road. The era of the truck show or motorized circus had dawned in the United States in 1918 with the short-lived Great United States Motorized Circus. Motor transport proved successful in the 1920s for Downie Bros. Circus, Seils-Sterling Circus, and some others. At first there was a tendency within the profession to belittle circuses transported by truck, but the 1930s produced several outstanding shows that moved by motor transport, including one owned by Tom Mix. Motorization enabled the canvas-tent circus to survive and continues to do so, only Ringling Bros. still moving the bulk of its show by rail.

Moving by truck

Circuses never entirely discontinued the use of indoor facilities for their exhibitions. Especially in Europe, circuses continued to perform in buildings throughout the heyday of the canvas shows—at the Circus Schumann in Copenhagen, for example, and at the venerable Cirque d'Hiver in Paris, which dates from 1852. Even the greatest

tent shows opened their seasons in metropolitan buildings, enabling them to expand their seasons into periods of weather not conducive to canvas operations. Such was the case with the annual opening of Bertram Mills Circus at Olympia in London, Ringling Bros. Circus at Chicago's Coliseum, and Barnum & Bailey at Madison Square Garden in New York City. In the Soviet Union, where the circus is regarded as an art form and receives lavish state support, nearly every city of any size boasts a permanent circus building. The New Moscow Circus possesses two additional rings, for water and ice shows, that can be separately raised by a mechanism from below to replace the circus ring as occasion demands. Most major American and western European circuses continue to tour under canvas, however, and even in the Soviet Union and eastern Europe tenting circuses are by no means uncommon.

**Preservation of the past.** Since the early part of the 20th century there has been a growing interest in preserving the circus's colourful past. Pursuing this end are collectors and historians throughout the world, as well as associations, including the Circus Fans Association (England and the United States), the Club du Cirque (France), the Society of Friends of the Circus (West Germany and Austria), and the Circus Historical Society and the Circus Model Builders Association (United States). Sarasota, Fla., developed two circus museums, the Circus Hall of Fame (closed in 1980) and the Ringling Museum of the Circus. At Baraboo, Wis., the extensive Circus World Museum, including a Circus Library and Research Center, is operated by the State Historical Society of Wisconsin. Each July 4, for a number of years, a train of 24 circus railroad cars transported more than 70 restored circus parade wagons from the Circus World Museum to Milwaukee, where they were joined by hundreds of privately owned baggage horses for a parade through the streets of the city. At Bridgeport, Conn., the Barnum Museum features exhibits related to the circus and the life of the world-famous showman. Most major European collections of circus materials remain in private hands, although important holdings are available to researchers at such institutions as the British Library in London, the library and museum of the Paris Opera, and the Museum of Circus Art attached to the Leningrad Circus.

**Contemporary developments.** Despite anxiety over its future during the 1950s and '60s, and the disappearance or reorganization of such stalwarts as the Bertram Mills and Ringling Bros. circuses during the same period, the circus continues to give every indication of flourishing. During the mid-1980s, for instance, more than 30 major circuses toured the United States and Canada, while dozens more—some lasting an entire season, some for only a few weeks or for single engagements sponsored by local groups such as the Shriners—also performed. Beginning in 1969 the Ringling organization fielded two complete circuses, the "red" and "blue" units, which perform for 11 months each year and between them annually visit nearly every major American and Canadian city. The story is much the same in western Europe, although in Great Britain circus owners have often been harassed by "antis," activists who are convinced that cruelty is involved in the training of circus animals and who consequently agitate to have such acts banned. The circus and its traditions have been preserved and renewed above all, however, in the Soviet Union and eastern European countries, whose state-supported artists and circus companies are increasingly seen in the West. Among the distinguishing characteristics of their acts are the originality of the apparatus, the costumes, and the presentation. In the high-wire act of the Voljankis, for example, the wire changed from being horizontal to an oblique angle, while the tension was maintained; the Koch sisters performed on a giant semaphore arm that revolved slowly as they balanced on the outside edge.

The Soviet circus

In the Soviet Union a professional school for the training of circus artists has been associated with the Moscow Circus since the 1920s. After four years of rigorous course work, graduates are assigned to the nearly 100 circuses that perform around the country; they are better paid then the average Soviet citizen and are entitled to retire after 20 years. In 1985 the French Ministry of Culture and Communication opened a four-year professional school for artists, the High School of Circus Arts (l'Ecole Supérieure des Arts du Cirque), at Châlons-sur-Marne. While nothing comparable exists in the United States, the Big Apple Circus of New York City runs a training program for circus artists, and Ringling Bros. annually sponsors a "Clown College" at its Venice, Fla., winter quarters. A number of secondary schools and universities around the United States also offer instruction in circus techniques. Florida State University continues to sponsor its "Flying High Circus" (begun in 1947), whose performers are drawn exclusively from the student body. In 1974 Prince Rainier III established an annual circus festival at Monte-Carlo. For several days each winter top acts and performers from circuses around the world (including North America and Asia) compete before a jury of knowledgeable experts. Prizes are awarded, the highest being the Clown d'Or ("Golden Clown"), which among circus artists has come to be regarded in the same light as the Hollywood Oscar.

For all its emphasis on tradition and discipline, the circus has always been a dynamic, growing institution. Clowns, acrobats, and the ring itself will probably forever be a part of it. Yet a willingness to experiment and innovate—whether with various modes of transport; newfangled tents and, eventually, air-conditioned sports arenas; or the introduction of exotic animals, the flying trapeze, and any number of other novel acts—has remained one of the most distinguishing characteristics of the circus. Essentially a nonverbal entertainment, wholesome and almost never political, its appeal easily bridges national boundaries and is virtually universal and inexhaustible. One might also term the circus the most democratic of entertainments, since, no matter what the spectator's age or level of sophistication, there is always something of interest to observe or laugh at, be it the hilarious farce of a group of clowns, the fearlessness of a diminutive trainer inside a cage of menacing polar bears, or the elegant combinations of somersaults, pirouettes, and twisters performed by artists on the flying trapeze. The sheer spectacle is itself a treat, mingling with the spirited music from the band, the smells of animals and peanuts, and the excitement of being crowded into a tent or building with thousands of other spectators, all of them intent on enjoying themselves. For these and other reasons the circus remains preeminently a "live" entertainment, whose pleasures and nuances can never be fully captured on film or television. For these reasons, too, the circus will undoubtedly continue to endure, thrilling people of all nations for ages to come.

(R.L.P./A.D.H.C./A.H.Sa.)

**BIBLIOGRAPHY.** Comprehensive works include MONICA J. RENEVEY (ed.), *Le Grand livre du cirque,* 2 vol. (1977); and HENRY THÉTARD, *La Merveilleuse histoire du cirque,* new ed., rev. and enl. by L.-R. DAUVEN (1978). For studies of the circus in the United States, see STUART THAYER, *Annals of the American Circus,* 2 vol. (1976–86); CHARLES PHILIP FOX and TOM PARKINSON, *The Circus in America* (1969); DAVID LEWIS HAMMARSTROM, *Behind the Big Top* (1980); JOHN BURKE, *Buffalo Bill: The Noblest Whiteskin* (1973); and DON RUSSELL, *The Wild West: or, A History of the Wild West Shows* (1970). Other countries are discussed in DAVID LEWIS HAMMARSTROM, *Circus Rings Around Russia* (1983); and A.H. SAXON, *The Life and Art of Andrew Ducrow & the Romantic Age of the English Circus* (1978). Individual acts and attractions are described in ANTONY HIPPISLEY COXE, *A Seat in the Circus,* rev. ed. (1980); A.H. SAXON, *Enter Foot and Horse: A History of Hippodrama in England and France* (1968); and BILL BALLANTINE, *Clown Alley* (1982), on the history of the Clown College and its personalities. The activities of famous impresarios are presented in GEORGE SANGER, *Seventy Years a Showman* (1910, reissued 1966); GEORGE SANGER COLEMAN and JOHN LUKENS, *The Sanger Story: The Story of His Life with His Grandfather "Lord" George Sanger* (1956, reissued 1974); CYRIL BERTRAM MILLS, *Bertram Mills Circus: Its Story* (1967, reissued 1983); HENRY RINGLING NORTH and ALDEN HATCH, *The Circus Kings: Our Ringling Family Story* (1960); P.T. BARNUM, *Struggles and Triumphs: or, The Life of P.T. Barnum Written by Himself,* ed. by GEORGE S. BRYAN, 2 vol. (1927), the best edition of the famous autobiography; and NEIL HARRIS, *Humbug: The Art of P. T. Barnum* (1973, reprinted 1981). For further research on the subject, see RAYMOND TOOLE-STOTT, *Circus and Allied Arts: A World Bibliography, 1500–1970,* 4 vol. (1958–71).

(A.H.Sa.)

# Cities

The name city is given to certain urban communities by virtue of some legal or conventional distinction. It also refers to a particular type of community, the urban community, and its generic culture, often called "urbanism." In legal terms, in the United States, for example, a city is an urban area incorporated by special or general act of a state legislature. Its charter of incorporation prescribes the extent of municipal powers and the frame of local government, subject to constitutional limitation and amendment. In common usage, however, the name is applied to almost every American urban centre, whether legally a city or not, and without much regard to actual size or importance. In Australia and Canada, city is a term applied to the larger units of municipal government under state and provincial authority respectively. New Zealand has followed British precedent since the abolition of the provinces in 1876; the more populous towns are called boroughs under the Municipal Corporations Act of 1933 and earlier legislation. In the United Kingdom itself, city is merely an official style accorded towns either in their historical identity as episcopal sees or as the beneficiaries *honoris causa* of a special act of the crown (the first town so distinguished was Birmingham in 1889). Except for the ancient City of London (an area of about 677 acres in central London under the jurisdiction of the lord mayor), the title has no significance in local government in the United Kingdom. In all the other countries of the world, the definition of city similarly follows local tradition or preference.

City government is almost everywhere the creation of higher political authority, state or national. Some European countries have adopted general municipal codes which permit centralized administrative control over subordinate areas through a hierarchy of departmental prefects and local mayors. Socialist countries also employ a hierarchical system of local councils that correspond to, and are under the authority of, governing bodies at higher levels of government. In English-speaking countries, devolution of powers to the cities occurs through legislative acts that delegate limited self-government to local corporations.

As a type of community, the city may be regarded as a relatively permanent concentration of population, together with its diverse habitations, social arrangements, and supporting activities, occupying a more or less discrete site, and having a cultural importance that differentiates it from other types of human settlement and association. In its elementary functions and rudimentary characteristics, however, a city is not clearly distinguishable from a town or even a large village. Mere size of population, surface area, or density of settlement are not in themselves sufficient criteria of distinction, while many of their social correlates (division of labour, nonagricultural activity, central-place functions, and creativity) characterize in varying degree all urban communities from the small country town to the giant metropolis.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 524 and 542.

This article is divided into the following sections:

The history of cities 425
- Initial requirements for urban development 425
- Early cities 426
  - Ancient world
  - Autonomous and dependent cities
- Medieval and early modern era 426
  - Medieval cities: from fortress to emporium
  - The city and the nation-state
- Industrialization and the modern world 428

Urban planning 428
- The development of urban planning 428
  - Early history
  - 19th century
  - 20th century
  - Goals of modern urban planning
- Planning and government 430
  - Zoning and subdivision controls
  - Large-scale planning
  - Planning jurisdictions

Modern city government 431
- Basic characteristics of city government 431
  - Jurisdictions of cities
  - City services
- Types of city government 432
  - Decentralized city government in federal systems
  - Decentralized city government in unitary systems
  - The Napoleonic supervisory system
  - The integrated system
  - Distribution of the various systems

Bibliography 435

## The history of cities

### INITIAL REQUIREMENTS FOR URBAN DEVELOPMENT

It was no accident that the earliest of man's fixed settlements are found in the rich subtropical valleys of the Nile, the Tigris, the Euphrates, the Indus, and the Yellow rivers or in such well-watered islands as Crete. Such areas provided favourable environmental factors making town living relatively easy: climate and soil favourable to plant and animal life, an adequate water supply, ready materials for providing shelter, and easy access to other peoples. Although man with ingenuity has been able to utilize almost any environment for town living, environments favourable to the production of food and shelter and ease and comfort of living clearly possessed advantages for the beginnings of urban life.

Environmental characteristics of early settlements

A distinguished historian, Ralph E. Turner, has suggested that various preurban developments made possible the technology and organization permitting city life. These included psychological elements such as recognition of "in-group" versus "out-group" interests; the notion of a universe, even if mysterious, that could be controlled; and belief in the existence of a soul. The in-group and out-group differentiation provided a basis for respect for the rights of others and for life, property, and family values. The notion that man could control the world in which he lived was of great importance, even if the methods of control were primitively based on magic and religion. The belief in a soul helped make life on Earth more acceptable, even if hard, for life became then only an incident in a long journey.

Preurban developments that paved the way for urban life also included such factors as traditionalism, a power structure, and a form of economic as well as social organization. Traditionalism lay in the acceptance and transmission of what had worked in the life of the group and was therefore "right" and to be retained. Some form of power structure involving subordination was necessary, for leadership was a vital element in urban living in that it was essential to the performance of such vital functions as sustenance, religious practices, social life, and defense. Also prerequisite to group life were new economic and social institutions and groupings such as property, work, the family, a system for distribution of commodities and services, record keeping, police for internal security, and armed forces for defense.

New value orientations and ideologies may also have affected the course of urbanization, though their importance is still highly conjectural. There are those who have felt that urbanization depended on a new outlook; it meant that people had become more rationalistic (and less mystical); it meant that, for purposes of building, they were more willing and able to defer immediate for more desirable later gratification; it meant more emphasis on achievement and success as distinguished from status and prestige; it meant a cosmopolitan as distinguished from a parochial outlook; and it meant that relations between people were more ordered, impersonal, and utilitarian, rather than only personal and sentimental.

## EARLY CITIES

**Ancient world.** About 10,000 years ago in the Neolithic Period, man achieved relatively fixed settlement, but for perhaps 5,000 years such living was confined to the semipermanent peasant village—semipermanent because, when the soil had been exhausted by the relatively primitive methods of cultivation, the entire village was usually compelled to pick up and move to another location. Even when the village prospered in one place and the population grew relatively large, the village usually had to split in two, so that all cultivators would have ready access to the soil.

The evolution of the Neolithic village into a city took at least 1,500 years—in the Old World from 5000 to 3500 BC. The technological developments making it possible for man to live in urban places were, at first, mainly advances in agriculture. Neolithic man's domestication of plants and animals eventually led to improved methods of cultivation and stock breeding and the proliferation of the crafts, which in turn eventually produced a surplus and freed some of the population to work as artisans, craftsmen, and service workers.

Transportation development

As human settlements increased in size, by reason of the technological advances in irrigation and cultivation, the need for improving the circulation of goods and people became ever more acute. Pre-Neolithic man leading a nomadic existence in his never-ending search for food moved largely by foot and carried his essential goods with the help of his wife and children. Neolithic man, upon achieving the domestication of animals, used them for transportation as well as for food and hides. Then came the use of draft animals in combination with a sledge equipped with runners for carrying heavier loads. The major technological achievement in the early history of transportation, however, was obviously the invention of the wheel, used first in the Tigris–Euphrates Valley about 3500 BC and constructed first with solid materials and only later with hubs, spokes, and rims. Wheels, to be used efficiently, required roads, and thus came road building, an art most highly developed in ancient times by the Romans. Parallel improvements were made in water transport—with rafts, dugouts, the Egyptian reed float, eventually wooden boats, and of course canals used for both navigation and irrigation. (Ed.)

By 3500 BC urban populations were distinguished by literacy, technological progress (notably in metals), social controls, political organization, and emotional focus (formalized in religious-legal codes and symbolized in temples and walls). Such places, dated by historical means, existed on the Sumerian coast at Ur and in the Indus Valley at Mohenjo-daro during the 3rd millennium and, before 2000 BC, had also appeared in the Nile and Wei-ho valleys. Cities proliferated along overland trade routes from Turkestan to the Caspian and then to the Persian Gulf and eastern Mediterranean. Their economic base in agriculture (supplemented by trade) and their political-religious institutions made for an unprecedented degree of occupational specialization and social stratification. From central vantage points, cities already gave some coherence and direction to life and society in their hinterlands.

The growth of cities, however, was by no means the inevitable outcome of a succession from primitive life to civilization. As S. Piggott pointed out in "Role of the City in Ancient Civilizations" (in *Metropolis in Modern Life,* ed. by E.M. Fisher [1955]), an alternative and, in some ways, inimical type of community had arisen in the steppe-lands of Asia based upon animal husbandry: the nomadic encampment. Like their urban contemporaries, the nomads were no longer "primitive" men. In addition to pastoralism, they had developed great oral traditions, abstract art styles, and numerous crafts, albeit no formal architecture. Led by warrior chiefs, these self-sustaining migratory peoples encroached upon the settled agricultural-trading areas to the south.

During the 2nd millennium the Indus civilization was engulfed by an onslaught of Aryan nomads, while other peoples, using horses and chariots, penetrated the urban heartland from Mesopotamia to Egypt. In these circumstances of prolonged upheaval, survival required the perfection of warlike arts and predatory supply systems, which transformed the urban communities into paramilitary states—*e.g.,* the Hittite, Egyptian, and Mycenaean empires. Citizenship, though still a ceremonial service, was increasingly associated with the bearing of arms. After 1200 BC even the city-empires (a city-camp hybrid) lapsed into chaos and disorder until the lifting of the Hellenic "dark ages" during the 8th century BC and the transplanting of the syncretic city-state beyond the eastern Mediterranean by Phoenicians and Greeks.

**Autonomous and dependent cities.** The heterogeneous peoples that created the Greco-Roman world inherited a technological and nonmaterial culture from southwestern Asia which helped mollify barbarism and nourish the growth of cities. Their trading colonies, from the Crimea to Cadiz, eventually brought the entire Mediterranean within the orbit of civilization. It was in the Greek city-state, or *polis*, however, that the city idea reached its peak. Originally a devout association of patriarchal clans, the *polis* came to be a small self-governing community of citizens in contrast to the Asian empires and nomadic hordes. For citizens, at least, the city and its laws constituted a moral order symbolized in magnificent buildings and public assemblies. It was, in Aristotle's phrase, "a common life for a noble end."

The Greek *polis*

When the old exclusive citizenship was relaxed and as new commercial wealth surpassed that of the older landed citizenry, social strife at home and rivalry abroad gradually weakened the common life of the city-republics. The creativity and variety of the *polis* gave way before the unifying forces of king-worship and empire epitomized by Alexander the Great and his successors. To be sure, many new cities were planted between the Nile and the Indus through which the amenities and forms of city-culture were carried back to the east, but the city itself ceased to be an autonomous body politic and became a dependent member of a larger political-ideological whole.

The Romans, who fell heirs to the Hellenistic world, transplanted the city into the technologically backward areas beyond the Alps inhabited by pastoral-agricultural Celtic and Germanic peoples. But, if Rome brought order to civilization and carried both to barbarians along the frontier, it made of the city a means to empire (a centre for military pacification and bureaucratic control) rather than an end in itself. The enjoyment of the imperial Roman peace entailed the acceptance of the status of *municipium*—a dignified but subordinate rank. Initiatives passed to the centre; and, in the east, the culture of provincial cities became imitative, their politics trivial. They contributed little to the larger economic life beyond the needs of their social elites and the payment of taxes; they tapped the surpluses created by local agriculture and trade in rents and tribute. As Roman citizenship became more universal and formal, the idea of public duty gave way to private ambition. Municipal functions atrophied; and, except for their fiscal duties, it was in a passive role that the city survived into the Byzantine era.

Roman citizenship and provincial cities

## MEDIEVAL AND EARLY MODERN ERA

**Medieval cities: from fortress to emporium.** In Latin Europe neither political nor religious reforms could sustain the Roman regime. The breakdown of public administration and the breach of the frontier led to a revival of parochial outlook and allegiance, but their focus was not upon the city. Community life now centred on the fortress (*burgum*) or castle (*castellum*); the term city (*civitas*) was

attached to the precincts of the episcopal throne, as in Merovingian Gaul.

Early medieval society was a creation of camp and countryside to meet the local imperatives of sustenance and defense. With Germanic variations on late Roman forms, communities were restructured into functional estates, each of which owned formal obligations, immunities, and jurisdictions. What remained of the city was comprehended in this feudal-manorial order, and the distinction between town and country was largely obscured when secular and ecclesiastical lords ruled over the surrounding counties (*comté, Grafschaft*) as the vassals of mock emperors or barbarian kings. Social ethos and organization enforced submission to the common good of earthly survival and heavenly reward; the true city, *civitas Dei,* was not of this world. The attenuation of city life in most of northern and western Europe was accompanied by provincial separatism, economic isolation, and religious otherworldliness. Not before the cessation of attacks by Magyars, Norsemen, and Saracens did urban communities again experience sustained growth.

Recovery after the 10th century was not confined to the city or to any one part of Europe. The initiatives of monastic orders, seigneurs, or lords of the manor, and merchants alike fostered a new era of increased tillage, enlarged manufacture, money economy, the growth of rural population, and the founding of "new towns," as distinguished from those "Roman" cities that had survived from the period of Germanic and other encroachments. In almost all the medieval towns the role of the merchant was central: his needs and aspirations had a catalytic effect and, largely as a consequence of mercantile enterprise in the long-distance staple trade, cities were to flourish once more. Under commercial stimulus, feudal obligations were relaxed and European society was made over anew by the city and the marketplace in pursuit of self-government and economic gain.

Before the year 1000 contacts with rich Byzantine and Islāmic areas in the Levant had revitalized the mercantile power in Venice, which commanded the profitable route to the Holy Land during the Crusades. Meanwhile, merchant communities had attached themselves to the more accessible castle towns and diocesan centres in northern Italy and on the main travelled routes to the Rhineland and Champagne. They later appeared along the rivers of Flanders and northern France and on the west–east road from Cologne to Magdeburg.

Rise of civic autonomy

It was no coincidence that the 12th and 13th centuries, which saw the founding of more new towns than any time between the fall of Rome and the Industrial Revolution, also witnessed a singular upsurge toward civic autonomy. Throughout western Europe, towns acquired various kinds of municipal institutions loosely grouped together under the designation "commune." Broadly speaking, the history of the medieval towns is that of the merchant elites seeking to free their communities from lordly jurisdiction and to secure their government to themselves. Wherever monarchical power was strong, they had to be content with a municipal status, but elsewhere they created city-states. Taking advantage of renewed conflict between popes and emperors, they allied with local nobility to establish communal self-government in the largest cities of Lombardy, Tuscany, and Liguria. In Germany the city councils sometimes usurped the rights of higher clergy and nobility; Freiburg im Breisgau obtained its exemplary charter of liberties in 1120. The movement spread to Lübeck and later to the net of Hanse towns on the Baltic and North seas, touching even the Christian "colonial" towns east of the Elbe–Saale rivers. In the 13th century the "Great Towns" of Bruges, Ghent, and Ypres, creditors of the counts of Flanders, virtually governed the entire province. In France, revolutionary uprisings, directed against nobility and clergy, sometimes established free communes, but most communities were perforce content with a franchise from their sovereign more limited than those enjoyed by English boroughs under the Norman Conquest. Finally, the corporate freedom of the towns brought emancipation to individuals. When bishops in the older German cities treated newcomers as serfs, the emperor Henry V affirmed the principle *Stadtluft macht frei* ("City air brings freedom") in charters for Speyer (or Spires) and Worms; "new towns" founded on the lands of lay and clerical lords offered freedom and land to settlers who took up residence for more than "a year and a day." In France the *villes neuves,* or "new towns" (*e.g.,* Lorris), and *bastides* (*e.g.,* Montauban) likewise conferred rights on servile persons.

In the 14th century, the urban movement subsided as Europe entered on a period of political anarchy and economic decline that did not much abate before the 16th century. At a time when local specialization and interregional exchange required more liberal trade policies, craft protectionism and corporate particularism in the cities tended to hobble the course of economic growth. The artisan and labouring classes, moreover, now challenged the oligarchical rule of the wealthy burghers and gentry, disrupted local government, and ultimately destroyed the basis of civic autonomy: prolonged social warfare led to "popular" despotisms and fiscal bankruptcy. Visitations of plague, fanatical crusades against heresy, and Turkish encroachments on the routes to Asia worsened conditions in town and country alike. Europe turned inward upon itself; and, except for a few large centres, activity in the marketplace was depressed: the cities surrendered their liberties and their population. These centuries of decline were relieved only by the slow process of individual emancipation and the cultural efflorescence of the Renaissance, which laid the intellectual basis for the great age of geographical and scientific discovery exemplified in the new technologies of gunpowder, mining, printing, and navigation. Not before the triumph of princely government, in fact, did political allegiance, economic interests, and spiritual authority again become centred in a viable unit of organization, the absolutist nation-state.

**The city and the nation-state.** The virtue of absolutism in the early modern period lay in its ability to utilize the new technologies. Through the centralization of power, economy, and belief it brought order and progress to Europe and provided a framework in which individual energies could once more be channeled to a common end. While the nation stripped the cities of their remaining pretensions to political and economic independence (symbolized in their walls and tariff barriers), it created larger systems of interdependence in which territorial division of labour could operate. Though new mercantilist policies built up national wealth, they did not necessarily foster the growth of cities. All too often the wealth of nations was dissipated in war. Much of the income produced in town and country went to bolster the monarch's power and advertise his fame; the splendour of court life and the baroque glory of palaces and churches were paid for by merchant enterprise and the toil of peasants and craftsmen. Only in colonial areas, notably the Americas, did the age of expansion see the planting of many new cities, and it is significant that the capitals and ports of the colonizing nations experienced their most rapid growth during these years. Under absolutist regimes, a few large political and commercial centres grew at the expense of smaller outlying communities and the rural hinterlands.

Technology and the centralization of power

By the 18th century, the mercantile classes were increasingly disenchanted with monarchical rule. They resented their lack of political influence and assured prestige. They objected to outmoded regulations that hindered their efforts to link commercial operations with the systematic improvement of production. Eventually, they would unite with other dissident groups to curb the excesses of absolutism, erase the vestiges of feudalism, and secure a larger voice in the shaping of public policy. In northwestern Europe, where these liberal movements went furthest, the city populations and their bourgeois elites played a critical role out of all proportion to their numbers. Elsewhere, as in Germany, the bourgeois were more reconciled to existing regimes or, as in northern Italy, had assumed a passive if not wholly parasitical role.

With the exceptions of Great Britain and the Netherlands, however, the proportion of national populations resident in urban areas nowhere exceeded 10 percent. As late as 1800 only 3 percent of world population lived in towns of more than 5,000 inhabitants. No more than 45

cities had populations over 100,000, and of these fewer than half were situated in Europe. Asia had almost two-thirds of the world's large-city population, and Edo (Tokyo), with 1,400,000 residents, was the largest 18th-century city, larger than ancient Rome or medieval Byzantium at their peaks. Clearly, the mere presence of large cities or merchant elites anywhere in the world did not ensure the development of a dynamic social economy: the decisive factor was industrialism.

### INDUSTRIALIZATION AND THE MODERN WORLD

Before 1800, innovations in agricultural and manufacturing technique had permitted a singular concentration of productive activity close to the sources of mechanical power—water and coal. A corresponding movement of population was accelerated by the perfection of the steam engine and the superiority of the factory over preindustrial business organization. From the standpoint of economy, therefore, the localization of differentiated but functionally integrated work processes near sources of fuel was the mainspring of industrial urbanism. Under conditions of belt-and-pulley power transmission, urban concentration was a means of (1) minimizing the costs of overcoming frictions in transport and communications and (2) maximizing internal economies of scale and external economies of agglomeration. Although the intellectual and social prerequisites for industrialization were not uniquely present in any one nation, an unusual confluence of commercial, geographic, and technological factors in Britain led to far-reaching changes in such strategic activities as textiles, transport, and iron. Britain became "the workshop of the world" and London its "head office." Differentiation went so far that the cotton, woollen, and iron districts became more specialized and productive, each proceeding within its own cycle of technical and organizational change. By the mid-19th century, similar if less comprehensive industrial organization was evident in parts of France, the Low Countries, and the northeastern United States.

ndustry nd urban oncentra- ion

The concentration of the manufacturing labour force in "mill towns" and "coke towns" gradually undermined traditional social structures and relations. Age-old problems of public order, health, housing, utilities, education, and morals were aggravated by the influx of newcomers from the countryside. High rural birth rates combined with the industrialization of agriculture to release not only the country's foods and fibres but its children as well. Though the lowering of mortality in the 19th century was later offset by declines in fertility, the population of the more industrialized nations boomed into the 20th century, and the greater part of the increment migrated to the larger towns. The outcome was rural depopulation and the urbanization of society. Local institutions, often of medieval origin, were unable to cope with conditions that exaggerated poverty, disrupted family life, and complicated personal adjustment. Piecemeal reforms did little to improve the new milieu because, in the last analysis, the "city problem" arose not so much from the lack of public authority as from an unwillingness to pay the costs of social planning and improvement. Generations of urbanites experienced a continuing disorganization of their lives and work before the rising productivity of machines and increasing popular pressures on government could arrest the worst effects of this profound transformation. Slowly and painfully, the city's population adapted to its norms and enjoyed its satisfaction. New economic and cultural opportunities in the city evidently compensated for its congestion and strain.

ocial isruption nd reform

In the century after 1850 world population doubled and the proportion living in cities of more than 5,000 inhabitants rose from less than 7 percent to almost 30 percent. Between 1900 and 1950 the population living in large cities (100,000 plus) rose by 250 percent, the rate of increase in Asia being three times that of Europe and the United States. Nevertheless, the pattern of industrial urbanization (an overwhelmingly nonagricultural economy organized in a hierarchical system of different-sized cities ranging from one or more metropolitan centres at the top to a broad base of smaller-sized cities underneath) was still largely confined to the economically advanced areas: Europe, North America, Japan, and to a lesser extent Australasia. Meanwhile, industrial urbanism had entered on its metropolitan phase. The widespread use of cheap electric power, the advent of rapid transit and communications, new building materials, the automobile, and rising levels of per capita personal income had led to some relaxation of urban concentration. City dwellers began moving out from older downtown areas to suburbs and satellite communities where conditions were thought to be less wearing on nerves and bodies. Rising central area land values, traffic congestion, increased taxation, and festering slums reinforced the exodus. At the city's core the composition of the resident population came to include growing proportions of the aged, minority groups, and the very poor.

In the reshaping of the 20th-century city, advantages for residence and consumption probably played a more decisive role than advantages for production. Thus, while its advantages for manufacturers have diminished, the city remains the only feasible locus for the mass of specialized service activity that forms so large a part of the modern economy: the city offers maximum access to people. The spread of the city, however, has further weakened the vitality of local government: the difficulty of defining appropriate administrative boundaries has been added to the older problems of powers and finance. The task is to find viable forms of government for vast metropolitan districts, sometimes identified as conurbations, which sprawl across the countryside without unity or identity. (E.E.La./Ed.)

## Urban planning

Urban planning and redevelopment is aimed at fulfilling social and economic objectives that go beyond the physical form and arrangement of buildings, streets, parks, utilities, and other parts of the urban environment. Urban planning takes effect largely through the operations of government and requires the application of specialized techniques of survey, analysis, forecasting, and design. It may thus be described as a social movement, as a governmental function, or as a technical profession. Each aspect has its own concepts, history, and theories. Together they fuse into the effort of modern society to shape and improve the environment within which increasing proportions of humanity spend their lives: the city.

### THE DEVELOPMENT OF URBAN PLANNING

**Early history.** There are examples from the earliest times of efforts to plan city development. Evidence of planning appears repeatedly in the ruins of cities in China, India, Egypt, Asia Minor, the Mediterranean world, and South and Central America. There are many signs: orderly street systems that are rectangular and sometimes radial; divisions of a city into specialized functional quarters; development of commanding central sites for palaces, temples, and what would now be called civic buildings; and advanced systems of fortifications, water supply, and drainage. Most of the evidence is in smaller cities, built in comparatively short periods as colonies. Often the central cities of ancient states grew to substantial size before they achieved governments capable of imposing controls. In Rome, for example, the evidence points to no planning prior to late applications of remedial measures.

For several centuries during the Middle Ages, there was little building of cities in Europe. There is conflicting opinion on the quality of the towns that grew up as centres of church or feudal authority, of marketing or trade. They were generally irregular in layout, with low standards of sanitation. Initially, they were probably uncongested, providing ready access to the countryside and having house gardens and open spaces used for markets and fairs or grazing livestock. But as the urban population grew, the constriction caused by walls and fortifications led to overcrowding and to the building of houses wherever they could be fitted in. It was customary to allocate certain quarters of the cities to different nationalities, classes, or trades, as in cities of the Far East in the present day. As these groups expanded, congestion was intensified.

The physical form of medieval and Renaissance towns

and cities followed the pattern of the village, spreading along a street, a crossroad, in circular patterns or in irregular shapes—though rectangular patterns tended to characterize some of the newer towns. Most streets were little more than footpaths—more a medium for communication than for transportation—and even in major cities paving was not introduced until 1184 in Paris, 1235 in Florence, and 1300 in Lübeck. As the population of the city grew, walls were often expanded, but few cities at the time exceeded a mile in length. Sometimes sites were changed, as in Lübeck, and many new cities emerged with increasing population—frequently about one day's walk apart. Towns ranged in population from several hundred to perhaps 40,000 (London in the 14th century). Paris and Venice were exceptions, reaching 100,000.

Housing varied from elaborate merchant houses to crude huts and stone enclosures. Dwellings were usually two to three stories high, aligned in rows, and often with rear gardens or inner courts formed by solid blocks. Windows were small apertures with shutters, at first, and later covered with oiled cloth, paper, and glass. Heating improved from the open hearth to the fireplace and chimney. Rooms varied from the single room for the poor to differentiated rooms for specialized use by the wealthy. Space generally was at a premium. Privacy was rare and sanitation primitive.

Urban design in the Renaissance

During the Renaissance, however, there were conscious attempts to plan features, such as logistically practical circulation patterns and encircling fortifications, which forced overbuilding as population grew. As late as the 1860s, the radial boulevards in Paris had military as well as aesthetic purposes. The grand plan, however, probably had as its prime objective the glorification of a ruler or a state. From the 16th to the end of the 18th century, many small cities and parts of large cities were laid out and built with monumental splendour. The result may have pleased and inspired the citizens, but it rarely contributed to the health or comfort of their homes or to the efficiency of manufacturing, distribution, or marketing.

The planning concepts of the European Renaissance were transplanted to the New World. In particular, Pierre l'Enfant's plan for Washington, D.C. (1791), illustrated the strength and weakness of these concepts; it was a plan ably designed to achieve monumentality and grandeur in the siting of public buildings but was in no way concerned with the efficiency of residential, commercial, or industrial development. More prophetic of the layout of U.S. cities was the rigid, gridiron plan of Philadelphia, designed by William Penn (1682), with a layout of streets and lots (plots) adaptable to rapid changes in land use but wasteful of land and inefficient for traffic. The gridiron plan travelled westward with the pioneers, since it was the simplest method of dividing surveyed territory. Its special advantage was that a new city could be planned in the eastern offices of land companies and lots sold without buyer or seller ever seeing the site.

The New England town also influenced later settlement patterns in the United States. The central commons, initially a cattle pasture, provided a focus of community life and a site for meetinghouse, tavern, smithy, and shops. It became the central square in county seats from the Alleghenies to the Pacific and remained the focus of urban activity. Also from the New England town came the tradition of the freestanding, single-family house. Set well back from the street and shaded by trees, it had an ornamental front yard and a working backyard and became the norm of American residential development. This was in contrast to the European town house, with its party wall and tiny fenced backyard.

**19th century.** In both Europe and the United States, the surge of industry during the 19th century was accompanied by rapid population growth, unfettered individual enterprise, great speculative profits, and remarkable lapses of community responsibility. During this era, sprawling, giant metropolitan cities developed, offering wealth and adventure, variety and change. Their slums, congestion, disorder, and ugliness, however, provoked a reaction in which housing reform was the first demand. Industrial slums in European and American cities were unbelievably congested, overbuilt, unsanitary, and unpleasant. The early regulatory laws set standards that improved upon the slums of the time but seemed a century later to be impossibly low. Progress was very slow, for the rent-paying ability of slum dwellers did not make it profitable to invest in better housing for them. Housing improvement as an objective, however, recurred continually. Early significant improvements in public health resulted from engineering improvements in water supply and sewerage, which were essential to the later growth of urban populations.

Industrial slums

Toward the end of the 19th century, another effort to improve the urban environment emerged from the recognition of the need for recreation. Parks were developed to provide visual relief and places for healthful play or relaxation. Later, playgrounds were carved out in congested areas, and facilities for games and sports were established not only for children but also for adults, whose workdays gradually shortened.

Concern for the appearance of the city had long been manifest in Europe, in the imperial tradition of court and palace and in the central plazas and great buildings of church and state. In Paris, Georges-Eugène, Baron Haussman, became the greatest of the planners on a grand scale, advocating straight arterial boulevards, advantageous vistas, and a symmetry of squares and radiating roads. The resurgence of this European tradition had a counterpart in the "city beautiful" movement in the United States following Chicago's World Columbian Exposition of 1893. This movement expressed itself widely in civic centres and boulevards, contrasting with and in protest against the surrounding disorder and ugliness.

**20th century.** Early in the 20th century, during the sprawling growth of industrial cities, factories invaded residential areas, tenements crowded in among small houses, and skyscrapers overshadowed other buildings. To preserve property values and achieve economy and efficiency in the structure and arrangement of the city, the need was felt to sort out incompatible activities, to set some limits upon height and density, and to protect established areas from despoilment. Zoning (see below) was the result.

Consequences of developments in transportation

As transportation evolved from foot and horse to street railway, underground railway or subway, elevated railroad, and automobile, the new vehicles made possible tremendous urban territorial expansion. Workers were able to live far from their jobs, and complex systems of communications developed. The new vehicles also rapidly congested the streets in the older parts of cities. By threatening strangulation, they dramatized the need to establish orderly circulation systems of new kinds.

Metropolitan growth so intensified these and other difficulties that the people living in cities—who for the first time outnumbered the rural population in many countries—began to demand an attack upon all of these problems. In response, city planning by midcentury aimed not at any single problem but at the improvement of all aspects of the urban physical environment through unified planning of the whole metropolitan area. This introduced issues of national planning and in many countries brought city planning into the field of planning the nation's economic and social resources as a whole.

**Goals of modern urban planning.** The ultimate goals had always been social, even during the period when city plans themselves related only to physical change. They had been and continued to be deeply involved with intermediate economic objectives. The expression of the goals was, of course, coloured by the culture of the society seeking them. Of increasing weight was the goal of equality of opportunity and the redress of the grievances of disadvantaged minorities. Within this value system the physically oriented urban planning of the first half of the 20th century had evolved a set of environmental objectives that continued to be valid: (1) the orderly arrangement of parts of the city—residential, business, industrial—so that each part could perform its functions with minimum cost and conflict; (2) an efficient system of circulation within the city and to the outside world, using to the maximum advantage all modes of transportation; (3) the development of each part of the city to optimum standards, in terms of lot size, sunlight, and green space in residential areas,

Six objectives of urban planning

and parking and building spacing in business areas; (4) the provision of safe, sanitary, and comfortable housing in a variety of dwelling types to meet the needs of all families; (5) the provision of recreation, schools, and other community services of adequate size, location, and quality; (6) the provision of adequate and economical water supply, sewerage, utilities, and public services.

Even these superficially clear objectives, however, were not fully operational. They involve such terms as "adequate" and "high standard," which are relative rather than absolute, and change with new insights from experience or research (medical, psychological, social) and with new technological achievements. Inherent in the concept of city planning was the recognition that an ideal is not a fixed objective but will itself change, that the ideal city can be striven toward but never achieved. This turned the focus of planning away from the "master plan" and toward a stress upon the process and the directions of change.

## PLANNING AND GOVERNMENT

As a normal and identifiable function of government, city planning for the physical environment has been recognized in Europe and the United States since the early years of the 20th century. The year 1909 was a milestone. It saw the passage of Britain's first town planning act and, in the United States, the first national conference on city planning, the publication of Daniel Burnham's plan for Chicago, and the appointment of Chicago's Plan Commission (the first official planning agency in the U.S. was in Hartford, Connecticut, in 1907). Germany, Sweden, and other European countries also developed planning administration and law at this time.

Varying approaches to the administration of planning

City planning as a government function involves the coordination of all governmental activities that bear upon community growth and change, especially those that influence private development, so that they all work toward comprehensive objectives. The place of the city-planning function in the structure of urban government has developed in different ways in different countries. On the continent of Europe, where municipal administration was strongly centralized, city planning became the sphere of an executive department with substantial authority. In Great Britain, the local planning authority was a local legislative body (the county or county borough in England and Wales, the county or burgh in Scotland), advised by a planning committee of local councillors and with a planning department to act in an executive and advisory capacity. In the United States, with its tradition of tripartite government, it was recognized that decisions of importance to community development were made both by the executive branch (mayor) and the legislative (council). Rather than impinge on the authority of either, planning was allotted to a separate commission, advisory to both, with no authority beyond the right to be consulted before any action affecting the plan was taken.

**Zoning and subdivision controls.** Zoning, the regulation of the use of land and buildings, the density of population, and the height, bulk, and spacing of structures, was the principal means of putting into effect a comprehensive scheme for land use. It is generally dated from the adoption of New York City's first comprehensive ordinance in 1916. Though zoning was used in Great Britain and other European countries, it was developed furthest in the United States. The first ordinances were simple regulations, intended to protect existing property values and preserve light and air. As planning itself broadened its objectives and evolved its techniques during the 1930s, zoning developed into a more precise and sensitive tool.

Parallel to the evolution of zoning in the United States was the development of subdivision controls—subjecting the initial laying out of vacant land to public regulation. It was realized, after bitter experience with suburban land speculations in the 1920s, that the interest of the owner and developer of raw land is sometimes temporary and purely financial, while the urban community must live with the results for generations afterward. Subdivision regulations in many United States cities specified that new streets conform to the overall city plan and that new lots be properly laid out for building sites. Some required the developer to give the land needed for streets, playgrounds, and school sites and to pay all or most of the cost of development of these facilities.

Zoning and subdivision control offered adequate controls over the growth of new parts of cities, where they were used by enlightened legislative bodies. It was realized, however, that they were insufficient to correct past mistakes and especially to bring about the rebuilding of the obsolete parts of cities.

Experimental new towns of the early 20th century

**Large-scale planning.** All over the Western world, in the first half of the 20th century, new towns were built, constituting a very small part of the total or urban growth but serving as experiments and as examples of what could be done. This was largely the product of England's garden-city movement, which proposed preplanned new cities, on land held by the community and limited to 30,000 population, complete with business services and employment centres and surrounded by permanent greenbelts of rural land. The initial experimental cities were undertaken in England by private initiative, motivated by a spirit of reform; Letchworth was started in the early 1900s and Welwyn Garden City in the 1920s.

The concept had substantial influence in the United States. Kingsport, Tennessee, was a new city built by industrial interests. Some of the design ideas were used in suburban real-estate developments, outstanding being that of Radburn, New Jersey, which pioneered the super-block scheme as the "town for the motor age." U.S. examples, however, omitted the community-ownership feature, and almost all omitted employment centres, balanced income groups, and effective greenbelts. The federal government undertook a few large-scale housing developments for immigrant industrial workers during World Wars I and II, as make-work projects during the depression of the 1930s, and as examples of sound urban design. During the 1960s a number of private-enterprise developments at the scale of new cities were undertaken, primarily integral to expanding metropolitan areas rather than as truly independent cities. Notable were Reston, Virginia, and Columbia, Maryland, both near Washington, D.C., and the Irvine Ranch area near Los Angeles.

Also during the 1930s a number of European countries, especially France, The Netherlands, Germany, and the Soviet Union, undertook the building of new towns as governmental enterprises. Most of them (except in the Soviet Union) were residential suburbs rather than complete urban units. During the period following World War II, many European countries made strides in the regulation of new growth and in planned rebuilding of bomb-torn city centres.

After World War II, Great Britain embarked on a bold program. It reorganized the planning districts of the country; established sweeping new powers over private land use, almost nationalizing the right to develop undeveloped land; and undertook to build new towns to receive population and industry from congested great cities, which were planned for building at lower densities. By 1960, 15 new towns were under way, but the national program had suffered reverses. At first, economic exigencies interfered with the relocation of industry suggested by long-range environmental planning, and some of the controls over private land development, which appeared to impede investment and construction, had to be relaxed. Most of the new towns, nevertheless, had become centres of rapid industrial and population expansion and constituted important new work in city plan effectuation.

Postwar urban redevelopment in the United States

By the early 1970s urban redevelopment and renewal in the United States had achieved some major successes in revitalizing the economy of central-city areas. Such programs were bitterly criticized, however, for displacing low-income families and for disregarding the network of social relationships that had meant more to these families than the squalor and danger of their physical shelter. Among the new programs evolved after the establishment of a new cabinet-level Department of Housing and Urban Development was the Model Cities program. This was an experiment, in several dozen U.S. cities, in attacking the problems of major blighted areas with massive federal financial aid. It included programs of physical improvement

coordinated with social and economic upgrading through job training, school improvement, encouragement of economic enterprise, and a complete panoply of self-help and outside-help measures aimed at reducing poverty and all of its adverse concomitants.

**Planning jurisdictions.** Where a single municipal government included all of an urban area, tools for physical planning and effectuation seemed, in the second half of the 20th century, to be approaching adequacy. This condition, however, was exceptional. In Europe and the Americas, the metropolitan area was the typical urban form, composed of many independent municipalities, with overlapping jurisdictions of counties, school districts, and special authorities. During this period, European countries were groping toward solutions of the metropolitan planning and development problem, with some progress in Great Britain, Scandinavia, West Germany, and The Netherlands. In the 1950s a limited metropolitan government was established for Toronto, Ontario, with planning as an integral function. As late as the early 1970s metropolitan planning efforts in the United States were still largely ineffective. Planning agencies had little voice in the decisions of not only the separate cities and suburbs but also larger public agencies, such as state highway departments, sewer and water supply authorities, and port and airport authorities. The U.S. planning movement had not yet evolved the governmental machinery for reconciling in a democratic way the conflicting interests of all of the constituents of a metropolitan area.

In Asia, the emerging industrial economies of the post-World War II period produced cities following many of the patterns of the West. These rapidly developing countries, however, are still preoccupied with political and economic problems and have made little progress in establishing an environmental planning function in city or metropolitan government effective enough to prevent the mistakes made earlier in Western cities. There are a few outstanding examples of planned new cities in such widely scattered places as India, Israel, and South America. There are also signs of increasing concern in Puerto Rico, India, Indonesia, and elsewhere for regional development programs. (Ed.)

## Modern city government

A city cannot operate without a government of some kind. There are, indeed, no known examples of a city without government, however far back one goes in history. In some European countries cities were for centuries virtually independent political entities in themselves. Although the city-states of ancient Greece are the most famous illustration of this phenomenon, local government in such countries as England, France, Italy, Spain, and Germany is much older than national government. In the modern world, however, cities are contained within the boundaries of national states, and city government forms part of a much larger and more complex constitutional regime.

City government as a reflection of national government

City government invariably reflects the general characteristics of this national regime. When political democracy exists at the national level, as in most of the Western world and Japan today, cities enjoy a substantial degree of local autonomy and have democratic systems of government. When the regime is authoritarian, central control is likely to diminish or extinguish local self-government and to suppress democratic forms. Similarly, a country that prefers to concentrate power and responsibility in a single individual—or, conversely, in a committee—is likely to display this preference in its cities no less than in its central government.

### BASIC CHARACTERISTICS OF CITY GOVERNMENT

**Jurisdictions of cities.** Hitherto, the jurisdiction of city governments has been limited to the built-up urban area although there are exceptions to this in Brazil, South Africa, and Yugoslavia. A clear distinction between the city and the surrounding countryside no longer exists. The built-up central area extends without any sharp dividing line to the suburbs, then to the farther fringe composed of housing estates and villages for commuters, interspersed with small produce farms, recreational areas, industrial estates, and so forth, all of which may form a single area of interrelated activities. An army of commuters daily invades the main city, and at the close of the day they retreat to their homes in the suburbs or beyond. They and their families use the city for such purposes as recreation, trade, shopping, professional services, and higher or technical education. The city depends for its economic health on their services and their purchasing power. But commuters also have to be provided with costly daytime services such as police and fire protection, water supply, sewage, public health, highways, and public transport, although those who live outside the city limits usually contribute little or nothing to the municipal revenue.

Urban technology and the patterns of behaviour of contemporary life have made it difficult or impossible for municipalities to cope with the mounting problems of the city region, and particularly those of the metropolitan areas, unless drastic changes of structure and scope are carried out.

**City services.** Certain functions must be performed in every city. Law and order must be maintained; there must be some regulation of building to ensure a minimum of safety and to ensure that houses or workshops are not constructed on public land or in improper places; there must be regular methods of preventing, controlling, and extinguishing fires; and there must be regulations and executive action to protect the health of the citizens. The services now provided by city governments are different in nature and wider in scope than in the past. Generalization is impossible, but the most widespread functions today are the environmental and personal health services, including clinics and hospitals; primary, secondary, and further education; water supply, sewage, refuse collection and disposal; construction, maintenance, and lighting of streets; public housing; welfare services for the old, destitute, physically and mentally handicapped, orphans and abandoned children, unemployed and disabled workers, and other categories needing help; cemeteries and crematoriums; markets and abattoirs. The traditional services have been transformed beyond recognition.

The primary functions of law, order, safety, and health

Many cities have had museums and art galleries for a century or more. Today such institutions are often part of extensive programs for recreation and culture sponsored by the municipality. Public parks and playgrounds are not a new feature of city life, but they too have become part of the comprehensive programs of outdoor recreation organized by the municipality.

A group of public-utility services comprising the supply of gas, electricity, water, and public transport are frequently provided by the city government itself, by a public corporation closely connected with it, or by a commercial company operating under a concession granted by the municipality. In some countries, municipal enterprise in the public utility field has been supplanted by larger regional or national schemes.

A city council inevitably takes an interest in the economic well-being of the city that it governs. Every city government wishes to assist industry and promote trade, but there are great differences in the role assigned to local authorities in this respect in different countries. In the Communist regimes of eastern Europe, nearly the whole of local trade and much local industry is directly or indirectly under the control of the city government. On the other hand, in the United States a city government can control local industry only by means of zoning regulations, restrictions imposed under public health legislation, and so forth. In any case, municipal governments can do much to assist industry and commerce by good planning and physical development, by providing for trade fairs and exhibition centres, and by designing and developing roads, public housing, schools, and other municipal services to meet the needs of employers and employees.

The attraction of tourists also has become an almost universal goal of every city that has the slightest pretension to be of interest to visitors, and here too the municipality can do much to attract tourists by providing not only publicity and information but also convenient and agreeable facilities.

TYPES OF CITY GOVERNMENT

There are today four principal types of municipal systems of government: (1) the decentralized system found in federal constitutions; (2) the decentralized system found in unitary constitutions; (3) the supervisory system found under the "Napoleonic," or French-type, administration; and (4) the integrated system found in eastern Europe.

**Decentralized city government in federal systems.** In federal constitutions, local government is usually a matter lying within the jurisdiction of the state or provincial government rather than that of the national government. This is the position in the United States, and it accounts for the great diversity of municipal organization existing in that country.

Mayor and council

*United States.* The mayor-and-council form is the traditional type of city government in the United States. It prevails in a majority of American cities whether large or small. The relations between the mayor and the council are by no means uniform, but in general the borough or city council, which was the dominant partner in colonial days, has lost power as the role of the mayor has expanded. Bicameral councils have disappeared.

In the weak-mayor form, the council retains a good deal of administrative power that it exercises through committees. The mayor has few administrative powers but possesses a number of legislative and judicial functions. Many of the municipal officers in such cities are directly elected. In the past the result was often a lack of organized leadership because power and responsibility were too widely diffused. The only person able to coordinate the fragmented authority of these several parts of the city government was the party boss. It has been truly remarked that the price paid for his services was high (in graft) even though his product was of low quality.

The strong-mayor system

It is against this background that the rise of the strong-mayor system is to be seen—a system that now exists in most of the larger American cities and many of the smaller ones. In this type, the mayor presides over the council and usually has the right to veto its ordinary legislative acts. The veto may be absolute, or it may be a suspensory veto that can be overcome if the measure in question is again passed by the council by a specified majority. The mayor usually prepares the budget for submission to the council; convenes special sessions of the council to consider particular questions; appoints and dismisses heads of departments and can give them instructions or directions; appoints the chairmen and members of boards or commissions; and decides or participates in the appointment of other city employees, such as policemen, firemen, and clerks, though patronage of this kind may be restricted by a civil service commission.

In the commission form the council is replaced by a small body of elected commissioners who decide general policies of municipal administration in addition to performing the usual functions of a council. Each of the commissioners also serves as head of one or more departments. The commission may also appoint various boards and committees to work with it in such spheres as health, libraries, and recreation. A member of the commission is chosen to be mayor either by the citizens or by his fellow commissioners. He is not the chief executive but only *primus inter pares* ("first among equals"). He seldom has a veto power and is distinguished from the other commissioners only on ceremonial occasions.

The commission pattern of city government has never made great inroads on the mayor-and-council form. It exists in fewer than 10 percent of American municipalities at the present time and is now on the decline. Its lack of success is due to the divided authority among the elected commissioners and its inability to concentrate responsibility in a single officer.

City manager

A more serious challenge to the traditional form of city government came from the city-manager system. The most recently developed and the most rapidly spreading, it is derived from the method of organizing business corporations that was in favour in the first quarter of the 20th century; a general manager was entrusted with operating activities by a board of directors to whom he was responsible. When the concept was applied to city government, there emerged a small council numbering from three to nine members, all elected at large. The council passes ordinances, adopts the budget, decides rates of taxation, and engages the manager. The mayor (if there is one) has a role that is chiefly ceremonial. The city manager is the real chief executive; his position is generally set out in the city charter, which states that the council shall not interfere with his administrative functions. He has a duty to provide the council with whatever information they need to determine matters of policy.

Independent boards and commissions

In addition to these principal forms of municipal government, there are in most American cities a considerable number of ad hoc boards and commissions possessing varying degrees of independence. The school board is invariably separate from the rest of the city administration, and there are usually many other independent boards. Some of them can appoint and dismiss their own staff. The mayor may have the power to appoint the chairmen and members, but thereafter his power to direct or influence may become slight or even negligible. A widespread characteristic of all forms of American city government is the fragmentation of authority caused by such devices.

State governments in the United States exercise many different kinds of supervision and control over the cities within their jurisdiction. The structure of local government is determined by state law, and every municipality owes its corporate status and machinery of self-government to the state.

*West Germany.* The primary unit of local government in the Federal Republic of Germany is the commune or municipality (*Gemeinde*). It may be either rural (*Landgemeinde*) or urban (*Stadt*). Above the municipality is the *Kreis,* which corresponds to a district or county containing villages, hamlets, and small towns; it is an upper tier of local self-government that exercises supervisory powers and also provides services. Some 90 towns, most of them with more than 50,000 population, however, are independent of this *Kreis* (*Kreisfreie Städte,* or Kreisfree towns). Above the *Kreis* is the province or administrative district (*Regierungsbezirk*) of the state (*Land*), though the *Regierungsbezirk* does not exist in the small states of the Saarland and Schleswig-Holstein. The three great cities of Bremen, Hamburg, and West Berlin rank as states as well as municipalities within the federation of West Germany.

Several fundamental changes were introduced when the republic was constituted after World War II. As in the past, municipalities are the creatures of their respective states (*Länder*); but the Bonn constitution requires that the representative organs of the municipalities shall be elected by universal, direct, free, equal, and secret ballot. Furthermore, the elected council is declared to be the supreme organ in the municipality, and chief executives must be chosen by either the citizens or the council.

The structure of local government varies among the different states. In Bayern and Baden-Württemberg, the burgomaster (*Bürgermeister*) and deputy burgomasters, who form the executive organ, are directly elected by the citizens. In Nordrhein-Westfalen and Niedersachsen, the burgomaster retains only political functions, his former administrative functions having been transferred to a city director (*Stadtdirektor*) or chief executive officer. The *Bürgermeisterverfassung* prevails in Rheinland-Pfalz and the Saarland and entails a functional distinction between the respective powers of the council and the mayor. The *Magistratsverfassung* obtains in Hesse, Schleswig-Holstein, and Bremen. The council elects its own chairman and also appoints a board consisting of six to 12 members to act as the executive organ of the municipality. The council designates one of the members of this board to serve as its chairman and as burgomaster. Whatever the form of the executive organ, it must prepare and carry out the resolutions, policies, and administrative decisions of the council. Only in urgent matters or emergencies can the executive decide in place of the council.

**Decentralized city government in unitary systems.** *Great Britain.* The Local Government Act of 1972 created a two-tier system of counties and districts. Both counties and districts have independent, locally elected councils that perform separate functions: county authorities are

generally responsible for large-scale services, while district authorities are generally responsible for more local ones. In certain heavily populated areas, counties and districts are designated metropolitan counties and districts, and provision of certain large-scale services falls upon these more powerful metropolitan district authorities.

In Britain, as elsewhere, central control has grown during the 20th century. Each department of the national government uses its own methods, but typical among them are the approval (or rejection) of schemes submitted by local authorities for town planning, education, and highways; approval of the appointment or qualifications of chief officers; and approval of slum clearance schemes and purchases of land. They also hear appeals by citizens dissatisfied with certain decisions of local authorities, and they inspect the police forces and the schools administered by local authorities. The overriding feature of central–local relations is the fact that local authorities have become increasingly dependent on central grants for their revenue. Their only source of tax revenue consists of rates levied on real property (leaving aside rents and fees for services). The result is that central grants are now larger than income from rates. Unless local authorities are given additional sources of fiscal revenue, they are unlikely to achieve a position of greater independence—despite strong movements for the reform of local controls and for a reduction in central control.

*Japan.* After World War II, a radical reform of local government took place directed toward the creation of political democracy in Japan. The new constitution provided that local authorities should be organized and operated in accordance with the principle of local autonomy, that both the chief executives and the local assemblies should be directly elected by popular vote, and that local authorities should have the right to manage their property and affairs and to make their own regulations within the law. A Local Autonomy Law passed in 1947 prescribed in detail the organization and functions of local government. Legislation followed on finance and the public service.

Local government structure

The municipalities consist of cities (*shi*), towns (*machi*), and villages (*mura*). All have the same structure and legal status but differ in powers. A city must have a population of not less than 50,000 (formerly 30,000), of which at least 60 percent must engage in commerce and industry; and it must possess civic halls, a sewage system, libraries, and other public amenities. In 1953 a compulsory amalgamation of local government units reduced the number of towns and villages from 9,610 to 2,915, while the number of cities was increased to 556. These figures have changed by only small percentages since.

The separation of powers is applied in municipalities of all types. The mayor is directly elected by the voters and so too is the assembly, of which he is not a member; the only exceptions are the special wards in which the mayors are appointed by the ward council with the concurrence of the metropolitan governor. The assembly passes the budget, enacts bylaws, approves the accounts, decides the local taxes, disposes of property, and can demand reports or carry out investigations. The mayor controls the entire administration, except certain functions that have been entrusted to separate administrative boards. These include education boards, election administrative commissions to manage both national and local elections, audit and inspection commissions, civil service commissions, and police boards.

The mayor has the right to convene the council and to place bylaws before it. He is the ceremonial head of the city as well as its chief executive. He is also entrusted with certain duties on behalf of the central government, such as the maintenance of national roads and the census.

The control over the cities exercised by the central government is considerable. The Ministry of Autonomy is the department mainly concerned with local government, but the ministries of Finance, Education, Construction, and Transport exercise varying degrees of control over local authorities. One of the unresolved problems in Japan is the reluctance of central departments to coordinate their activities, and this often leads to serious inconsistencies and conflicts of policy.

**The Napoleonic supervisory system.** *France.* Until the administrative reforms of 1982, the system of local government in France was derived mainly from the French Revolution and the Napoleonic era, when the basic organization was a highly centralized administrative state in which the communes were local units of the central government. Historically this French pattern has had far-reaching influence in Europe, Africa, and Asia.

The mayor and the prefect

The French system had two tiers of authorities. In each of the communes (which comprise municipal units of all sizes, from tiny villages to large cities), there was an elected council and a mayor (*maire*), appointed by the councillors from among their own members. In each of the regional *départements* into which France is divided, there was a prefect (*préfet*) who was appointed by and responsible to the central government and who was assisted by an elected departmental council. The separation of powers was applied at both levels, and the mayor and the prefect were executive officers, whereas the councils were deliberative and legislative bodies.

The mayor was the responsible head of the municipality. He represented the local community. He presided at council meetings and carried out its decisions. He prepared the budget and submitted it to the council for its consideration. He initiated proposals for new measures or policies and approved all expenditures. In the large towns, the mayor had the help of several assistant mayors; he could delegate particular tasks to one or more of them, and he also could delegate functions to a councillor. In the larger municipalities, the mayor appointed a secretary general who was answerable to him for the day-to-day administration, thus leaving the mayor free to devote himself to political leadership and policy making. Mayors were not infrequently among the leading national figures in French politics.

A city government was legally authorized to deal with all matters of local interest, but the extent to which it did so depended on the attitude of the council and its mayor, the resources at its disposal, and the degree of control exercised by higher authority. The prefect was the supervisory agent who possessed the authority of all the national ministries, including even those that had their own specialists working in the field. When the mayor was carrying out his duties as an agent of the state, the prefect was his hierarchical superior and could amend or set aside his decisions.

Under a law passed on March 2, 1982, a process of decentralization was begun in which administrative and financial powers are transferred from the appointed departmental prefects to the presidents of the locally elected departmental councils. The departmental councils are still deliberative and legislative bodies. The presidents of the councils, however, are now the chief executives of the *départements,* responsible for, among other things, budgetary matters and provision of services. The prefects were renamed Commissioners of the Republic and now act as representatives of the central government, overseeing the local administrative bodies and maintaining public law and order. Mayors, who are representatives of the communes and also agents of the central government, and elected councils are still the governing bodies of the communes.

More than 80 percent of the communes have a population of less than 1,000. Because units of this size cannot hope to provide the services demanded by advanced nations today, they are permitted under French law to form joint associations for common tasks, or to merge and constitute a new municipality, or to establish a *syndicat,* or corporate body, with its own budget to carry out particular functions. Although these devices have been used in cities of medium size, they have not solved the problems of the metropolitan areas. Under a law enacted in 1966, metropolitan authorities must be established in the great urban concentrations of Lyon, Lille-Roubaix Tourcoing, Bordeaux, and Strasbourg; similar action is permitted in any other concentrated area of 50,000 or more inhabitants. Under this law, each constituent municipality continues to be responsible for administering and developing a wide range of services concerning its own area; but the metropolitan authority undertakes a number of functions

of interest to the metropolis as a whole. These relate to such matters as town planning and the public services involving roads, water, sanitation, housing, industrial estates, transport, secondary schools, and the like. Also included are public investments for assisting the arts, sports, hospitals, and other social and educational facilities. The governing body of the metropolitan authority is a council of between 50 and 90 members, who are representatives of the local authorities.

*Latin America.* The tendency in Latin-American countries is to adopt the basic principles of the supervisory system. This involves appointing central government officers of the prefectoral type who exercise control over local authorities.

Despite this general trend, there are large differences in the degree of local autonomy existing in the various countries of Latin America. In Argentina, for example, under a succession of military governments, the former elected city councils were supplanted by advisory committees appointed from above. Democratic rule was returned to the country in 1983, however, when general elections were held, and a civilian government assumed power.

In Ecuador the municipalities have elected councils and indirectly elected mayors who undertake local government functions. They are supervised by the centrally appointed "political chief" of the municipality (*jefe político*); his immediate superior is the provincial governor, who is appointed by the president of the republic. The *jefe político* supervises the city council and reports any illegal or irregular acts to the provincial governor; he also performs a number of state functions in the municipality. The provincial governor is charged with maintaining public order and upholding the law and constitution and is also responsible for developing the education, health, welfare, and cultural services and the public works projects.

Chile generally had elected municipal councils that appointed their own mayors, but in the large cities of Santiago, Valparaíso, Viña del Mar, and Concepción the mayors were appointed and removable by the president. At each level of administration—provincial, departmental, municipal, and district—there was a representative of the president assisted by a small elected council. This followed the essential principle of the French system. Since a military government assumed power in 1973, however, officials appointed by the president have been in charge of administration at the various levels.

Even more subordinate to the central government are the cities in countries where the mayor is appointed by either the president or the prefect on his behalf. This occurs in such countries as Bolivia and Colombia. The municipalities have elected councils, but their functions are mainly advisory. A high degree of local autonomy exists in Mexico, where both mayors and councils are elected.

The soviets under Communism

**The integrated system.** *Soviet Union.* The Soviet Union was from the beginning based on local "soviets" (elected councils) set up by the Communist Party in every province, district, city, town, or village throughout the land. These soviets were responsible not only for managing the affairs of their own areas but also for electing the soviet of the next higher level of government. Thus, the entire hierarchy of councils except those at the lowest level was based on indirect election. For this reason, the Soviet Union was originally described as a state of soviets. When direct election was introduced at all levels in 1936 this ceased to be a correct description, but the local soviets remained organs of local government. The regime has, however, retained a feature of great importance known as "democratic centralism." This means that each local authority is responsible to and must carry out the directions of the corresponding organ at the next higher level of government. Although some policies may originate locally, they have to be submitted to the next higher soviet for approval or, if of sufficient importance, to even higher soviets. This hierarchical and integrated structure explains why every soviet is considered a local organ of the state.

The executive committee and the presidium

A city soviet elects an executive committee (*ispolkom*), which in turn appoints a presidium, the principal executive organ. The presidium formulates operational and financial policies and directs the conduct of the city's services, whereas the full executive committee confines itself mainly to confirming the actions of the presidium. The executive committee is also responsible for preparing the agenda for the soviet, seeing that its decisions are carried out, ensuring that directions from above are obeyed, and giving advice or help to deputies and their committees. The president is the most important public figure in the municipality; he supervises personnel policy, convenes meetings of the executive committee, sees that complaints and petitions from citizens come to the executive committee, allocates functions to the administrative departments, and generally supervises the work of the city government. He is likely to be held personally responsible if things go wrong or if plans are not fulfilled.

The city soviet wields comparatively little power. Its short sessions, large membership, and crowded agenda preclude detailed discussion of problems. The deputies nevertheless do form a vital link between the citizens and the city government; and, moreover, the city soviets do have standing commissions composed of deputies and interested citizens called activists who are prepared to take an interest in a particular service such as housing or public health. In a large city, there may be as many as 15 of these commissions inspecting and reporting grievances or defects to the soviet, including alleged improper or criminal actions on the part of officials. About 2,500,000 activists are engaged in this work of the commissions. This attempt to achieve mass participation in the conduct of local government is a special feature of the soviet system. It exists side by side with a high degree of central control.

Functions of soviet city governments

A city government in the Soviet Union has in theory an almost unlimited jurisdiction. It administers the services for which municipal authorities are responsible in most developed countries, but in addition it is engaged in retail distribution, local industry of many kinds, public utilities, and many other kinds of municipal enterprises. It controls the entire construction industry and employs all of the architects. A city soviet will even control many aspects of industrial enterprises that are not directly subordinate to it. This extensive range of functions is liable to restriction, expansion, or direction by the soviet organs at the next higher level of government or by the Communist Party apparatus. With increasing knowledge, professional skill, and experience, however, the governments of the larger cities have been accorded a greater degree of discretion than formerly in dealing with municipal affairs.

Cities with a population exceeding 100,000 may have a lower tier of wards or *rayon* soviets. They are, in most respects, replicas of the main city government except that their executive committees do not appoint presidia; they also have fewer administrative departments. A *rayon* soviet shares many functions with the main city government. It may run the local shops, schools, clinics, and even hospitals. It has its own planning organ and its own budget, but its plans must be approved by the city government, and the expenditure and revenues of the *rayon* are embodied in the budget of the main city.

Metropolitan control

The problem of reorganizing local government in metropolitan areas to adjust it to changes in the distribution of population, the location of employment, and patterns of commuting has been solved more easily in the Soviet Union than in Western countries. One reason is that the large city government exercises a much greater degree of control over the surrounding area. Another advantage is the ease with which changes in the arrangement or boundaries of areas and authorities are carried out; such matters, under the direction of republican ministries, are effected with little difficulty or opposition.

*Poland.* Poland has adopted the soviet model with certain modifications. All levels of government are completely integrated, and the various people's councils are regarded as links in a unified system of state power. For the realization of its policies, the central government relies entirely on the people's councils in the provinces, districts, towns, and so on, and the principle of democratic centralism applies throughout.

In Warsaw and other large cities a two-tier structure has been established, consisting of a people's council for the entire city and a number of local councils for the sepa-

rate wards or boroughs. Poland has also developed small neighbourhood units in the form of committees that play a significant part in such matters as the development of housing, the care of gardens, and the improvement of amenities, and generally in securing widespread participation by the citizens in the city government. The soviet concept of mass participation in municipal affairs has been vigorously applied in Poland.

Each people's council elects from its members an executive board and also sets up numerous committees, which include nonelected persons, to carry out functions similar to those undertaken by the standing commissions in Soviet cities. They do not have the right to make decisions or to lay down rules or regulations, but they can supervise the activities of the city government and give advice or submit reports to the people's council.

The city government has extensive functions: it operates the shops, schools, hospitals, transport, gas and electricity supply, local industries, theatres, museums, and other forms of indoor and outdoor recreation. It may undertake, alone or jointly with neighbouring local authorities, the building of roads, bridges, and other public works. Although the main industries are administered by separate agencies that are not controlled by the city government, these agencies must keep the city informed of any policy decisions or activities that would affect the inhabitants of the city, and they must obtain the city's approval of such changes.

**Distribution of the various systems.** The four types of city government described in this article form the basis of nearly every system of local self-government in the civilized world. The British system was transplanted to the United States, Canada, Australia, New Zealand, Ireland, and South Africa, although in each of those countries significant divergences appeared. The most important was the emergence in the United States of a strong, directly elected mayor as the chief executive responsible for the administration of the city, subject to certain overriding controls by the city council. The British system also was established in the former British colonies in Africa and Asia under the tutelage of British-appointed officers resembling prefects. (In the largest Indian cities, such as Calcutta and Bombay, however, the state government appoints a municipal commissioner who controls the administration.) The system of directly elected mayor and council has been transplanted from the United States to parts of West Germany, Japan, the Philippines, and one or two Latin-American states. The French system was established in all of the former French colonies, but it has also been widely imitated in Latin America and elsewhere. The soviet system has been adopted by all of the Communist regimes of eastern Europe except that of Yugoslavia.

The principal problems confronting city governments are broadly similar irrespective of the constitutional type. They concern the planning and development of large cities, particularly those classed as metropolitan areas; the continual erosion of local autonomy by the increase of central governmental control; the municipal dependence on grants and subsidies from the central government; and the immense difficulty of providing adequate traffic and transport facilities, housing, education, and welfare services at an acceptable standard. Finally, the deterioration of the environment has become a matter of serious concern that is likely to persist for many years. (W.A.R./Ed.)

**BIBLIOGRAPHY.** LEWIS MUMFORD, *The City in History* (1961), a historical review of the city with special attention to physical features and with philosophical observations and excellent illustrations; R.E. PARK, *Human Communities* (1952), a pioneer sociological analysis of urbanization; HENRI PIRENNE, *Medieval Cities* (1925), a historical treatment of the emergence of cities from the collapse of the Roman Empire through the Middle Ages; GIDEON SJOBERG, *The Pre-Industrial City* (1960), a comparative study of preindustrial cities with special focus on common characteristics; R.E. TURNER, *The Great Cultural Traditions,* vol. 1, *The Ancient Cities* (1941), a historical synthesis of the origin of cities from Neolithic times to Greek civilization; UNITED NATIONS, *Urbanization: Development Policies and Planning* (1968), a review of world urbanization trends, and *Growth of the World's Urban and Rural Population, 1920–2000* (1969), a statistical analysis; RAYMOND VERNON, *Metropolis 1985* (1960), an interpretation of the results of the New York Metropolitan Region study; A.F. WEBER, *The Growth of Cities in the Nineteenth Century* (1899, reprinted 1963), history and statistics of urban growth in the 19th century in the U.S., Europe, and selected Asian, African, and Latin-American countries. CARL ABBOTT, *The New Urban America: Growth and Politics in Sunbelt Cities* (1981), describes the post-World War II growth of cities in the southern United States.

For historical background on urban planning, see ERWIN A. GUTKIND, *International History of City Development,* 5 vol. (1964–70). For modern practice, see W.I. GOODMAN and ERIC C. FREUND (eds.), *Principles and Practice of Urban Planning,* 4th ed. (1968); MICHAEL P. BROOKS, *Social Planning and City Planning* (1970); KEVIN LYNCH, *Site Planning,* 2nd ed. (1971); FREDERICK H. BAIR, JR., *Planning Cities: Selected Writings on Principles and Practice* (1970); CHARLES ABRAMS, *Man's Struggle for Shelter in an Urbanizing World* (1964); and JOHN FRIEDMANN and WILLIAM ALONSO (eds.), *Regional Development and Planning* (1964).

The most useful general description of city government in many countries is S. HUMES and E.M. MARTIN, *The Structure of Local Government: A Comparative Survey of 81 Countries* (1969). A much more detailed study of a selected group of large cities may be found in W.A. ROBSON and D.E. REGAN (eds.), *Great Cities of the World: Their Government, Politics and Planning,* 3rd rev. ed. (1971). Local government in Europe is considered in W. HAUS and A. KREBSBACH, *Gemeindeordnungen in Europa* (1967), a reference volume in English, French, German, and Italian; INSTITUT EUROPÉEN D'ETUDES ET DE RELATIONS INTERCOMMUNALES, *Pouvoirs locaux en Europe* (1964); and W. ANDERSON (ed.), *Local Government in Europe* (1939). For the United States, see H.F. ALDERFER, *American Local Government and Administration* (1956); H. ZINK, *Government of Cities in the United States* (1939); and E.C. BANFIELD and J.Q. WILSON, *City Politics* (1963). The history of English city government may be studied in SIDNEY and BEATRICE WEBB, *The Manor and the Borough,* 2 vol. (1908, reprinted 1963); J. REDLICH and F.W. HIRST, *The History of Local Government in England* (1903, reprinted 1958); and J.H. THOMAS, *Town Government in the Sixteenth Century* (1933). Modern history is covered in H.J. LASKI, W.I. JENNINGS, and W.A. ROBSON (eds.), *A Century of Municipal Progress 1835–1935* (1935); J.H. WARREN, *The English Local Government System* (1968); G. RHODES, *The Government of London: The Struggle for Reform* (1970); and W.A. ROBSON, *Local Government in Crisis,* 2nd ed. (1968). The best historical work dealing with Germany is W.H. DAWSON, *Municipal Life and Government in Germany* (1914). R.H. WELLS, *German Cities* (1932), carries the story up to the collapse of democracy in face of the Nazi regime. For a recent statement, see W. BOCKELMANN, W. HAUS, and H. VON HERZFELD (eds.), *Kommunalwissenschaftliche Forschung* (1966). See also KEVIN LYNCH, *A Theory of Good City Form* (1981).

On French city government, see B. CHAPMAN, *Introduction to French Local Government* (1953), and *Prefects and Provincial France* (1955); and F. RIDLEY and J. BLONDEL, *Public Administration in France* (1964). For Italy the only relevant work is R.C. FRIED, *The Italian Prefects* (1963). On Latin America there is very little of value, but some information may be found in JACQUES LAMBERT, *Amérique Latine, structures sociales et institutions politiques* (1963; Eng. trans., *Latin America: Social Structure and Political Institutions,* 1967). Local government in the Soviet Union may be found in L.G. CHURCHWARD, *Contemporary Soviet Government* (1968); L. SCHAPIRO, *The Government and Politics of the Soviet Union,* 2nd ed. (1967); H. MCCLOSKY and J.E. TURNER, *The Soviet Dictatorship* (1960); and D.J.R. SCOTT, *Russian Political Institutions,* 4th ed. (1969). See also E. SIMON *et al., Moscow in the Making* (1937); and L.G. CHURCHWARD, "Soviet Local Government Today," *Soviet Studies,* 17:431–452 (1966). The special problems of metropolitan areas are analyzed in R.E. DICKINSON, *City Region and Regionalism* (1947); and in the reports of the WORLD CONFERENCE OF LOCAL GOVERNMENT, *Local Government Structure and Organization: Problems of Metropolitan Areas* (1962).

# Climate and Weather

From the ancient Greek origins of the word (*klíma,* "an inclination or slope"—*e.g.,* of the Sun's rays; a latitude zone of the Earth; a clime), and from its earliest usage in English, climate has been understood to mean the atmospheric conditions that prevail in a given region or zone. In the older form, "clime," it was sometimes taken to include all aspects of the environment, including the natural vegetation. The best modern definitions of climate regard it as constituting the total experience of weather and atmospheric behaviour over a number of years in a given region. Climate is not just the "average weather" (an obsolete, and always inadequate, definition). It should include not only the average values of the climatic elements, such as temperature, humidity, rainfall, and wind, that prevail at different times but also their extreme ranges, variability, and the frequency of various occurrences. Just as one year differs from another, decades and centuries are found to differ from one another by a smaller but sometimes significant amount: climate is, therefore, time-dependent, and climatic values or indices should not be quoted without specifying what years they refer to.

This article treats the factors that produce weather and climate and the complex processes that cause variations in both. Other major points of coverage include global climatic types and microclimates. The article also considers both the impact of climate on human life and the effects of human activities on the climate. Due attention is given to the scientific measurement of atmospheric conditions and weather forecasting. For details concerning the disciplines of meteorology and climatology, see EARTH SCIENCES, THE. See also the article ATMOSPHERE for further information about the properties and behaviour of the atmospheric system. Relevant data on the influence of the oceans and of atmospheric moisture on climate can be found in HYDROSPHERE, THE. (Ed.)

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 221, 222, 223, and 10/33, and the *Index.*

This article is divided into the following sections:


Solar radiation and temperature 437
  Solar radiation 437
    Distribution of radiant energy from the Sun
    Effects of the atmosphere
    Average radiation budgets
    Surface-energy budgets
  Temperature 439
    Global variation of mean temperature
    Diurnal, seasonal, and extreme temperatures
    Variation with height
    Circulation, currents, and ocean–atmosphere interaction
    Short-term temperature changes
Atmospheric humidity and precipitation 441
  Atmospheric humidity 441
    Humidity indices
    Relation between temperature and humidity
    Humidity and climate
  Precipitation 445
    Origin of precipitation in clouds
    Types of precipitation
    World distribution of precipitation
    Effects of precipitation
Atmospheric pressure and wind 452
  Atmospheric pressure 452
  Wind 453
    Relationship of wind to pressure and governing forces
    Cyclones and anticyclones
    Local winds
    Zonal surface winds
    Monsoons
    Upper-air waves
    Jet streams
    Stratospheric and mesospheric wind systems
Major forms of weather disturbances 468
  Thunderstorms 468
    Visual and aural phenomena
    Causes of thunderstorms
    Types of thunderstorms
    Energy of thunderstorms
    Physical characteristics
    Weather under thunderstorms
    Occurrence of thunderstorms
    Observations of thunderstorms
  Tornadoes, waterspouts, and whirlwinds 474
    General characteristics
    Occurrence and distribution
    Physical characteristics of the vortices
    Theories of vortex formation: the generation of tornadoes, waterspouts, and whirlwinds
  Hurricanes and typhoons 479
    Origins
    Physical characteristics
    Distribution
    Damage caused by tropical cyclones
    Warning and tracking of tropical cyclones
Climatic variations and change 483
  Seasonal cycle 483
  Climatic change 484
    Pre-Pleistocene climatic change
    Pleistocene climatic change
    Late Pleistocene and Holocene climatic change
    Causes of climatic variation
Climatic classification 497
  General considerations 497
  Approaches to climatic classification 498
    Genetic classifications
    Empiric classifications
  World distribution of major climatic types 499
    Type A climates
    Type B climates
    Type C and D climates
    Type E climates
    Highland climates
Climate and life 504
  Coevolution of climate and life 504
    The developing atmosphere
    Causes of climatic change
  Impact of climate on human life 505
    General observations
    Weather and technology
    World food production and climate
  Impact of human activities on climate 507
    Ozone depletion
    Acid rain
    Greenhouse effect induced by carbon dioxide and other trace gases
Meteorological measurement and weather forecasting 511
  General considerations 511
    Measurements and ideas as the basis for weather prediction
    Practical applications of weather forecasting
  History of weather forecasting 513
    Early measurements and ideas
    The emergence of synoptic forecasting methods
    Progress during the early 20th century
    Modern trends and developments
  Principles and methodology of weather forecasting 516
    Short-range forecasting
    Long-range forecasting
Scientific weather modification 518
  General considerations 518
  Methods of modifying atmospheric phenomena 519
    Cloud seeding
    Fog dissipation
    Precipitation modification
    Modification of other weather phenomena
    Changes in the radiation balance near the ground
Bibliography 521

# SOLAR RADIATION AND TEMPERATURE

Air temperatures have their origin in the absorption of radiant energy from the Sun. They are subject to many influences, including those of the atmosphere, ocean, and land, and are modified by them. As variation of solar radiation is the single most important factor affecting climate, it is considered here first.

## Solar radiation

### DISTRIBUTION OF RADIANT ENERGY FROM THE SUN

Nuclear fusion deep within the Sun releases a tremendous amount of energy that is slowly transferred to the solar surface, from which it is radiated into space. The planets intercept minute fractions of this energy, the amount depending on their size and distance from the Sun. A one-square-metre (11-square-foot) area perpendicular to the Sun at the top of the Earth's atmosphere, for example, receives about 1,365 watts of solar power. (This amount is comparable to the power consumption of a typical electric heater.) Because of the slight ellipticity of the Earth's orbit around the Sun, the amount of solar energy intercepted by the Earth steadily rises and falls by ±3.4 percent throughout the year, peaking on January 3 when the Earth is closest to the Sun. Although about 31 percent of this energy is not used as it is scattered back to space, the remaining amount is sufficient to power the movement of atmospheric winds and oceanic currents and to sustain nearly all biospheric activity.

Most surfaces are not perpendicular to the Sun, and the energy they receive depends on the solar elevation angle (90° for the overhead Sun). This angle changes systematically with latitude, the time of year, and the time of day. The noontime elevation angle reaches a maximum at all latitudes north of the Tropic of Cancer (23.5° N) around June 22 and a minimum around December 22 (Figure 1). South of the Tropic of Capricorn (23.5° S) the opposite holds true, and between the two tropics the maximum occurs twice a year. When the Sun has a lower elevation angle, the solar energy is less intense because it is spread out over a larger area (Figure 2). Variation of solar elevation is thus one of the main factors that accounts for the dependence of climatic regime on latitude. The other main factor is the length of day. For latitudes poleward of 66.5°, the length ranges from zero (winter solstice) to 24 hours (summer solstice), whereas the Equator has a constant 12-hour day throughout the year. The seasonal range of temperature consequently decreases from high latitudes to the tropics, where it becomes less than the diurnal range of temperature.

Variation of solar elevation and length of day

### EFFECTS OF THE ATMOSPHERE

Of the radiant energy reaching the top of the atmosphere, 46 percent is absorbed by the Earth's surface on average, but this value varies significantly from place to place, depending on cloudiness, surface type, and elevation. If there is persistent cloud cover, as exists in some equatorial regions, much of the incident solar radiation is scattered back to space and very little is absorbed by the Earth's surface. Water surfaces have low reflectivity (4–10 percent), except for low solar elevations, and are the most efficient absorbers. Snow surfaces, on the other hand, have high reflectivity (40–80 percent) and so are the poorest absorbers. High-altitude desert regions consistently absorb higher than average amounts of solar radiation because of the reduced effect of the atmosphere above them.

Absorption of solar radiation by the atmosphere

An additional 23 percent or so of the incident solar radiation is absorbed on average in the atmosphere, especially by water vapour and clouds at lower altitudes and by ozone ($O_3$) in the stratosphere. Absorption of solar radiation by ozone shields the terrestrial surface from harmful ultraviolet light and warms the stratosphere, producing maximum temperatures of −15° to 10° C (5° to 50° F) at an altitude of 50 kilometres (30 miles). Most atmospheric absorption takes place at ultraviolet and infrared wavelengths, so that more than 90 percent of the visible portion of the solar spectrum, with wavelengths between 0.4 and 0.7 micrometre (0.00002 to 0.00003 inch), reaches the surface on a cloud-free day. Visible light, however, is scattered in varying degrees by cloud droplets, air molecules, and dust particles. Blue skies and red sunsets are in effect attributable to the preferential scattering of short (blue) wavelengths by air molecules and small dust particles. Cloud droplets scatter visible wavelengths impartially (hence clouds usually appear white) but very efficiently, so that the reflectivity of clouds to solar radiation is typically about 50 percent and may be as high as 80 percent for thick clouds.

The constant gain of solar energy by the Earth is systematically returned to space in the form of thermally emitted radiation in the infrared portion of the spectrum. The emitted wavelengths are mainly between five and 100 micrometres, and they interact differently with the atmosphere compared to the shorter wavelengths of solar radiation. Very little of the radiation emitted by the Earth's

From G.T. Trewartha and L.H. Horn, *An Introduction to Climate* (1980), McGraw-Hill Book Company; reproduced with permission of the publisher

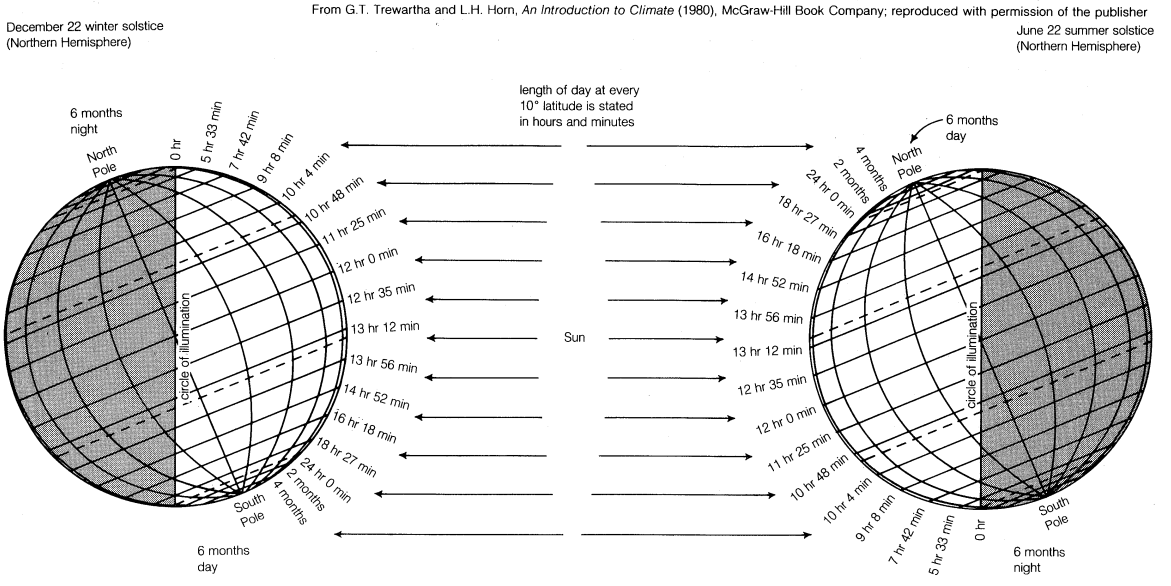



Figure 1: Vertical noontime solar rays reach 23.5° S and 23.5° N on December 22 and June 22, respectively. The circle of illumination then cuts all parallels, except the Equator, unequally, so that days and nights are unequal in length except at a latitude of 0°.

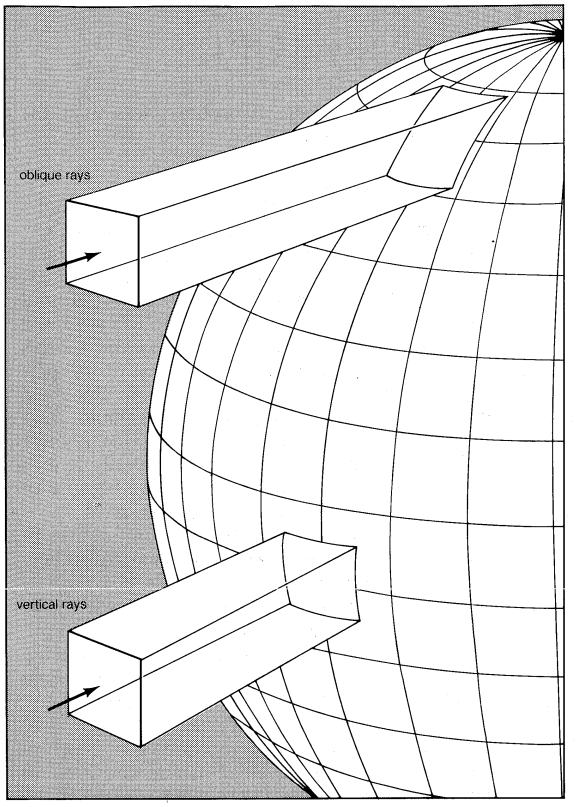



Figure 2: Less energy is received per unit area from oblique solar rays than from vertical ones.

From G.T. Trewartha and L.H. Horn, *An Introduction to Climate* (1980), McGraw-Hill Book Company; reproduced with permission of the publisher

surface passes directly through the atmosphere. Most of it is absorbed by clouds, carbon dioxide, and water vapour and is then reemitted in all directions. The atmosphere thus acts as a radiative blanket over the Earth's surface, hindering the loss of heat to space. The blanketing effect is greatest in the presence of low clouds and weakest for clear, cold skies that contain little water vapour. Without this effect, the mean surface temperature of 15° C would be some 30 degrees colder. Conversely, as atmospheric concentrations of carbon dioxide, methane, chlorofluorocarbons, and other absorbing gases continue to increase, in large part due to human activities, surface temperatures should rise because of the capacity of such gases to trap infrared radiation. The exact amount of this temperature increase, however, remains uncertain because of unpredictable changes in other atmospheric components, especially cloud cover. An extreme example of such an effect (commonly dubbed the greenhouse effect) is that produced by the dense atmosphere of the planet Venus, which results in surface temperatures of about 475° C. This condition exists in spite of the fact that the high reflectivity of the Venusian clouds causes the planet to absorb less solar radiation than the Earth.

The so-called greenhouse effect

## AVERAGE RADIATION BUDGETS

The difference between the solar radiation absorbed and the thermal radiation emitted to space determines the radiation budget of the Earth. Since there is no appreciable long-term trend in planetary temperature, it may be concluded that this budget is essentially zero on a global, long-term average. Latitudinally, it has been found that much more solar radiation is absorbed at low latitudes than at high latitudes. On the other hand, thermal emission does not show nearly as strong a dependence on latitude, so that the planetary radiation budget decreases systematically from the Equator to the poles. It changes from being positive to negative at a latitude of about 40° (Figure 3). The atmosphere and oceans, through their general circulation, act as vast heat engines, compensating for this imbalance by providing non-radiative mechanisms for the transfer of heat from the Equator to the poles.

The atmosphere and oceans as heat engines

While the Earth's surface absorbs a significant amount of thermal radiation due to the blanketing effect of the atmosphere, it loses even more through its own emission, and thus experiences a net loss of long-wave radiation. This loss is only about 14 percent of the amount emitted by the surface and is less than the average gain of absorbed solar energy. Consequently, the surface has on average a positive radiation budget.

By contrast, the atmosphere emits thermal radiation both to space and to the surface, yet it receives long-wave radiation back from only the latter. This net loss of thermal energy cannot be compensated for by the modest gain of absorbed solar energy within the atmosphere. The atmosphere thus has a negative radiation budget, equal in magnitude to the positive radiation budget of the surface but opposite in sign. Non-radiative heat transfer again compensates for the imbalance, this time largely by vertical atmospheric motions involving the evaporation and condensation of water.

## SURFACE-ENERGY BUDGETS

The rate of temperature change in any region is directly proportional to the region's energy budget and inversely proportional to its heat capacity. While the radiation budget may dominate the average energy budget of many surfaces, non-radiative energy transfer and storage also are generally important when local changes are considered.

Significance of non-radiative energy transfer

Foremost among the cooling effects is the energy required to evaporate surface moisture, which produces atmospheric water vapour. Most of the latent heat contained in water vapour is subsequently released to the atmosphere during the formation of precipitating clouds, although a minor amount may be returned directly to the surface during dew or frost deposition. Evaporation increases with rising surface temperature, decreasing relative humidity, and increasing surface wind speed. Transpiration by plants also increases evaporation rates, which explains why the temperature in an irrigated field is usually lower than that over a nearby dry road surface.

Another important non-radiative mechanism is the exchange of heat that occurs when the temperature of the air is different from that of the surface. Depending on whether the surface is warmer or cooler than the air next to it, heat is transferred to or from the atmosphere by turbulent air motion (more loosely, by convection). This effect also increases with increasing temperature difference and with increasing surface wind speed. Direct heat transfer may be an important cooling mechanism that limits the maximum temperature of hot, dry surfaces. Alternatively, it may be an important warming mechanism that limits the minimum temperature of cold surfaces. Such warming is sensitive to wind speed, so that calm conditions promote lower minimum temperatures.

In a similar category, whenever a temperature difference occurs between the surface and the medium beneath the

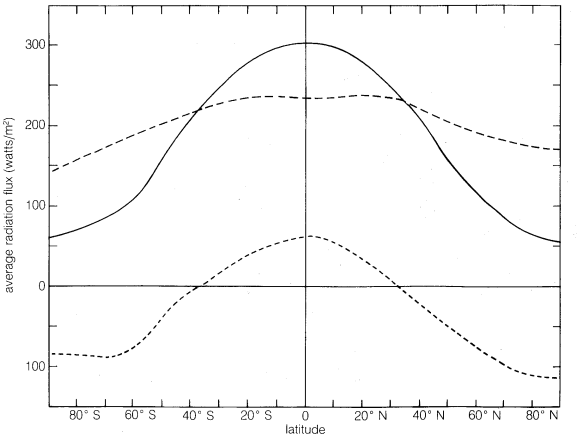



Figure 3: *Dependence of each component of the average planetary radiation budget on latitude.*
The solid curve represents absorbed solar radiation; the dashed curve, the net emission of long-wave radiation; and the dotted curve, the net radiation.

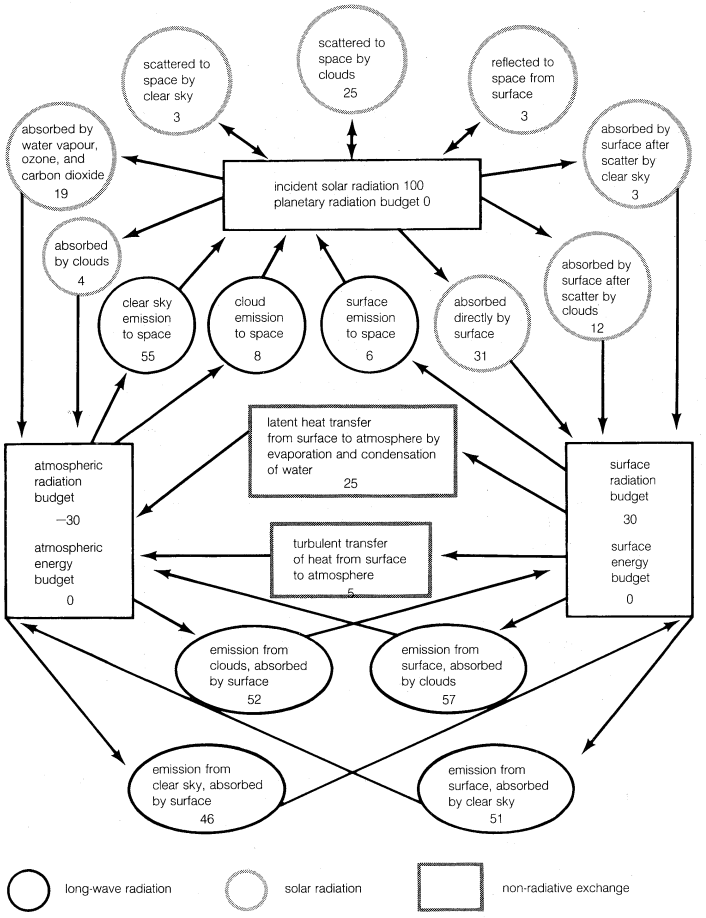



Figure 4: Average exchange of energy between the surface, atmosphere, and space, as percentages of incident solar radiation (1 unit = 3.4 watts per square metre).

Heat transfer by conduction and convection

surface, there is a transfer of heat to or from the medium. In the case of land surfaces, heat is transferred by conduction. In the case of water surfaces, the transfer is by convection and may consequently be affected by the horizontal transport of heat within large bodies of water.

Average values of the different terms in the energy budgets of the atmosphere and surface are given in Figure 4. The individual terms may be adjusted to suit local conditions and may be used as an aid to understanding the various temperature characteristics discussed in the next section. (Ro.D.)

## Temperature

### GLOBAL VARIATION OF MEAN TEMPERATURE

Figures 5 and 6 show the global variation of average surface air temperatures for January and July. These temperatures have been reduced to their corresponding sea-level values to remove the immediate effects of altitude. The values range from above 35° C in North Africa to below −50° C in Antarctica and show the major influences of latitude, continentality, ocean currents, and prevailing winds.

The effect of latitude is evident in the large north–south gradients in average temperature that occur at middle and high latitudes in each winter hemisphere. These gradients are due mainly to the rapid decrease of available solar radiation but also in part to the higher surface reflectivity at high latitudes associated with snow and ice and low solar elevations. A broad area of the tropical ocean, by contrast, shows little temperature variation.

Effect of continentality

Continentality is a measure of the difference between continental and marine climates and is mainly the result of the increased range of temperatures that occurs over land compared to water. This difference is a consequence of the much lower effective heat capacities of land surfaces as well as of their generally reduced evaporation rates. Heating or cooling of a land surface takes place in a thin layer, the depth of which is determined by the ability of the ground to conduct heat. The greatest temperature changes occur for dry, sandy soils because they are poor conductors with very small effective heat capacities and contain no moisture for evaporation. By far the greatest effective heat capacities are those of water surfaces due to the mixing of water near the surface and the penetration of solar radiation that distributes heating to depths of several metres. In addition, about 90 percent of the radiation budget of the ocean is used for evaporation. Ocean temperatures are thus slow to change.

The effect of continentality may be moderated by proximity to the ocean, depending on the direction and strength of the prevailing winds. Contrast with ocean temperatures at the edges of each continent may be further modified by the presence of a north- or south-flowing ocean current. For most latitudes, however, continentality explains much of the variation in average temperature at a fixed latitude as well as variations in the difference between January and July temperatures.

### DIURNAL, SEASONAL, AND EXTREME TEMPERATURES

The diurnal range of temperature generally increases with distance from the sea and toward those places where solar radiation is strongest, in dry tropical climates and on high mountain plateaus (owing to the reduced thickness of the atmosphere to be traversed by the Sun's rays). The average difference between the day's highest and lowest temperatures in January is 3° C (in July it is 5° C) in those parts of the British Isles nearest the Atlantic. It is 4.5° C in January and 6.5° C in July on the small island of Malta. At Tashkent, U.S.S.R., it is 9° C in January and 15.5° C in July, and at Khartoum, The Sudan, the corresponding figures are 17° C and 13.5° C. At Qandahār, Afg., which lies more than 1,000 metres above sea level, it is 14° C in January, 20° C in July, and exceeds 23° C in September and October when there is less cloudiness than in July. Near the ocean at Colombo, Sri Lanka, the figures are 8° C in January and 4.5° C in July.

The seasonal variation of temperature and the magnitudes of the differences between the same month in different years and different epochs generally increase toward high latitudes and with distance from the ocean. Extreme temperatures observed in different parts of the world are listed in Table 1.

### VARIATION WITH HEIGHT

There are two main levels where the atmosphere is heated—namely, at the Earth's surface and at the top of the ozone layer (about 50 kilometres up) in the stratosphere. Radiation balance shows a net gain at these levels in most cases. Prevailing temperatures tend to decrease with distance from these heating surfaces (apart from the ionosphere and outer atmospheric layers where other processes are at work). The world's average lapse rate of temperature (change with altitude) in the lower atmosphere is 0.6 to 0.7° C per 100 metres. Lower temperatures prevail with increasing height above sea level for two reasons: (1) because there is a less favourable radiation balance in the free air, and (2) because rising air—whether lifted by convection currents above a relatively warm surface or forced up over mountains—undergoes a reduction of temperature associated with its expansion as the pressure of the overlying atmosphere declines. This is the adiabatic lapse rate of temperature, which equals about 1° C per 100 metres for dry air and 0.5° C per 100 metres for saturated air in which condensation (with liberation of latent heat) is produced by adiabatic cooling. The difference between these rates of change of temperature (and therefore density) of rising air currents and the state of the surrounding air determines whether the upward currents are accelerated or retarded—*i.e.*, whether the air is unstable so that vertical convection with its characteristically attendant tall cumulus cloud and shower development is encouraged, or whether it is stable and convection is damped down.

Temperature lapse rate

For these reasons, the air temperatures observed on hills and mountains are generally lower than on low ground except in the case of extensive plateaus, which present a raised heating surface (and on still, sunny days, when even

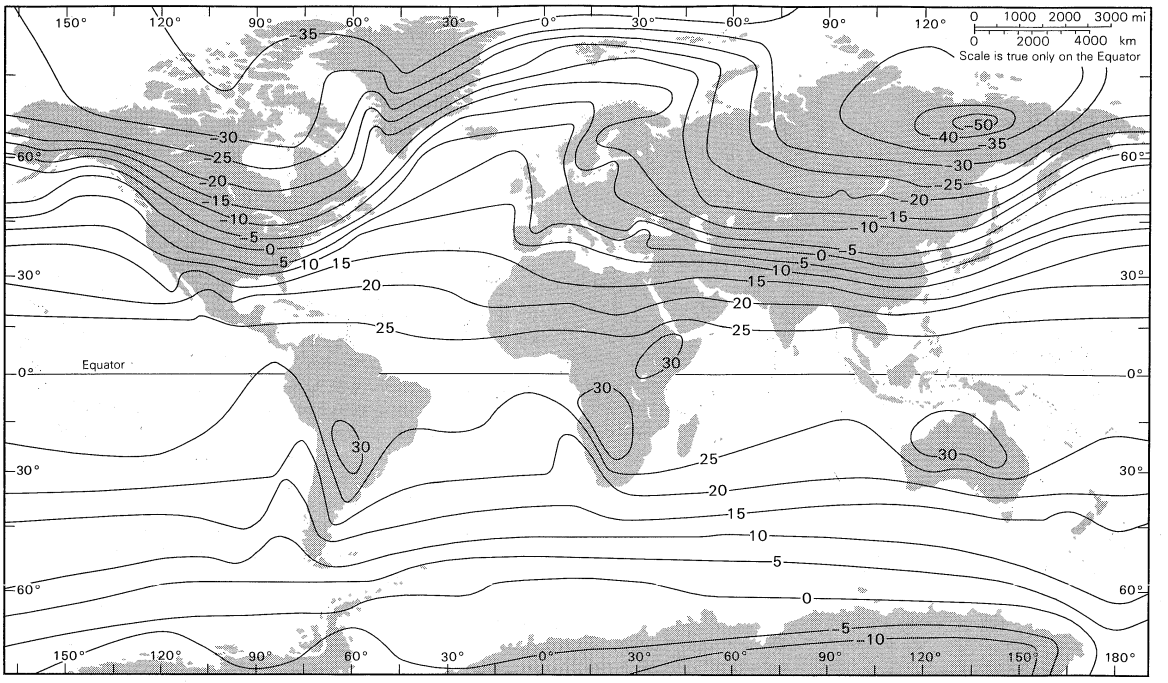



Figure 5: Average surface temperatures for January reduced to sea-level values (° C).
Adapted from P.W. Suckling and R.R. Doyon, *Studies in Weather and Climate* (1981); Contemporary Publishing Company of Raleigh, Inc.

a mountain peak is able to warm appreciably the air that remains in contact with it).

## CIRCULATION, CURRENTS, AND OCEAN–ATMOSPHERE INTERACTION

The circulation of the ocean is a key factor in air temperature distribution. Ocean currents that have a northward or southward component, such as the warm Gulf Stream in the North Atlantic or the cold Humboldt Current off South America, effectively exchange heat between low and high latitudes. In tropical latitudes the ocean accounts for a third or more of the poleward heat transport; at latitude 50° N the ocean's share is about one-seventh. In the particular sectors where the currents are located, their importance is of course much greater than these figures, which represent hemispheric averages.

Effects of currents on air temperature

A good example of the effect of a warm current is that of the Gulf Stream in January, which causes a strong east–west gradient in temperatures across the eastern edge of the North American continent (Figure 5). The relative warmth of the Gulf Stream affects air temperatures all the way across the Atlantic, and prevailing westerlies extend the warming effect deep into northern Europe. As a result, January temperatures of Tromsø, Nor. (69°40′ N), for example, average 24° C above the mean for that latitude. The Gulf Stream maintains a warming influence in July, but it is not as noticeable because of the effects of continentality.

The ocean, particularly in areas where the surface is warm, also supplies moisture to the atmosphere. This in turn contributes to the heat budget of those areas in which the water vapour is condensed into clouds, liberating latent

Adapted from P.W. Suckling and R.R. Doyon, *Studies in Weather and Climate* (1981); Contemporary Publishing Company of Raleigh, Inc.

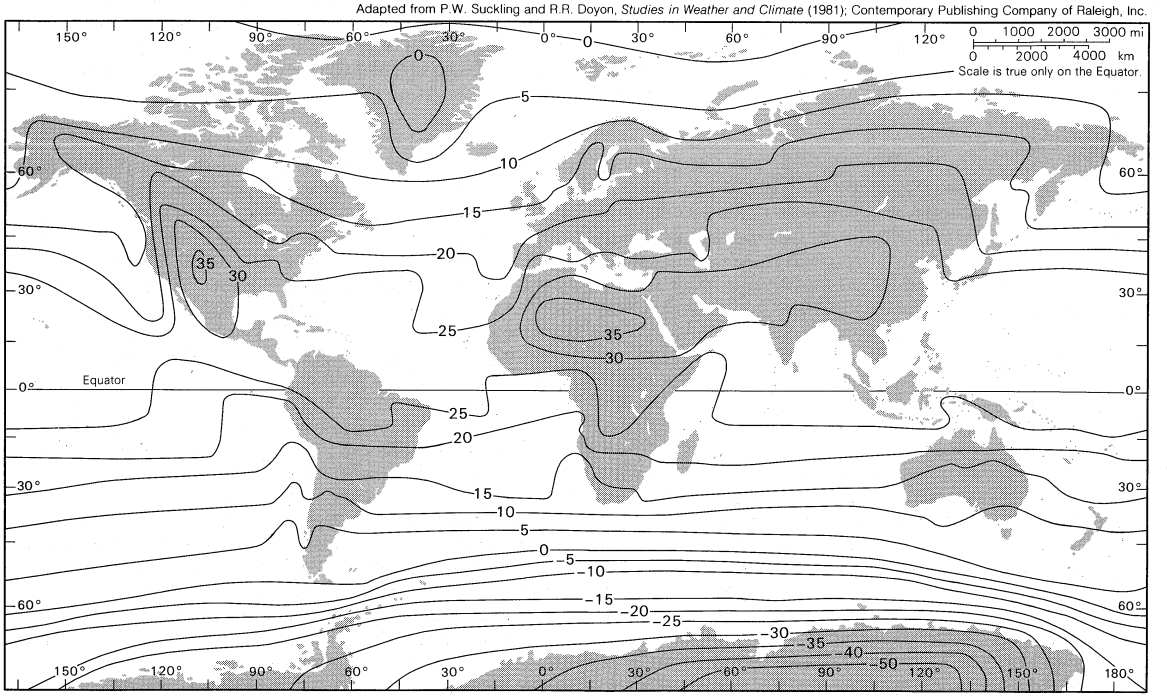



Figure 6: Average surface temperatures for July reduced to sea-level values (° C).

**Table 1: World Temperature Extremes**

| continent or region | highest recorded air temperature | | | lowest recorded air temperature | | |
|---|---|---|---|---|---|---|
| | place | temperature | | place | temperature | |
| | | (°C) | (°F) | | (°C) | (°F) |
| Africa | Al-'Aziziyah, Libya (112 m or 367 ft asl.*) | 57.7 | 136 | Ifrane, Morocco (1,635 m or 5,363 ft asl.) | −23.9 | −11 |
| Antarctica | Bahía Esperanza (Hope Bay) 63°23′ S, 57°00′ W (7m or 23 ft asl.) | 14.6 | 58 | Vostok II 78°27′ S, 106°52′ E (3,420 m or 11,218 ft asl.) | −88.3 | −127 |
| Asia | Jacobābād, Pakistan (57 m or 187 ft asl.*) | 52.8 | 127 | Verkhoyansk, NE Siberia, U.S.S.R. (137 m or 449 ft asl.*) | −69.8 | −94 |
| Australia | Cloncurry, Queensland (193 m or 633 ft asl.) | 53.1 | 127.5 | Charlotte Pass, New South Wales | −22.2 | −8 |
| Europe | Seville, Spain (39 m or 128 ft asl.*) | 50.0 | 122 | Ust-Shchugor, U.S.S.R. (85 m or 279 ft asl.) | −59.0 | −74 |
| N. America | Death Valley (Greenland Ranch), California (−54 m or −177 ft asl.) | 56.7 | 134.5 | Snag, Yukon (646 m or 2,119 ft asl.) | −62.8 | −81 |
| S. America | Rivadavia, Argentina (205 m or 672 ft asl.*) | 48.9 | 120 | Colonia, Sarmiento, Argentina (268 m or 879 ft asl.*) | −33.0 | −27 |
| Tropical Pacific islands | Echague, Luzon, Republic of the Philippines (78 m or 257 ft asl.) | 40.5 | 105 | Haleakala, Hawaii (2,972 m or 9,748 ft asl.) | −7.8 | 18 |

*asl.—above sea level.

heat in the process, often in high latitudes and in places remote from the ocean where the moisture was taken up.

The great ocean currents are themselves wind-driven—set in motion by the drag of the winds over vast areas of the sea surface, especially where waves increase the friction. At the limits of the warm currents, especially where they abut directly upon a cold current, as at the left flank of the Gulf Stream in the neighbourhood of the Grand Banks off Newfoundland and at the subtropical and Antarctic convergences in the oceans of the Southern Hemisphere, the strong thermal gradients in the sea surface result in marked differences in the heating of the atmosphere on either side of the boundary. These tend to position and guide the strongest flow of the jet stream (see below) in the atmosphere above and thereby influence the development and steering of weather systems.

Interactions between ocean and atmosphere proceed in both directions. They also operate at different rates. Some interesting lag effects, which are of value in long-range weather forecasting, arise through the much slower circulation of the ocean. Thus, enhanced strength of the easterly trade winds over low latitudes of the Atlantic, north and south of the Equator, impels more water toward the Caribbean and Gulf of Mexico, producing a stronger flow and greater warmth in the Gulf Stream some six months later. Anomalies in the position of the Gulf Stream–Labrador Current boundary, which produce a greater or lesser extent of warm water near the Grand Banks, so affect the energy supply to the atmosphere and the development and steering of weather systems from that region that they are associated with rather persistent anomalies of weather pattern over the British Isles and northern Europe. Anomalies in the equatorial Pacific and in the northern limit of the Kuroshio (also called the Japan Current) seem to have effects on a similar scale. Indeed, through their influence on the latitude of the jet stream and the wavelength (spacing of cold trough and warm ridge regions) in the upper westerlies, these ocean anomalies exercise an influence over the atmospheric circulation that spreads to all parts of the hemisphere.

Sea temperature anomalies

Sea-surface temperature anomalies that recur in the equatorial Pacific at variable intervals of two to seven years can sometimes produce major climatic perturbations. Such an anomaly is known as El Niño (Spanish for "the child"; it was so named by Peruvian fishermen who noticed its onset during the Christmas season). During an El Niño event, warm surface water flows eastward from the equatorial Pacific, in at least partial response to weakening of the equatorial easterly winds, and replaces the normally cold upwelling surface water off the coast of Peru and Ecuador that is associated with the northward propagation of the cold Peru (or Humboldt) Current. The change in sea-surface temperature transforms the coastal climate from arid to wet. The event also affects atmospheric circulation in both hemispheres and is associated with changes in precipitation in regions of North America, Africa, and the western Pacific.

El Niño event

### SHORT-TERM TEMPERATURE CHANGES

Many interesting short-term temperature fluctuations also occur, usually in connection with local weather disturbances. The rapid passage of a mid-latitude cold front, for example, can drop temperatures by 10° C in a few minutes and, if followed by a sustained movement of a cold air mass, by as much as 50° in 24 hours, with life-threatening implications for the unwary. Temperature increases of up to 40° in a few hours also are possible downwind of major mountain ranges when air that has been warmed by the release of latent heat on the windward side of a range is forced to descend rapidly on the other side (such a wind is variously called chinook, foehn, or Santa Ana). Changes of this kind, however, involve a wider range of meteorological processes than discussed in this section. For a more detailed treatment, see below *Atmospheric pressure and wind.* (H.H.L./Ro.D.)

# ATMOSPHERIC HUMIDITY AND PRECIPITATION

Atmospheric humidity, which is the amount of water vapour or moisture in the air, is another leading climatic element, as is precipitation. All forms of precipitation, including drizzle, rain, snow, ice crystals, and hail, are produced as a result of the condensation of atmospheric moisture to form clouds in which some of the particles, by growth and aggregation, attain sufficient size to fall from the clouds and reach the ground.

## Atmospheric humidity

At 30° C, 4 percent of the volume of the air may be occupied by water molecules, but where the air is colder than −40° C, less than one-fifth of 1 percent of the air molecules can be water. Although the water vapour content may vary from one air parcel to another, these limits can be set because vapour capacity is determined

by temperature. Temperature has profound effects upon some of the indices of humidity regardless of the presence or absence of vapour.

The connection between an effect of humidity and an index of humidity requires simultaneous introduction of effects and indices. Vapour in the air is a determinant of weather because it first absorbs the thermal radiation that leaves and cools the Earth and then emits thermal radiation that warms the planet. Calculation of absorption and emission requires an index of the mass of water in a volume of air. Vapour also affects the weather because, as indicated above, it condenses into clouds and falls as rain or other forms of precipitation. Tracing the moisture-bearing air masses requires a humidity index that changes only when water is removed or added.

## HUMIDITY INDICES

**Absolute humidity.** Absolute humidity is the vapour concentration or density in the air. If $m_v$ is the mass of vapour in a volume of air, then absolute humidity $d_v$ is simply $d_v = m_v / V$, in which $V$ is the volume and $d_v$ is expressed in grams per cubic metre (g/m³). This index indicates how much vapour a beam of radiation must pass through. The ultimate standard in humidity measurement is made by weighing the amount of water gained by an absorber when a known volume of air passes through it, and this measures absolute humidity, which may vary from 0 g/m³ in dry air to 30 g/m³ when the vapour is saturated at 30° C. The $d_v$ of a parcel of air changes, however, with temperature or pressure even though no water is added or removed because, as the gas equation states, the volume $V$ increases with the absolute, or Kelvin, temperature and decreases with the pressure.

**Specific humidity.** The meteorologist requires an index of humidity that does not change with pressure or temperature. A property of this sort will identify an air mass when it is cooled or when it rises to lower pressures aloft without losing or gaining water vapour. Because all the gases will expand equally, the ratios of the weight of water to the weight of dry air, or the dry air plus vapour, will be conserved during such changes and will continue identifying the air mass.

The mixing ratio $r$ is the dimensionless ratio $r = m_v/m_a$, where $m_a$ is the mass of dry air, and the specific humidity $q$ is another dimensionless ratio $q = m_v/(m_a + m_v)$. Because $m_v$ is less than 3 percent of $m_a$ at normal pressure and temperatures cooler than 30° C, $r$ and $q$ are practically equal. These indices are usually expressed in grams per kilogram (g/kg) because they are so small; the values range from 0 in dry air to 28 g/kg in saturated air at 30° C. Absolute and specific humidity indices have specialized uses, and so they are not familiar to most people.

**Relative humidity.** Relative humidity ($U$) is so commonly used that a statement of humidity, without a qualifying adjective, can be assumed to be relative humidity. $U$ can be defined, then, in terms of the mixing ratio $r$ that was introduced above. $U = 100r/r_w$, which is a dimensionless percentage. The divisor $r_w$ is the saturation mixing ratio, or the vapour capacity. Relative humidity is, therefore, the water vapour content of the air relative to its content at saturation. Because the saturation mixing ratio is a function of pressure, and especially of temperature, the relative humidity is a combined index of the environment that reflects more than water content. In many climates the relative humidity rises to about 100 percent at dawn and falls to 50 percent by noon. A relative humidity of 50 percent may reflect many different quantities of vapour per volume of air or gram of air, and it will not likely be proportional to evaporation.

An understanding of relative humidity thus requires a knowledge of saturated vapour, which will be discussed later in the section on the relation between temperature and humidity. At this point, however, the relation between $U$ and the absorption and retention of water from the air must be considered. Small pores retain water more strongly than large pores; thus, when a porous material is set out in the air, all pores larger than a certain size (which can be calculated from the relative humidity of the air) are dried out.

The water content of a porous material at air temperature is fairly well indicated by the relative humidity. The complexity of actual pore sizes and the viscosity of the water passing through them makes the relation between $U$ and moisture in the porous material imperfect and slowly achieved. The great suction also strains the walls of the capillaries, and the consequent shrinkage is used to measure relative humidity.

The absorption of water by salt solutions is also related to relative humidity without much effect of temperature. The air above water saturated with sodium chloride is maintained at 75 to 76 percent relative humidity, at a temperature between 0° and 40° C.

In effect, relative humidity is a widely used environmental indicator, but $U$ does respond drastically to changes in temperatures as well as moisture, a response caused by the effect of temperature upon the divisor $r_w$ in $U$.

## RELATION BETWEEN TEMPERATURE AND HUMIDITY

Tables that show the effect of temperature upon the saturation mixing ratio $r_w$ are readily available. Humidity of the air at saturation is expressed more commonly, however, as vapour pressure. Thus, it is necessary to understand vapour pressure and in particular the gaseous nature of water vapour.

The pressure of the water vapour, which contributes to the pressure of the atmosphere, can be calculated from the absolute humidity $d_v$ by the gas equation:

$$e = \frac{m_v}{V}\frac{RT}{M_w} = d_v\frac{RT}{M_w},$$

in which $R$ is the gas constant, $T$ the absolute temperature, $M_w$ the molecular weight of water, and $e$ water vapour pressure in millibars.

Relative humidity can be defined as the ratio of the vapour pressure of a sample of air to the saturation pressure at the existing temperature. Further, the capacity for vapour and the effect of temperature can now be presented in the usual terms of saturation vapour pressure.

Within a pool of liquid water some molecules are continually escaping from the liquid into the space above, while more and more vapour molecules return to the liquid as the concentration of vapour rises. Finally, equal numbers are escaping and returning, the vapour is then saturated, and its pressure is known as the saturation vapour pressure $e_w$. If the liquid and vapour are warmed, relatively more molecules escape than return, and $e_w$ rises. There is also a saturation pressure with respect to ice. The vapour pressure curve of water has the same form as the curves for many other substances. Its location is fixed, however, by the boiling point of 100° C, where the saturation vapour pressure of water vapour is 1,013 millibars (mb), the standard pressure of the atmosphere at sea level. The decrease of the boiling point with altitude can be calculated. For example, the saturation vapour pressure at 40° C is 74 mb, standard atmospheric pressure near 18,000 metres above sea level is also 74 mb, and that is where water boils at 40° C.

The everyday response of relative humidity to temperature can be easily explained. On a summer morning the temperature might be 15° C and the relative humidity 100 percent. The vapour pressure would be 17 mb and the mixing ratio about 11 g/kg. During the day the air could warm to 25° C, while evaporation added little water. At 25° C the saturation pressure is fully 32 mb. If, however, little water has been added to the air, its vapour pressure will still be about 17 mb. Thus with no change in vapour content, the relative humidity of the air has fallen from 100 to only 53 percent, illustrating why relative humidity does not identify air masses.

**Dew-point temperature.** The meaning of dew-point temperature can be illustrated by a sample of air with a vapour pressure of 17 mb. If an object at 15° C is brought into the air, dew will form on the object. Hence, 15° C is the dew-point temperature of the air—*i.e.*, the temperature at which the vapour present in a sample of air would just cause saturation, or the temperature whose saturation

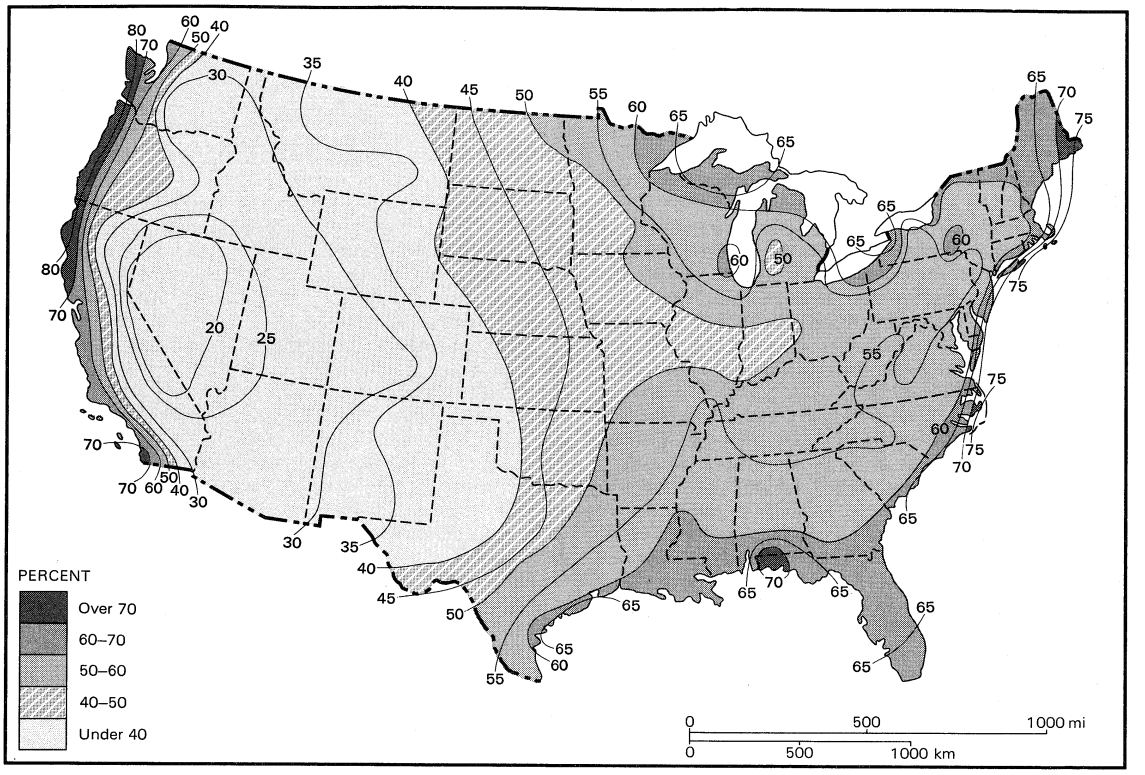



Figure 7: Average relative humidity, local noon, July.

vapour pressure equals the present vapour pressure in a sample of air, is the dew point. Below freezing, this index is called the frost point. There is a one-to-one correspondence between vapour pressure and dew point. The dew point has the virtue of being easily interpreted because it is the temperature at which a blade of grass or a pane of glass will become wet with dew from the air. Ideally, it is also the temperature of fog or cloud formation.

The clear meaning of dew point suggests a means of measuring humidity. A dew-point hygrometer was invented in 1751. In this instrument, cold water was added to water in a vessel until dew formed on the vessel, and the temperature of the vessel, the dew point, provided a direct index of humidity. The greatest use of the condensation hygrometer has been to measure humidity in the upper atmosphere where a vapour pressure of less than a thousandth millibar makes other means impractical.

The saturation deficit

Another index of humidity, the saturation deficit, can also be understood by considering air with a vapour pressure of 17 mb. At 25° C the air has (31 − 17) or 14 mb less vapour pressure than saturated vapour at the same temperature; that is, the saturation deficit is 14 mb.

The saturation deficit has the particular utility of being proportional to the evaporation capability of the air. The saturation deficit can be expressed as

$$e_w - e = e_w\left(1 - \frac{U}{100}\right),$$

and, because the saturation vapour pressure $e_w$ rises with rising temperature, the same relative humidity will correspond to a greater saturation deficit and evaporation at warm temperatures.

## HUMIDITY AND CLIMATE

The small amount of water in atmospheric vapour, relative to water on the Earth, belies its importance. Compared to one unit of water in the air, the seas contain at least 100,000 units, the great glaciers 1,500, the porous earth nearly 200, and rivers and lakes four or five units. The effectiveness of the vapour in the air is magnified, however, by its role in transferring water from sea to land by the media of clouds and precipitation and that of absorbing radiation.

The vapour in the air is the invisible conductor that carries water from sea to land, making terrestrial life possible. Fresh water is distilled from the salt seas and carried over land by the wind. Water evaporates from vegetation, and rain falls on the sea, too, but the sea is the bigger source, and rain that falls on land is most important to humans. The invisible vapour becomes visible near the surface as fog when the air cools to the dew point. The usual nocturnal cooling will produce fog patches in cool valleys. Or the vapour may move as a tropical air mass over cold land or sea, causing widespread and persistent fog, such as occurs over the Grand Banks off Newfoundland. The delivery of water by means of fog or dew is slight, however.

When air is lifted it is carried to a region of lower pressure where it will expand and cool as described by the gas equation. It may rise up a mountain slope or over the front of a cooler, denser air mass. If condensation nuclei are absent, the dew point may be exceeded by the cooling air, and the water vapour becomes supersaturated. If nuclei are present or if the temperature is very low, however, cloud droplets or ice crystals form, and the vapour is no longer in the invisible guise of atmospheric humidity.

The invisible vapour has another climatic role—namely, absorbing and emitting radiation. The temperature of the Earth and its daily variation is determined by the balance between incoming and outgoing radiation. The wavelength of the incoming radiation from the Sun is mostly shorter than three micrometres. It is scarcely absorbed by water vapour and its receipt depends largely upon cloud cover. The radiation exchanged between the atmosphere and Earth and the eventual loss to space is in the form of long waves. These long waves are strongly absorbed in the 3- to 8.5-micrometre band and in the greater than 11-micrometre range, where vapour is either partly or wholly opaque. As noted above, much of the radiation that is absorbed in the atmosphere is emitted back to Earth, and the surface receipt of long waves, primarily from water vapour and carbon dioxide in the atmosphere, is slightly more than twice the direct receipt of solar radiation at the surface. Thus the invisible vapour in the atmosphere combines with clouds and the advection (horizontal movement) of air from different regions to control the surface temperature.

The world distribution of humidity can be portrayed for different uses by different indices. To appraise the quantity

World distribution of humidity

of water carried by the entire atmosphere, the moisture in an air column above a given point on Earth is expressed as a depth of liquid water. It varies from 0.5 millimetre (0.02 inch) over the Himalayas and 2 mm over the poles in winter, to 8 mm over the Sahara, 54 mm in the Amazon region, and 64 mm over India during the wet season. During summer, the air over the United States transports 16 mm of water vapour over the Great Basin and 45 mm over Florida.

The humidity of the surface air may be mapped as vapour pressure, but a map of this variable looks much like that of temperature. Warm places are moist and cool ones are dry; even in deserts the vapour pressure is normally 13 millibars, whereas over the northern seas it is only about four millibars. Certainly the moisture in materials in two such areas will be just the opposite, and relative humidity is a more widely useful index.

The average relative humidity for July reveals the humidity provinces of the Northern Hemisphere when aridity is at a maximum. At other times the relative humidity generally will be higher. The humidities over the Southern Hemisphere in July indicate the humidities that comparable regions in the Northern Hemisphere will attain in January, just as July in the Northern Hemisphere suggests the humidities in the Southern Hemisphere during January. A contrast is provided by comparing a humid cool coast to a desert. The midday humidity on the Oregon coast, for example, falls only to 80 percent at midday, whereas in the Nevada desert it falls to 20 percent. At night the contrast is less, with averages being over 90 and about 50 percent in these two places.

Although the dramatic regular decrease of relative humidity from dawn to midday has been attributed largely to warming rather than declining vapour content, the content does vary regularly. In humid environments, daytime evaporation increases the water vapour content of the air, and the mixing ratio, which may be about 12 grams per kilogram, rises 1 or 2 g/kg in temperate places and may attain 16 g/kg in a tropical rain forest. In arid environments, however, little evaporation moistens the air and daytime turbulence tends to bring down dry air; this decreases the mixing ratio by as much as 2 g/kg.

Humidity also varies regularly with altitude. On the average, fully half the water in the atmosphere lies below 0.25 kilometre, and satellite observations over the United States in April revealed one millimetre or less of water in all the air above six kilometres. A cross section of the atmosphere along 75° W longitude shows a decrease in humidity with height and toward the poles. The mixing ratio is 16 g/kg just north of the Equator, but it decreases to 1 g/kg at 50° N latitude or eight kilometres above the Equator. The transparent air surrounding mountains in fair weather is very dry indeed.

Closer to the ground, the water vapour content also changes with height in a regular pattern. When water vapour is condensing on the Earth at night, the content is greater aloft than at the ground; during the day the content is in most cases less aloft than at the ground because of evaporation.

Evaporation and humidity

Evaporation, mostly from the sea and from vegetation, replenishes the humidity of the air. It is the change of liquid water into gaseous state, but it may be analyzed as diffusion. The rate of diffusion, or evaporation, will be proportional to the difference between the pressure of the water vapour in the free air and the vapour that is next to, and saturated by, the evaporating liquid. If the liquid and air have the same temperature, evaporation is proportional to the saturation deficit. It is also proportional to the conductivity of the medium between the evaporator and the free air. If the evaporator is open water, the conductivity will increase with ventilation. But if the evaporator is a leaf, the diffusing water must pass through the still air within the minute pores between the water within and the dry air outside. In this case, the porosity may modify the conductivity more than ventilation.

The temperature of the evaporator is rarely the same as the air temperature, however, because each gram of evaporation consumes about 600 calories and thus cools the evaporator. The availability of energy to heat the evaporator, therefore, is as important as the saturation deficit and conductivity of the air. Outdoors, some of this heat may be transferred from the surrounding air by convection, but much of it must be furnished by radiation. Evaporation is faster on sunny days than on cloudy ones not only because the sunny day may have drier air but also because the Sun warms the evaporator, thus raising the vapour pressure at the evaporator. In fact, according to the well-known Penman calculation of evaporation, this loss of water is essentially determined by the net radiation balance during the day. (P.E.W./Ed.)

From H.H. Lamb, *Climate: Present, Past and Future;* Methuen & Co. Ltd.

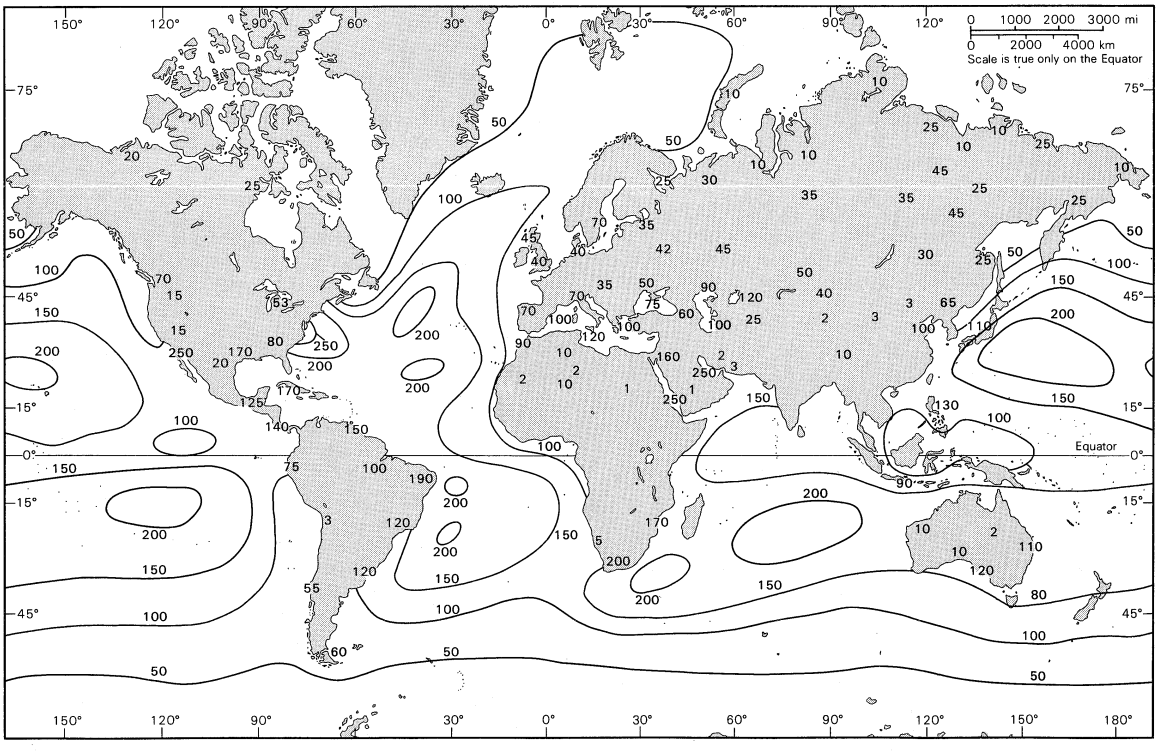



Figure 8: Average annual evaporation in centimetres.

## Precipitation

Precipitation is one of the three main processes (evaporation, condensation, and precipitation) that constitute the hydrologic cycle, the continual exchange of water between the atmosphere and the surface of the Earth. Water is evaporated from ocean, land, and freshwater surfaces, is carried aloft as vapour by the air currents, condenses to form clouds, and ultimately is returned to the Earth's surface as precipitation. The average global stock of water vapour in the atmosphere is equivalent to a layer of water 2.5 centimetres (one inch) deep covering the whole Earth. Because the Earth's average annual rainfall is about 100 centimetres, the average time that the water spends in the atmosphere, between its evaporation from the surface and its return as precipitation, is about 1/40 of a year, or about nine days. Of all the water vapour that is carried at all heights across a given region by the winds, only a small percentage is converted into precipitation and reaches the ground in that area. In deep and extensive cloud systems, the conversion is more efficient, but even in thunderclouds the quantities of rain and hail released amount to only some 10 percent of the total moisture entering the storm.

Measurement of amount and intensity

In the measurement of precipitation, it is necessary to distinguish between the amount—defined as the depth of precipitation (calculated as though it were all rain) that has fallen at a given point during a specified interval of time—and the rate or intensity—which specifies the depth of water that has fallen at a point during a particular interval of time. Persistent moderate rain, for example, might fall at an average rate of five millimetres per hour and thus produce 120 millimetres of rain in 24 hours. A thunderstorm might produce this total quantity of rain in 20 minutes, but at its peak intensity the rate of rainfall might become much greater—perhaps 120 millimetres per hour, or two millimetres per minute, for a minute or two.

The amount of precipitation falling during a fixed period is measured regularly at many thousands of places on the Earth's surface by rather simple rain gauges. Measurement of precipitation intensity requires a recording rain gauge, in which water falling into a collector of known surface area is continuously recorded on a moving chart or a magnetic tape. Investigations are being carried out on the feasibility of obtaining continuous measurements of rainfall over large catchment areas by means of radar.

Apart from the trifling contributions made by dew, frost, and rime and by desalination plants, the sole source of fresh water for sustaining rivers, lakes, and all life on Earth is provided by precipitation from clouds. Precipitation is therefore indispensable and overwhelmingly beneficial to humankind, but extremely heavy rainfall can cause great harm: soil erosion, landslides, and flooding. And hailstorm damage to crops, buildings, and livestock can prove very costly.

### ORIGIN OF PRECIPITATION IN CLOUDS

**Cloud formation.** Clouds are formed by the lifting of damp air, which cools by expansion as it encounters the lower pressures existing at higher levels in the atmosphere. The relative humidity increases until the air becomes saturated with water vapour, then condensation occurs on any of the aerosol particles suspended in the air. A wide variety of these exist in concentrations ranging from only a few per cubic centimetre in clean maritime air to perhaps 1,000,000 per cubic centimetre (16,000,000 per cubic inch) in the highly polluted air of an industrial city. For continuous condensation leading to the formation of cloud droplets, the air must be slightly supersaturated.

Condensation nuclei

Among the highly efficient condensation nuclei are sea-salt particles and the particles produced by combustion (*e.g.,* natural forest fires and man-made fires). Many of the larger condensation nuclei over land consist of ammonium sulfate. These are produced by cloud and fog droplets absorbing sulfur dioxide and ammonia from the air. Condensation onto the nuclei continues as rapidly as water vapour is made available through cooling; droplets about 10 micrometres in diameter are produced in this manner. These droplets constitute a nonprecipitating cloud.

**Cloud types.** The meteorologist classifies clouds mainly by their appearance, according to an international system similar to one proposed in 1803. But because the dimensions, shape, structure, and texture of clouds are influenced by the kind of air movements that result in their formation and growth, and by the properties of the cloud particles, much of what was originally a purely visual classification can now be justified on physical grounds.

Classification of clouds

The first *International Cloud Atlas* was published in 1896. Developments in aviation during World War I stimulated interest in cloud formations and in their importance as an aid in short-range weather forecasting. This led to the publication of a more extensive atlas, the *International Atlas of Clouds and States of Sky,* in 1932 and to a revised edition in 1939. After World War II, the World Meteorological Organization published a new *International Cloud Atlas* (1956) in two volumes. It contained 224 plates, describing 10 main cloud genera (families) subdivided into 14 species based on cloud shape and structure. Nine general varieties, based on transparency and geometric arrangement, also are described. The genera, listed according to their height, are as follows:

1. High: mean heights from 5 to 13 kilometres
   - *a.* Cirrus
   - *b.* Cirrocumulus
   - *c.* Cirrostratus
2. Middle: mean heights 2 to 7 kilometres
   - *a.* Altocumulus
   - *b.* Altostratus
   - *c.* Nimbostratus
3. Low: mean heights 0 to 2 kilometres
   - *a.* Stratocumulus
   - *b.* Stratus
   - *c.* Cumulus
   - *d.* Cumulonimbus

Heights given are approximate averages for temperate latitudes. Clouds of each genus are generally lower in the polar regions and higher in the tropics. The definitions and descriptions of the cloud genera used in the *International Cloud Atlas* are given in the accompanying Figures 9–11, which illustrate some of their characteristic forms.

Four principal classes are recognized when clouds are classified according to the kind of air motions that produce them: (1) layer clouds formed by the widespread regular ascent of air; (2) layer clouds formed by widespread irregular stirring or turbulence; (3) cumuliform clouds formed by penetrative convection; (4) orographic clouds formed by ascent of air over hills and mountains.

Layer clouds formed by widespread regular ascent

The widespread layer clouds associated with cyclonic depressions (see below *Cyclones and anticyclones*), near fronts and other bad-weather systems, frequently are composed of several layers that may extend up to nine kilometres or more, separated by clear zones that become filled in as rain or snow develops. These clouds are formed by the slow, prolonged ascent of a deep layer of air, in which a rise of only a few centimetres per second is maintained for several hours. In the neighbourhood of fronts, vertical velocities become more pronounced and may reach about 10 centimetres per second.

Most of the high cirrus clouds visible from the ground lie on the fringes of cyclonic cloud systems, and, though due primarily to regular ascent, their pattern is often determined by local wave disturbances that finally trigger their formation after the air has been brought near its saturation point by the large-scale lifting.

On a cloudless night, the ground cools by radiating heat into space without heating the air adjacent to the ground. If the air were quite still, only a very thin layer would be chilled by contact with the ground. More usually, however, the lower layers of the air are stirred by motion over the rough ground, so that the cooling is distributed through a much greater depth. Consequently, when the air is damp or the cooling is great, a fog a few hundred metres deep may form, rather than a dew produced by condensation on the ground.

In moderate or strong winds the irregular stirring near the surface distributes the cooling upward, and the fog may lift from the surface to become a stratus cloud, which is not often more than 600 metres thick.

Radiational cooling from the upper surfaces of fogs and stratus clouds promotes an irregular convection within the

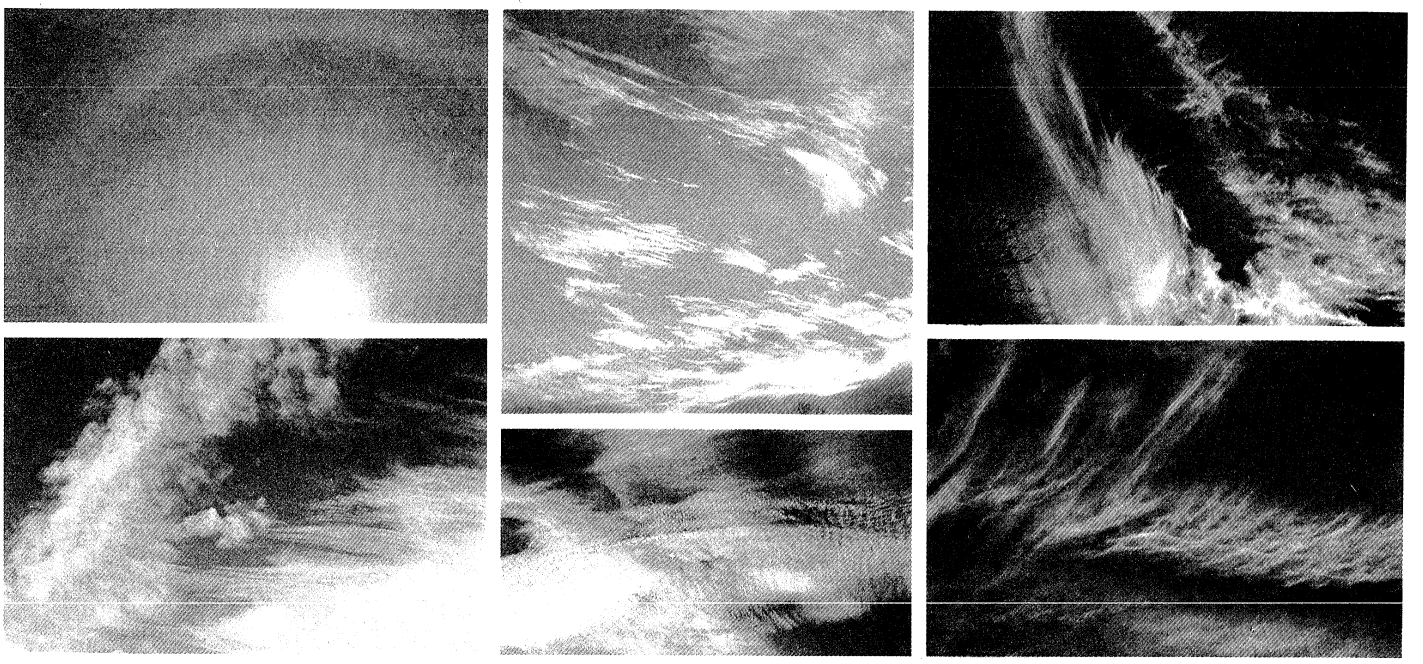


Figure 9: *High clouds.*
(Top left) Cirrostratus nebulosus, producing halo phenomenon. (Top centre) Cirrus fibratus, nearly straight or irregularly curved fine white filaments, generally distinct from one another. (Top right) Cirrus uncinus, detached clouds of delicate white filaments, often comma-shaped and ending at the top in a hook or tuft. (Bottom left) Cirrus spissatus, detached fibrous clouds of sufficient optical thickness to appear grayish when viewed against the Sun. (Bottom centre) Cirrocumulus, a thin white cloud patch composed of small elements in the form of ripples. (Bottom right) Cirrostratus fibratus, a thin whitish veil of nearly straight filaments.
(Top left, top right, bottom left, bottom centre) Louis D. Rubin, Richmond, Virginia, (top centre, bottom right) Photo Researchers, (top centre) Nick Impenna, (bottom right) John G. Ross

cloud layer and causes the surfaces to have a waved or humped appearance. When the cloud layer is shallow, billows and clear spaces may develop so that it is described as stratocumulus instead of stratus.

Convective clouds

Usually cumuliform clouds appearing over land are formed by the rise of discrete masses of air from near the Sun-warmed surface. These rising lumps of air, or thermals, may vary in diameter from a few tens to hundreds of metres as they ascend and mix with the cooler, drier air above them. Above the level of the cloud base the release of latent heat of condensation tends to increase the buoyancy of the rising masses, which tower upward and emerge at the top of the cloud with rounded upper surfaces.

At any moment a large cloud may contain a number of active thermals and the residues of earlier ones. A new thermal rising into a residual cloud will be partially protected from having to mix with the cool, dry environment and therefore may rise farther than its predecessor. Once a thermal emerges as a cloud turret at the summit or the flanks of the cloud, rapid evaporation of the droplets chills the cloud borders, destroys the buoyancy, and produces sinking. A cumulus thus has a characteristic pyramidal shape and, viewed from a distance, appears to have an unfolding motion, with fresh cloud masses continually emerging from the interior to form the summit and then sinking aside and evaporating.

In settled weather, cumulus clouds are well scattered and small; horizontal and vertical dimensions are only a kilometre or two. In disturbed weather they cover a large part of the sky, and individual clouds may tower as high as 10 kilometres or more, often ceasing their growth only upon reaching the stable stratosphere. These clouds produce heavy showers, hail, and thunderstorms (see below).

At the level of the cloud base the speed of the rising air masses is usually about one metre per second, but may reach five metres per second, and similar values are measured inside smaller clouds. The upcurrents in thunderclouds, however, often exceed five metres per second and may reach 30 metres per second or more.

The rather special orographic clouds are produced by the ascent of air over hills and mountains. The air stream is set into oscillation when it is forced over the hill, and the clouds form in the crests of the (almost) stationary waves. There may thus be a succession of such clouds stretching downwind of the mountain, which remain stationary relative to the ground in spite of strong winds that may be blowing through the clouds. The clouds have very smooth outlines and are called lenticular (lens-shaped) or wave clouds. Thin wave clouds may form at great heights (up to 10 kilometres, even over hills only a few hundred metres high) and occasionally are observed in the stratosphere (at 20 to 30 kilometres) over the mountains of Norway, Scotland, Iceland, and Alaska. These atmospheric wave clouds are known as nacreous, or "mother-of-pearl," clouds because of their brilliant iridescent colours.

Clouds produced by orographic disturbances

**Mechanisms of precipitation release.** Growing clouds are sustained by upward air currents, which may vary in strength from a few centimetres per second to several metres per second. Considerable growth of the cloud droplets (with falling speeds of only about one centimetre per second) is therefore necessary if they are to fall through the cloud, survive evaporation in the unsaturated air below, and reach the ground as drizzle or rain. The production of a few large particles from a large population of much smaller ones may be achieved in one of two ways. The first of these depends on the fact that cloud droplets are seldom of uniform size because droplets form on nuclei of various sizes and grow under slightly different conditions and for different lengths of time in different parts of the cloud. A droplet appreciably larger than average will fall faster than the smaller ones, and so will collide and fuse (coalesce) with some of those that it overtakes. Calculations show that, in a deep cloud containing strong upward air currents and high concentrations of liquid water, such a droplet will have a sufficiently long journey among its smaller neighbours to grow to raindrop size. This coalescence mechanism is responsible for the showers that fall in tropical and subtropical regions from clouds whose tops do not reach the 0° C level and therefore cannot contain ice crystals. Radar evidence also suggests that showers in temperate latitudes may sometimes be initiated by the coalescence of waterdrops, although the clouds may later reach heights at which ice crystals may form in their upper parts.

The coalescence process



Figure 10: *Middle clouds.*
(Top left) Altocumulus undulatus, a layer of shaded, regularly arranged rolls. (Top right) Altocumulus perlucidus, a white and gray layer in which there are spaces between the elements. (Bottom left) Altostratus translucidus, showing the Sun as if seen through ground glass. (Bottom right) Altocumulus radiatus, a layer with laminae arranged in parallel bands.
Photo Researchers, (top left) Richard Jepperson, (top right) H. von Meiss-Teuffen, (bottom left) Russ Kinne, (bottom right) Lawrence Smith

The ice crystal process

The second method of releasing precipitation can operate only if the cloud top reaches elevations at which temperatures are below 0° C and the droplets in the upper cloud regions become supercooled. At temperatures below −40° C the droplets freeze automatically or spontaneously. At higher temperatures they can freeze only if they are infected with special, minute particles called ice nuclei. The origin and nature of these nuclei are not known with certainty, but the most likely source is clay-silicate particles carried up from the ground by the wind. As the temperature falls below 0° C, more and more ice nuclei become active, and ice crystals appear in increasing numbers among the supercooled droplets. Such a mixture of supercooled droplets and ice crystals is unstable, however. The cloudy air is usually only slightly supersaturated with water vapour with respect to the droplets and is strongly oversaturated with respect to ice crystals; the latter thus grow more rapidly than the droplets. After several minutes the growing crystals acquire falling speeds of tens of centimetres per second, and several of them may become joined together to form a snowflake. In falling into the warmer regions of the cloud, this flake may melt and hit ground as a raindrop.

Precipitation from layer cloud systems

The deep, extensive, multilayer cloud systems, from which precipitation of a widespread, persistent character falls, are generally formed in cyclonic depressions (lows) and near fronts. Cloud systems of this type are associated with feeble upcurrents of only a few centimetres per second that last for at least several hours. Although the structure of these great rain-cloud systems is being explored by aircraft and radar, it is not yet well understood. That such systems rarely produce rain, as distinct from drizzle, unless their tops are colder than about −12° C suggests that ice crystals are mainly responsible. This view is supported by the fact that the radar signals from these clouds usually take a characteristic form that has been clearly identified with the melting of snowflakes.

**Showers, thunderstorms, and hail.** Precipitation from shower clouds and thunderstorms, whether in the form of raindrops, pellets of soft hail, or true hailstones, is generally of great intensity and shorter duration than that from layer clouds and is usually composed of larger particles. The clouds are characterized by their large vertical depth, strong vertical air currents, and high concentrations of liquid water, all factors favouring the rapid growth of precipitation elements by the accretion of cloud droplets.

In a cloud composed wholly of liquid water, raindrops may grow by coalescence. For example, a droplet being carried up from the cloud base grows as it ascends by sweeping up smaller droplets. When it becomes too heavy to be supported by the upcurrents, the droplet falls, continuing to grow by the same process on its downward journey. Finally, if the cloud is sufficiently deep, the droplet will emerge from its base as a raindrop.

In a dense, vigorous cloud several kilometres deep, the drop may attain its limiting stable diameter (about six millimetres) before reaching the cloud base and thus will break up into several large fragments. Each of these may continue to grow and attain breakup size. The number of raindrops may increase so rapidly in this manner that after a few minutes the accumulated mass of water can no longer be supported by the upcurrents and falls as a heavy shower. These conditions occur more readily in tropical regions. In temperate regions where the 0° C level is much lower in elevation, conditions are more favourable for the ice-crystal mechanism.

Growth of hailstones

The hailstones that fall from deep, vigorous clouds in warm weather consist of a core surrounded by several al-

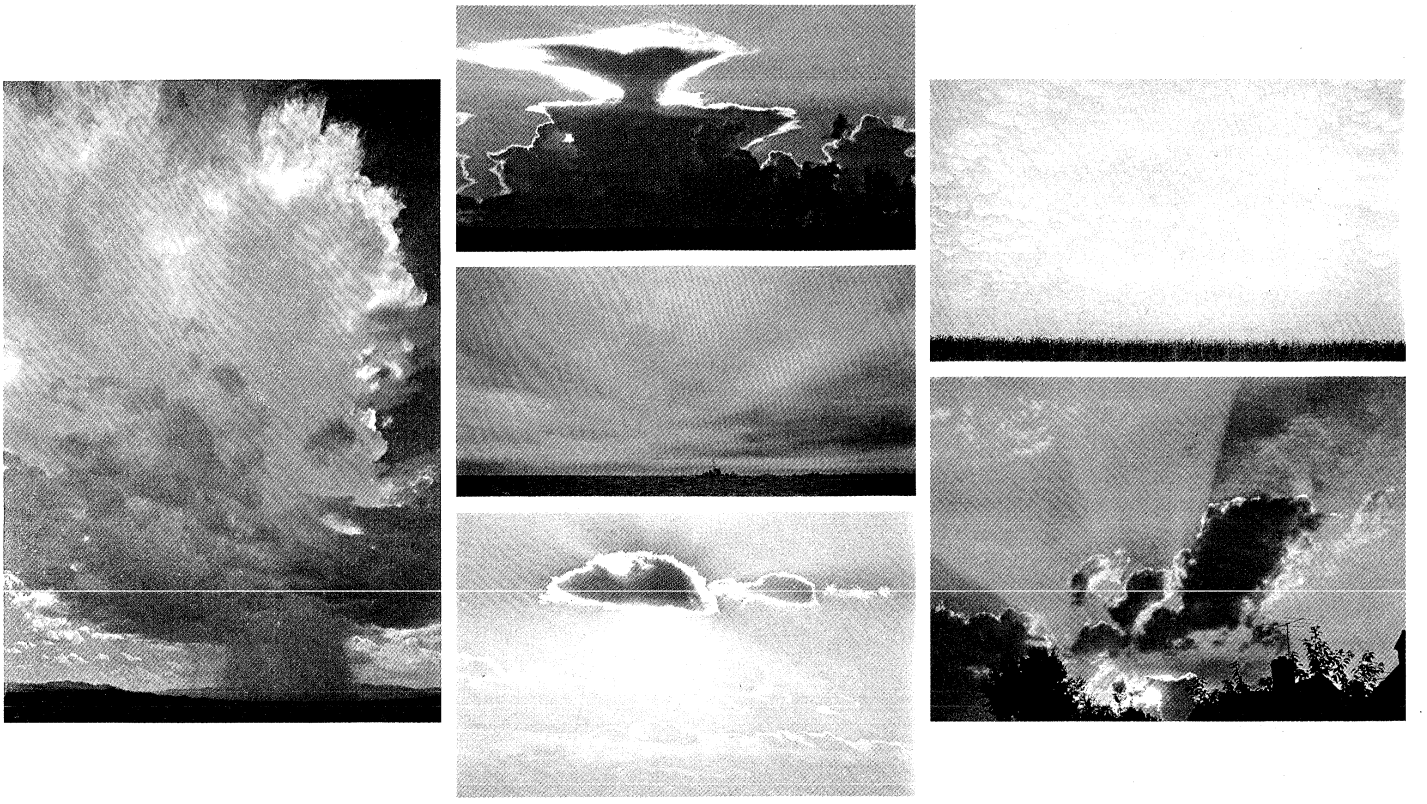


Figure 11: *Low clouds.*
(Left) Cumulonimbus calvus, a dense, heavy cloud with a considerable vertical extent, the upper portion of which has already lost its sharp outline. (Top centre) Cumulonimbus capillatus, showing the characteristic anvil-shaped upper portion, or thunderhead. (Top right) Cumulonimbus mamma, a light-coloured cloud sheet that has hanging protuberances on the undersurface. (Centre) Stratocumulus opacus, an extensive gray sheet with rounded masses, the greater part of which is sufficiently opaque to mask completely the Sun or Moon. (Bottom centre) Cumulus humilis, characterized by only a small vertical extent and appearing flattened. (Bottom right) Stratocumulus cumulogenitus (shown here with bright crepuscular rays), a gray layer with dark parts composed of elongated nonfibrous masses. These clouds represent a late stage of daytime development of cumulus.

(Left) William Belknap, Jr.—Rapho/Photo Researchers, (top centre, top right, centre, bottom centre) Photo Researchers, (top centre and centre) Russ Kinne, (top right) Irvin L. Oakes, (bottom centre) John G. Ross, (bottom right) Louis D. Rubin, Richmond, Virginia

ternate layers of clear and opaque ice. When the growing particle traverses a region of relatively high air temperature or high concentration of liquid water, or both, the transfer of heat from the hailstone to the air cannot occur rapidly enough to allow all of the deposited water to freeze immediately. This results in the formation of a wet coating of slushy ice, which may later freeze to form a layer of compact, relatively transparent ice. If the hailstone then enters a region of lower temperature and lower water content, the impacting droplets may freeze individually to produce ice of relatively low density with air spaces between the droplets. The alternate layers are formed as the stone passes through regions in which the combination of air temperature, liquid-water content, and updraft speed allows alternately wet and dry growth.

It is held by some authorities that lightning is closely associated with the appearance of precipitation, especially in the form of soft hail (see below), and that the charge and strong electric fields are produced by ice crystals or cloud droplets impacting on and bouncing off the undersurfaces of the hail pellets. For a detailed discussion of electrical effects in clouds, see below *Thunderstorms*.

### TYPES OF PRECIPITATION

**Drizzle.** Liquid precipitation in the form of very small drops, with diameters between 0.2 and 0.5 millimetre and terminal velocities between 70 and 200 centimetres per second, is defined as drizzle. It forms by the coalescence of even smaller droplets in low layer clouds containing weak updrafts of only a few centimetres per second. High relative humidity below the cloud base is required to prevent the drops from evaporating before reaching the ground; drizzle is classified as slight, moderate, or thick. Slight drizzle produces negligible runoff from the roofs of buildings, and thick drizzle accumulates at a rate in excess of one millimetre per hour.

**Rain and freezing rain.** Liquid waterdrops with diameters greater than those of drizzle constitute rain. Raindrops rarely exceed six millimetres in diameter because they become unstable when larger than this and break up during their fall. The terminal velocities of raindrops at ground level range from two metres per second for the smallest to about 10 metres per second for the largest. The smaller raindrops are kept nearly spherical by surface-tension forces, but, as the diameter surpasses about two millimetres, they become increasingly flattened by aerodynamic forces. When the diameter reaches six millimetres, the undersurface of the drop becomes concave because of the airstream and the surface of the drop is sheared off to form a rapidly expanding bubble or bag attached to an annular ring containing the bulk of the water. Eventually the bag bursts into a spray of fine droplets and the ring breaks up into a circlet of millimetre-sized drops.

Rain of a given intensity is composed of a spectrum of drop sizes, the average and median drop diameters being larger in rains of greater intensity. The largest drops, which have a diameter greater than five millimetres, appear only in the heavy rains of intense storms.

When raindrops fall through a cold layer of air (colder than 0° C) and become supercooled, freezing rain occurs. The drops may freeze on impact with the ground to form a very slippery and dangerous "glazed" ice that is difficult to see because it is almost transparent.

**Snow and sleet.** Snow in the atmosphere can be subdivided into ice crystals and snowflakes. Ice crystals generally form on ice nuclei at temperatures appreciably below

the freezing point. Below −40° C water vapour can solidify without the presence of a nucleus. Snowflakes are aggregates of ice crystals that appear in an infinite variety of shapes, mainly at temperatures near the freezing point of water.

In British terminology, sleet is the term used to describe precipitation of snow and rain together or of snow melting as it falls. In the United States, it is used to denote partly frozen ice pellets.

**Physical properties and conditions of formation**

Snow crystals generally have a hexagonal pattern, often with beautifully intricate shapes. Three- and 12-branched forms occur occasionally. The hexagonal form of the atmospheric ice crystals, their varying size and shape notwithstanding, is an outward manifestation of an internal arrangement in which the oxygen atoms form an open lattice (network) with hexagonally symmetrical structure. According to a recent internationally accepted classification there are seven types of snow crystals: plates, stellars, columns, needles, spatial dendrites, capped columns, and irregular crystals. They are shown in Figure 12. The size and shape of the snow crystals depend mainly on the temperature of their formation and on the amount of water vapour that is available for deposition. The two principal influences are not independent; the possible water vapour admixture of the air decreases strongly with decreasing temperature. The vapour pressure in equilibrium with a level surface of pure ice is 50 times greater at −2° C than at −42° C, the likely limits of snow formation in the air. Crystal shape and temperature at formation are related in Table 2.

**Table 2: Relation of Crystal Shape and Temperature at Formation**

| temperature (°C) | form | temperature (°C) | form |
|---|---|---|---|
| 0 to −3 | thin hexagonal plates | −12 to −16 | dendritic crystals |
| −3 to −5 | needles | −16 to −25 | hexagonal plates |
| −5 to −8 | hollow, prismatic columns | −25 to −50 | hollow prisms |
| −8 to −12 | hexagonal plates | | |

At temperatures above about −40° C, the crystals form on nuclei of very small size that float in the air (heterogeneous nucleation). The nuclei consist predominantly of silicate minerals of terrestrial origin, mainly clay minerals and micas. At still lower temperatures ice may form directly from water vapour (homogeneous nucleation). The influence of the atmospheric water vapour depends mainly on its degree of supersaturation with respect to ice.

**Snow and ice crystals in clouds**

If the air contains a large excess of water vapour, the snow particles will grow fast and there may be a tendency for dendritic (branching) growth. With low temperature the excess water vapour tends to be small, and the crystals remain small. In relatively dry layers the snow particles generally have simple forms. Complicated forms of crystals will cling together with others to form snowflakes that consist occasionally of up to 100 crystals; the diameter of such flakes may be as large as 2.5 centimetres. This process will be furthered if the crystals are near freezing point and wet, possibly by collision with undercooled water droplets. If a crystal falls into a cloud with great numbers of such drops, it will sweep some of them up. Coming into contact with ice, they freeze and form an ice cover around the crystal. Such particles are called soft hail or graupel (see below).

Snow particles constitute the clouds of cirrus type—namely cirrus, cirrostratus, and cirrocumulus, and many clouds of alto type. Ice and snow clouds originate normally only at temperatures some degrees below the freezing point; they predominate at −20° C. In temperate and low latitudes these clouds occur in the higher layers of the troposphere. In tropical regions they hardly ever occur below 4,570 metres. On high mountains and particularly in polar regions, they can occur near the surface and may appear as ice fogs. If cold air near the ground is overlain by warmer air (a very common occurrence in polar regions, especially in winter), mixture at the border leads to supersaturation in the cold air. Small ice columns and needles, "diamond dust," will be formed and float down, glittering, even from a cloudless sky. In the coldest parts of Antarctica where temperatures near the surface are below −50° C on the average and rarely above −30° C, the formation of diamond dust is a common occurrence. The floating and falling ice crystals produce in the light of the Sun and Moon the manifold phenomena of atmospheric optics, halos, arcs, circles, mock suns, some coronas, and iridescent clouds. Most of the different optical appearances can be explained by the shapes of the crystals and their position with respect to the light source.

**Diamond dust**

Most of the moderate to heavy rain in temperate latitudes depends on the presence of ice and snow particles in clouds. In the free atmosphere, droplets of fluid water can be undercooled considerably; typical ice clouds originate mainly at a temperature near −20° C. At an identical temperature below freezing point the water molecules are kept more firmly in the solid than in the fluid state. The equilibrium pressure of the gaseous phase is smaller in contact with ice than with water. At −20° C, which is the temperature of the formation of typical ice clouds (cirrus), the equilibrium pressure with respect to undercooled water (relative humidity 100 percent) is 22 percent greater than the equilibrium pressure of the water vapour in contact with ice. Hence, with an excess of water vapour beyond the equilibrium state, the ice particles tend to incorporate more water vapour and to grow more rapidly than the water droplets.

Being larger and so less retarded by friction, the ice par-

Vincent J. Schaefer

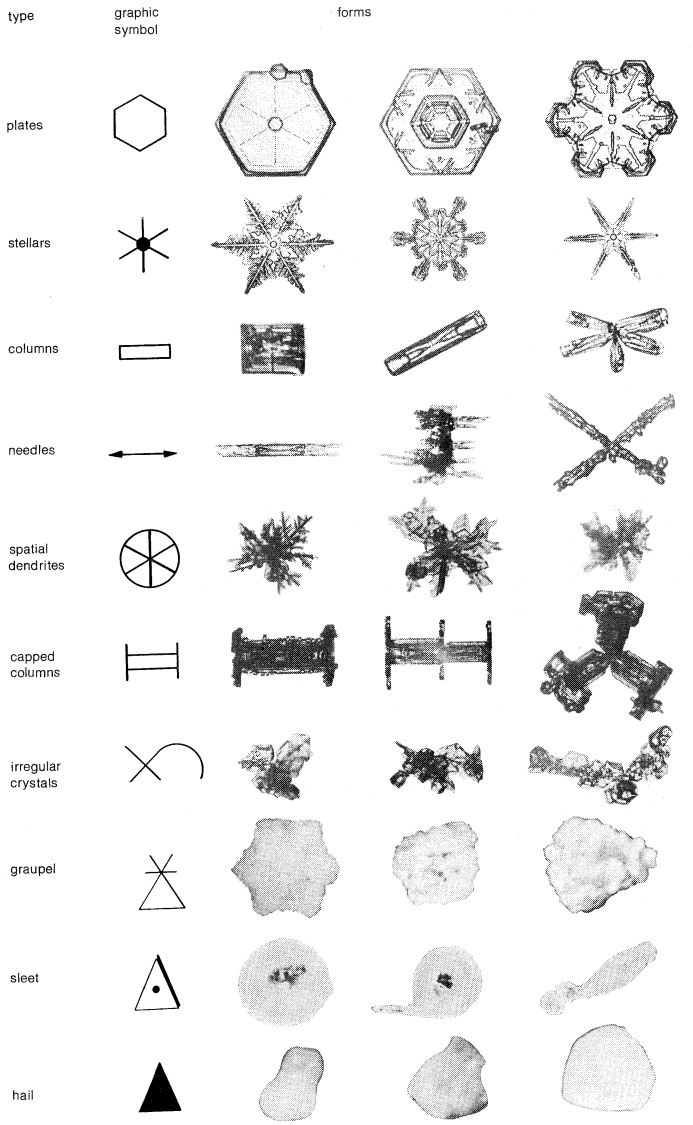



Figure 12: Classification of frozen precipitation.

ticles fall more rapidly. In their fall they sweep up some water droplets, which on contact become frozen. Thus a cloud layer originally consisting mainly of undercooled water with few ice crystals is transformed into an ice cloud. The development of the anvil shape at the top of a towering cumulonimbus cloud shows this transformation very clearly. The larger ice particles overcome more readily the rising tendency of the air in the cloud. Falling into lower levels they grow, aggregating with other crystals and possibly with water drops, melt, and form raindrops when near-surface temperatures permit.

Types of hail

**Hail.** Solid precipitation in the form of hard pellets of ice that fall from cumulonimbus clouds is hail. It is convenient to distinguish among three types of hail particles.

The first is soft hail, or snow pellets, which are white, opaque, rounded or conical pellets as large as six millimetres in diameter. They are composed of small cloud droplets frozen together, have a low density, and are readily crushed.

Second is small hail (ice grains or pellets), transparent or translucent pellets of ice that are spherical, spheroidal, conical, or irregular in shape, with diameters of a few millimetres. They may consist of frozen raindrops, of largely melted and refrozen snowflakes, or of snow pellets encased in a thin layer of solid ice.

True hailstones, the third type, are hard pellets of ice, larger than five millimetres in diameter, that may be spherical, spheroidal, conical, discoidal, or irregular in shape and often have a structure of concentric layers of alternately clear and opaque ice (Figure 13). A moderately severe storm may produce stones a few centimetres in diameter, whereas a very severe storm may release stones with a maximum diameter of 10 centimetres or more. Large damaging hail falls most frequently in the continental areas of middle latitudes (*e.g.*, in the Nebraska–Wyoming–Colorado area of the United States, in South Africa, and in northern India) but is rare in equatorial regions. Terminal velocities of hailstones range from about five metres per second for the smallest stones to perhaps 40 metres per second for stones five centimetres in diameter.

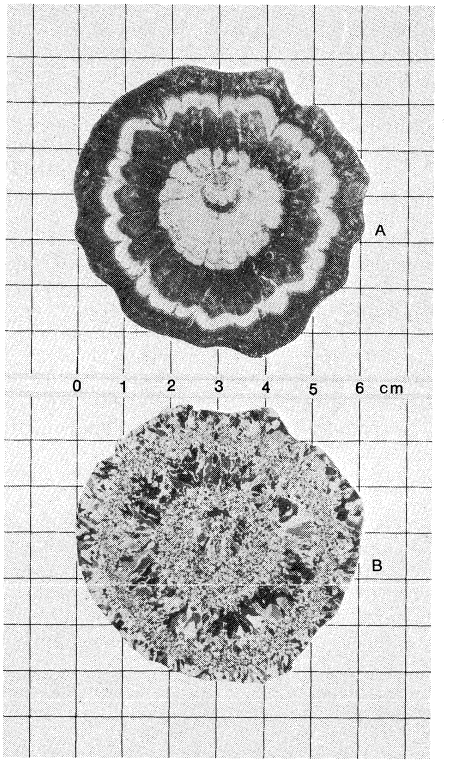



Figure 13: *Section through the centre of a large hailstone.*
(A) Alternate layers of opaque (white) and transparent (black) ice seen in reflected light. (B) The individual crystals seen in polarized light. The opaque layers are composed of small crystals and the transparent layers of much larger crystals.
By courtesy of K.A. Browning

## WORLD DISTRIBUTION OF PRECIPITATION

**Regional and latitudinal distribution.** The yearly precipitation averaged over the whole Earth is about 100 centimetres, but this is distributed very unevenly. The regions of highest rainfall are found in the equatorial zone and the monsoon area of Southeast Asia (Figure 14). Middle latitudes receive moderate amounts of precipitation, but little falls in the desert regions of the subtropics and around the poles.

From H.H. Lamb, *Climate: Present, Past and Future*; Methuen & Co. Ltd.

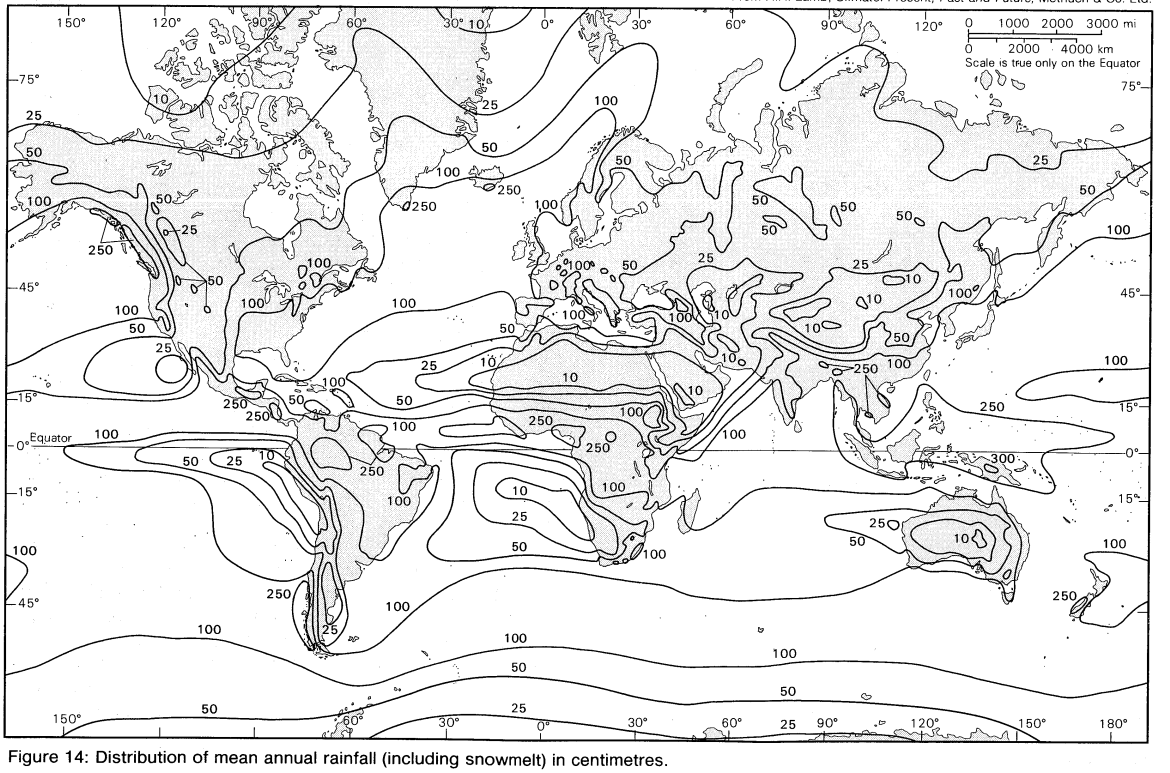



Figure 14: Distribution of mean annual rainfall (including snowmelt) in centimetres.

If the Earth's surface were perfectly uniform, the long-term average rainfall would be distributed in distinct latitudinal bands, but the situation is complicated by the pattern of the global winds, the distribution of land and sea, and the presence of mountains. Because rainfall results from the ascent and cooling of moist air, the areas of heavy rain indicate regions of rising air, whereas the deserts occur in regions in which the air is warmed and dried during descent. In the subtropics, the trade winds bring plentiful rain to the east coasts of the continents, but the west coasts tend to be dry. On the other hand, in high latitudes the west coasts are generally wetter than the east coasts. Rain tends to be abundant on the windward slopes of mountain ranges but sparse on the lee sides.

Equatorial and desert regions

In the equatorial belt, the trade winds from both hemispheres converge and give rise to a general upward motion of air, which becomes intensified locally in tropical storms that produce very heavy rains in the Caribbean, the Indian and southwest Pacific oceans, and the China Sea, and in thunderstorms that are especially frequent and active over the land areas. During the annual cycle, the doldrums move toward the summer hemisphere so that, outside a central region near the Equator, which has abundant rain at all seasons, there is a zone that receives much rain in summer but a good deal less in winter.

The dry areas of the subtropics, such as the desert regions of North Africa, the Arabian Peninsula, South Africa, Australia, and central South America, are due to the presence of semipermanent, subtropical anticyclones in which the air subsides and becomes warm and dry. These high-pressure belts tend to migrate with the seasons and cause summer dryness on the poleward side and winter dryness on the equatorward side of their mean positions (see below *Cyclones and anticyclones*). The easterly trade winds, having made a long passage over the warm oceans, bring plentiful rains to the east coasts of the subtropical landmasses, but the west coasts and the interiors of the continents, which are often sheltered by mountain ranges, are very dry.

Middle and high latitudes

In middle latitudes the weather and the rainfall are dominated by traveling depressions and fronts that yield a good deal of rain in all seasons and in most places except the far interiors of the Asian and North American continents. Generally the rainfall is more abundant in summer, except on the western coasts of North America, Europe, and North Africa, where it is higher during the winter.

At high latitudes and especially in the polar regions, the low precipitation is caused partly by subsidence of air in the high-pressure belts and partly by the low temperatures. Snow or rain occur at times (see Figures 14 and 15), but evaporation from the cold sea and land surfaces is slow, and the cold air has little capacity for moisture.

The influence of oceans and continents on rainfall is particularly striking in the case of the Indian monsoon. During the Northern Hemisphere winter, cool, dry air from the interior of the continent flows southward and produces little rain over the land areas. After the air has traveled some distance over the warm tropical ocean, however, it releases heavy shower rains over the East Indies. During the northern summer, when the monsoon blows from the southwest, rainfall is heavy over India and Southeast Asia. These rains are intensified where the air is forced to ascend over the windward slopes of the Western Ghāts and the Himalayas.

The combined effects of land, sea, mountains, and prevailing winds show up in South America. There the desert in southern Argentina is sheltered by the Andes from the westerly winds blowing from the Pacific Ocean, and the west-coast desert is not only situated under the South Pacific subtropical anticyclone but is also protected by the Andes against rain-bearing winds from the Atlantic.

**Amounts and variability.** The long-term average

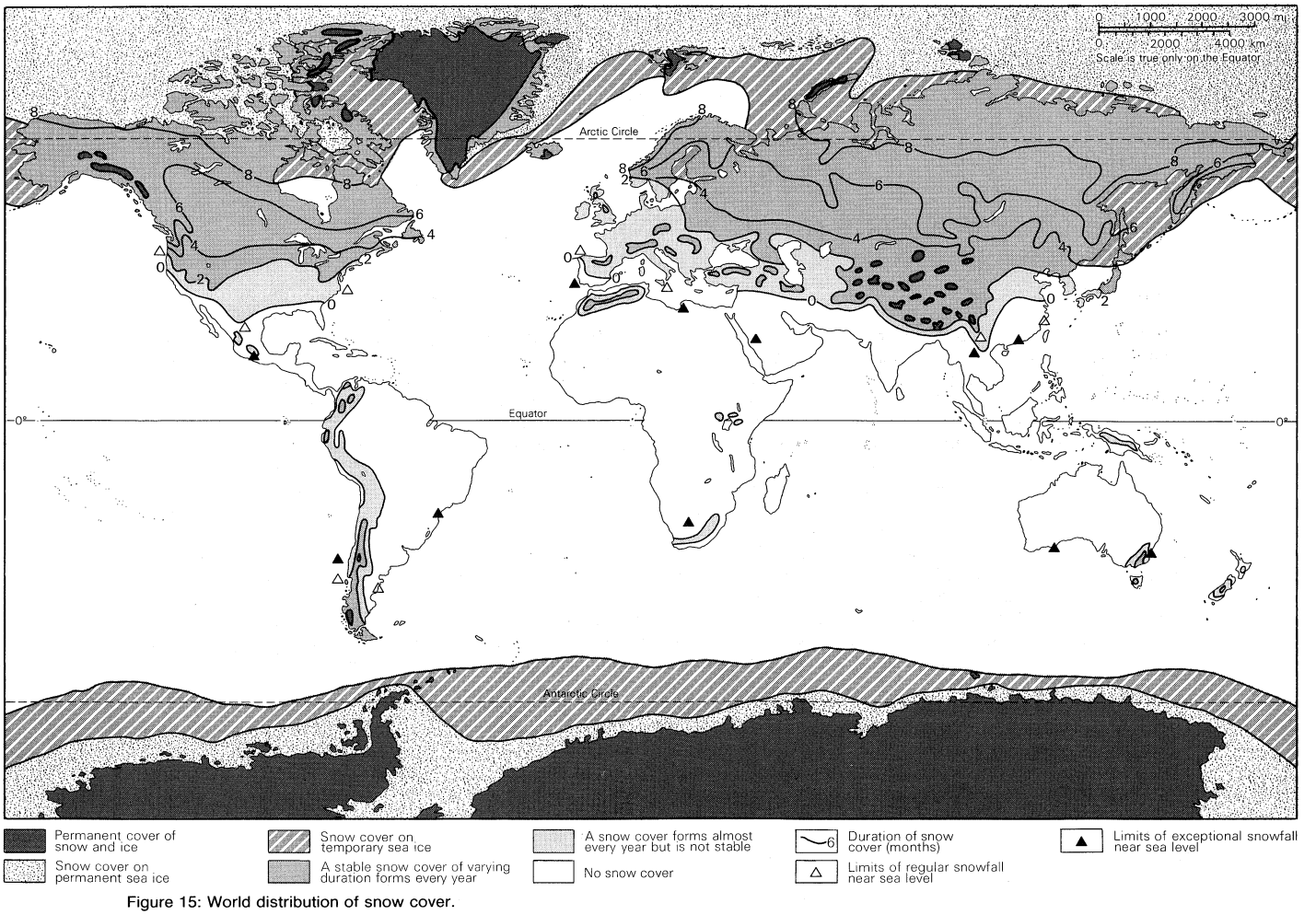



Figure 15: World distribution of snow cover.

amounts of precipitation for a season, or a year, give little information on the regularity with which rain may be expected, particularly for regions where the average amounts are small. For example, at Iquique, in northern Chile, four years once passed without rain, whereas the fifth year gave 15 millimetres; the five-year average was therefore three millimetres. Clearly, such averages are of little practical value, and the frequency distribution or the variability of precipitation also must be known.

Relation of variability and average amount

The variability of the annual rainfall is closely related to the average amounts. For example, over the British Isles, which have a very dependable rainfall, the annual amount varies by less than 10 percent above the long-term average value. A variability of less than 15 percent is typical of the mid-latitude cyclonic belts of the Pacific and Atlantic oceans and of much of the wet equatorial regions. In the interiors of the desert areas of Africa, Arabia, and Central Asia, however, the rainfall in a particular year may deviate from the normal long-term average by more than 40 percent. The variability for individual seasons or months may differ considerably from that for the year as a whole, but again the variability tends to be higher where the average amounts are low.

The heaviest annual rainfall in the world was recorded at Cherrapunji, India, where 26,470 millimetres fell between August 1860 and July 1861. The heaviest rainfall in a period of 24 hours was 1,870 millimetres recorded at Cilaos, Réunion, in the Indian Ocean on March 15–16, 1952. The lowest recorded rainfall in the world occurred at Arica, in northern Chile. An annual average, taken over a 43-year period, was only 0.5 millimetre.

Although past records give some guide, it is not possible to estimate very precisely the maximum possible precipitation that may fall in a given locality during a specified interval of time. Much will depend on a favourable combination of several factors, including the properties of the storm and the effects of local topography. Thus it is possible only to make estimates that are based on analyses of past storms or on theoretical calculations that attempt to maximize statistically the various factors or the most effective combination of factors that are known to control the duration and intensity of the precipitation. For many important planning and design problems, however, estimates of the greatest precipitation to be expected at a given location within a specified number of years are required. In the design of a dam, the highest 24-hour rainfall to be expected once in 30 years over the whole catchment area might be relevant. For dealing with such problems, a great deal of work has been devoted to determining, from past records, the frequency with which rainfalls of given intensity and total amount may be expected to reoccur at particular locations and also to determining the statistics of rainfall for a specific area from measurements made at only a few points.

### EFFECTS OF PRECIPITATION

**Raindrop impact and soil erosion.** Large raindrops, up to six millimetres in diameter, have terminal velocities of about 10 metres per second and so may cause considerable compaction and erosion of the soil by their force of impact. The formation of a compacted crust makes it more difficult for air and water to reach the roots of plants and encourages the water to run off the surface and carry away the topsoil with it. In hilly and mountainous areas, heavy rain may turn the soil into mud and slurry, which may produce enormous erosion by mudflow generation. Rainwater running off hard, impervious surfaces or waterlogged soil may cause local flooding.

**Surface runoff.** The rainwater that is not evaporated or stored in the soil eventually runs off the surface and finds its way into rivers, streams, and lakes or percolates through the rocks and becomes stored in natural underground reservoirs. A given catchment area must achieve an overall balance such that precipitation (P) less evaporation of moisture from the surface (E) will equal storage in the ground (S) and runoff (R). This may be expressed: $P - E = S + R$. The runoff may be determined by measuring the flow of water in the rivers with stream gauges, and the precipitation measured by a network of rain gauges, but storage and evaporation are more difficult to estimate.

Of all the water that falls on the Earth's surface, the relative amounts that run off, evaporate, or seep into the ground vary so much for different areas that no firm figures can be given for the Earth as a whole. It has been estimated, however, that in the United States 10 to 50 percent of the rainfall runs off at once, 10 to 30 percent evaporates, and 40 to 60 percent is absorbed by the soil. Of the entire rainfall, 15 to 30 percent is estimated to be used by plants, either to form plant tissue or in transpiration. (B.J.M./F.P.L./Ed.)

# ATMOSPHERIC PRESSURE AND WIND

## Atmospheric pressure

Atmospheric pressure and wind are both significant controlling factors in the Earth's weather and climate. Although these two physical variables may at first glance appear to be quite different, they are in fact closely related. Wind exists because of horizontal and vertical differences (gradients) in pressure, yielding a correspondence that often makes it possible to use the pressure distribution as an alternative representation of atmospheric motions. Pressure is the force exerted on a unit area, and atmospheric pressure is equivalent to the weight of air above a given area on the Earth's surface or within its atmosphere. This pressure is usually expressed in millibars (one mb equals 1,000 dynes per square centimetre) or in kilopascals (kPa; one kPa equals 10,000 dynes per square centimetre). Distributions of pressure on a map are depicted by a series of curved lines called isobars, each of which connects points of equal pressure.

At sea level, the mean pressure is about 1,000 millibars (100 kilopascals), varying by less than 5 percent from this value at any given location or time. Figures 16 and 17 show mean sea-level pressure for the mid-winter and mid-summer months. Since these charts represent average values over several days, pressure features that are relatively consistent day after day emerge, while more transient, short-lived features are removed. Those that remain are known as semi-permanent pressure centres and are the source regions for major, relatively uniform bodies of air known as air masses. Warm, moist maritime tropical (mT) air forms over tropical or subtropical ocean waters in association with the high-pressure regions prominent there. Cool, moist maritime polar (mP) air, on the other hand, forms over the colder subpolar ocean waters just south and east of the large, winter oceanic low-pressure regions. Over the continents, cold, dry continental polar (cP) air forms in the high-pressure regions that are especially pronounced in winter, while hot, dry continental tropical (cT) air forms over hot, desertlike continental domains in summer in association with low-pressure areas sometimes called heat lows.

Air masses

A closer examination of Figures 16 and 17 reveals some interesting features. First, it is clear that sea-level pressure is dominated by closed high- and low-pressure centres, which are largely caused by differential surface heating between low and high latitudes and between continental and oceanic regions. High pressure tends to be amplified over the colder surface features. Second, because of seasonal changes in surface heating, the pressure centres exhibit seasonal changes in their characteristics. For example, the Siberian High, Aleutian Low, and Icelandic Low that are so prominent in the winter virtually disappear in summer as the continental regions warm relative to surrounding bodies of water. At the same time, the Pacific and Atlantic highs amplify and migrate northward.

At altitudes well above the Earth's surface (Figure 18), the monthly average pressure distributions show much less tendency to form in closed centres but rather ap-

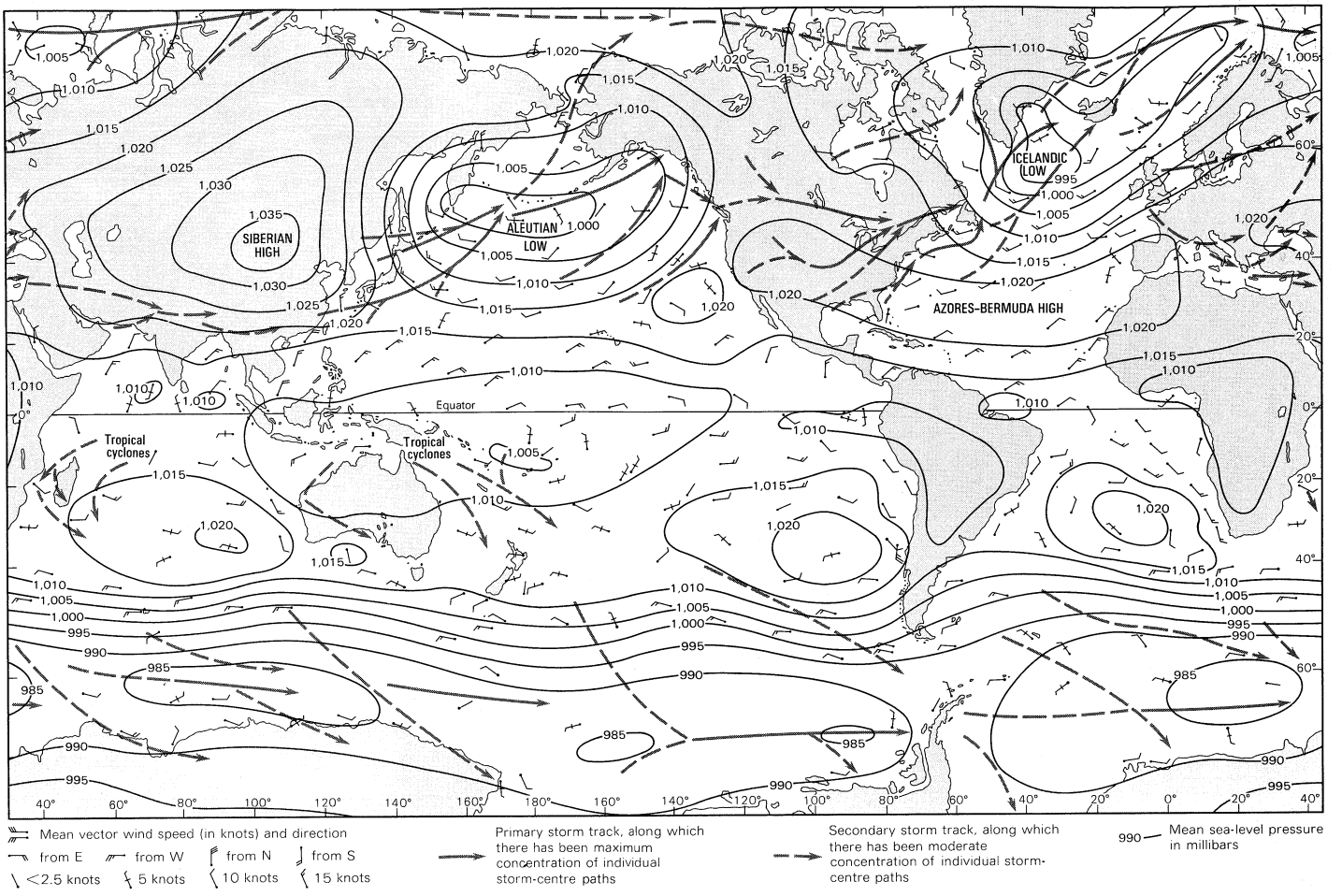



Figure 16: World distribution of mean sea-level pressure (in millibars) for January, and primary and secondary storm tracks. The general character of the global winds is also shown.
From H.L. Crutcher and O.M. Davis, *Navy Marine Climatic Atlas of the World*, vol. 8, NAVAIR 50-1C-54; U.S. Naval Weather Service Command

pear as quasi-concentric circles around the poles. This more symmetrical appearance reflects the dominant role of meridional radiative heating/cooling differences. Excess heating in tropical latitudes, compared to the polar, produces higher pressure at upper levels in the tropics. In addition, the greater heating/cooling contrast in winter yields stronger pressure differences during this season. Perfect symmetry is interrupted by superimposed wavelike disturbances. These are most pronounced over the Northern Hemisphere, with its more prominent land-ocean contrasts and orographic features.

## Wind

### RELATIONSHIP OF WIND TO PRESSURE AND GOVERNING FORCES

The changing wind patterns are governed by Newton's second law of motion, which states that the sum of the forces acting on a body equals the product of the mass of that body and the acceleration caused by those forces. The basic relationship between atmospheric pressure and horizontal wind is revealed by disregarding friction and any changes in wind direction and speed to yield the mathematical relationship

$$fu = -\frac{1}{\rho}\frac{\partial p}{\partial y} \quad \text{and} \quad fv = \frac{1}{\rho}\frac{\partial p}{\partial x}, \qquad (1)$$

where $u$ is the zonal wind speed (+ eastward), $v$ the meridional wind speed (+ northward), $f = 2\omega \sin \varphi$ (Coriolis parameter), $\omega$ the angular velocity of the Earth's rotation, $\varphi$ the latitude, $\rho$ the air density (mass per unit volume), $p$ the pressure, and $x$ and $y$ the distances toward the east and north, respectively. This simple, non-accelerating flow is known as geostrophic balance and yields a motion field known as the geostrophic wind. Equation (1) expresses, for both the $x$ and $y$ directions, a balance between the force created by horizontal differences in pressure (the horizontal pressure-gradient force) and the force that results from the Earth's rotation (the Coriolis force), as depicted is Figure 19. The pressure-gradient force expresses the tendency of pressure differences to effectuate air movement from higher to lower pressure. The Coriolis force is a more complicated concept that arises because the air motions are observed on a rotating, nearly spherical body. The total motion of a parcel of air has two parts: (1) the motion relative to the Earth as if the Earth were fixed, and (2) the motion given to the parcel by the planet's rotation. If an observer viewed the atmosphere from a fixed point in space, the rotation of the Earth would be apparent to him and he would observe the total motion. An observer on the ground, however, sees and measures only the relative motion and, as he also is rotating, cannot see directly the rotational motion applied by the Earth. Instead, he sees the effect of the rotation as a deviation applied to the relative motion. The quantity that causes this deviation is the so-called Coriolis force. More specifically, what the observer sees is a deflection of the relative motion to the right in the Northern Hemisphere and to the left in the Southern Hemisphere. Of particular significance in this simple model of wind–pressure relationships is the fact that the geostrophic wind blows in a direction parallel to the isobars with the low pressure on the observer's left as he looks downwind in the Northern Hemisphere and on his right in the Southern Hemisphere. Furthermore, the wind speed increases as the distance between isobars decreases (or pressure gradient increases). Curvature (*i.e.*, changes in wind direction) can be added to this model with relative

Geostrophic wind

Coriolis force

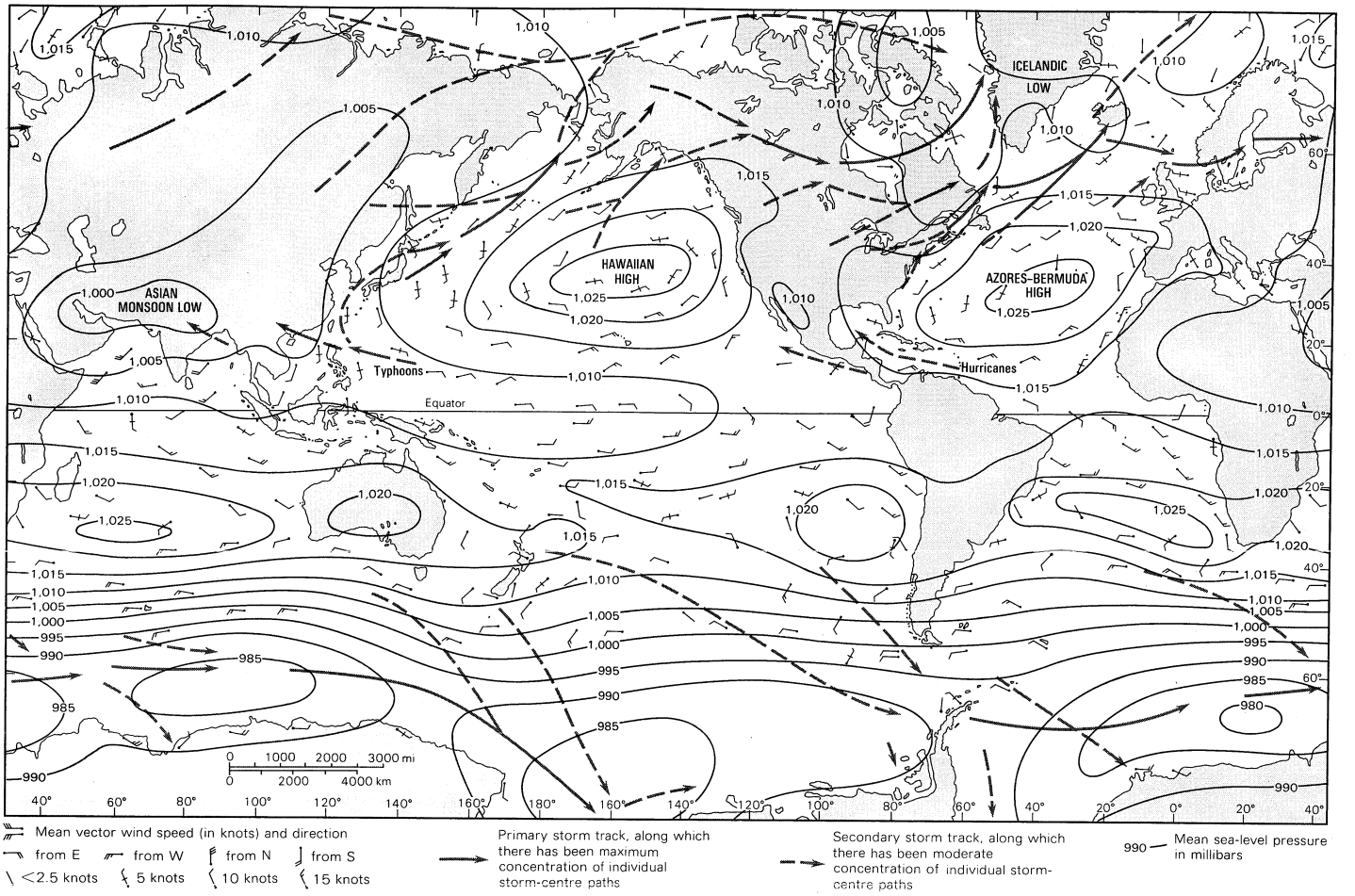



Figure 17: World distribution of mean sea-level pressure (in millibars) for July, and primary and secondary storm tracks. The general character of the global winds is also shown.
From H.L. Crutcher and O.M. Davis, *Navy Marine Climatic Atlas of the World*, vol. 8, NAVAIR 50-1C-54; U.S. Naval Weather Service Command

ease in a flow representation known as the gradient wind. The basic wind–pressure relationships, however, remain qualitatively the same. Of greatest importance is the fact that actual observed winds tend to behave much as the geostrophic- or gradient-flow models predict in most of the atmosphere.

The most notable exceptions are in low latitudes where the Coriolis parameter becomes very small and thus equation (1) cannot be used to provide a reliable wind estimate, and in the lowest kilometre of the atmosphere where friction becomes important. The friction induced by flow over the underlying surface reduces the wind speed and alters the simple balance of forces depicted in Figure 19 such that the wind blows with a component toward lower pressure, as shown in Figure 20.

## CYCLONES AND ANTICYCLONES

Simply stated, cyclones and anticyclones are regions of relatively low and high pressure, respectively. They occur over most of the Earth in a variety of sizes ranging from the very large, semipermanent examples described above to smaller, highly mobile systems. The latter are the focus of discussion in this section.

Common to both cyclones and anticyclones are the characteristic circulation patterns. The geostrophic-flow model (see above) dictates that in the Northern Hemisphere flow around a cyclone—cyclonic circulation—is counterclockwise and flow around an anticyclone—anticyclonic circulation—is clockwise (Figure 21). Circulation directions are reversed in the Southern Hemisphere (see Figures 16 and 17). In the presence of friction, the superimposed component of motion toward lower pressure produces a "spiraling" motion in toward the low centre and out away from the high centre.

The cyclones that form outside the equatorial belt, known as extratropical cyclones, may be regarded as large eddies in the broad air currents that flow in the general direction from west to east around the middle latitudes of both hemispheres (see below). They are an essential part of the mechanism by which the excess heat received from the Sun in the equatorial belt is conveyed toward higher latitudes. These higher latitudes radiate more heat to space than they receive from the Sun, and heat must reach them by winds from the equatorial belt if their temperature is to be maintained. If there were no cyclones and anticyclones, the north–south movements of the air would be much more limited, and there would be little opportunity for heat to be carried poleward by winds of subtropical origin. Under such circumstances the temperature of the equatorial regions would increase and the polar regions would cool; the temperature gradient between them would intensify.

Strong horizontal gradients of temperature are particularly favourable for the formation and development of cyclones. In the absence of cyclones and anticyclones the temperature difference between pole and Equator would tend to build up until it was sufficiently intense to generate new cyclones. The new cyclones themselves, however, would tend to reduce the temperature difference. Thus, the wind circulation on the Earth represents a balance between the effect of solar radiation in building a temperature difference between pole and Equator, and the effect of cyclones, anticyclones, and other wind systems in destroying this temperature contrast.

Cyclones of a somewhat different character occur closer to the Equator, forming in latitudes 10° to 15° N and S over the oceans. They generally are known as tropical cyclones but, more specifically, as hurricanes in the Atlantic and Caribbean, as typhoons in the western Pacific and China Sea, and as willy-willies off the coasts of Australia.

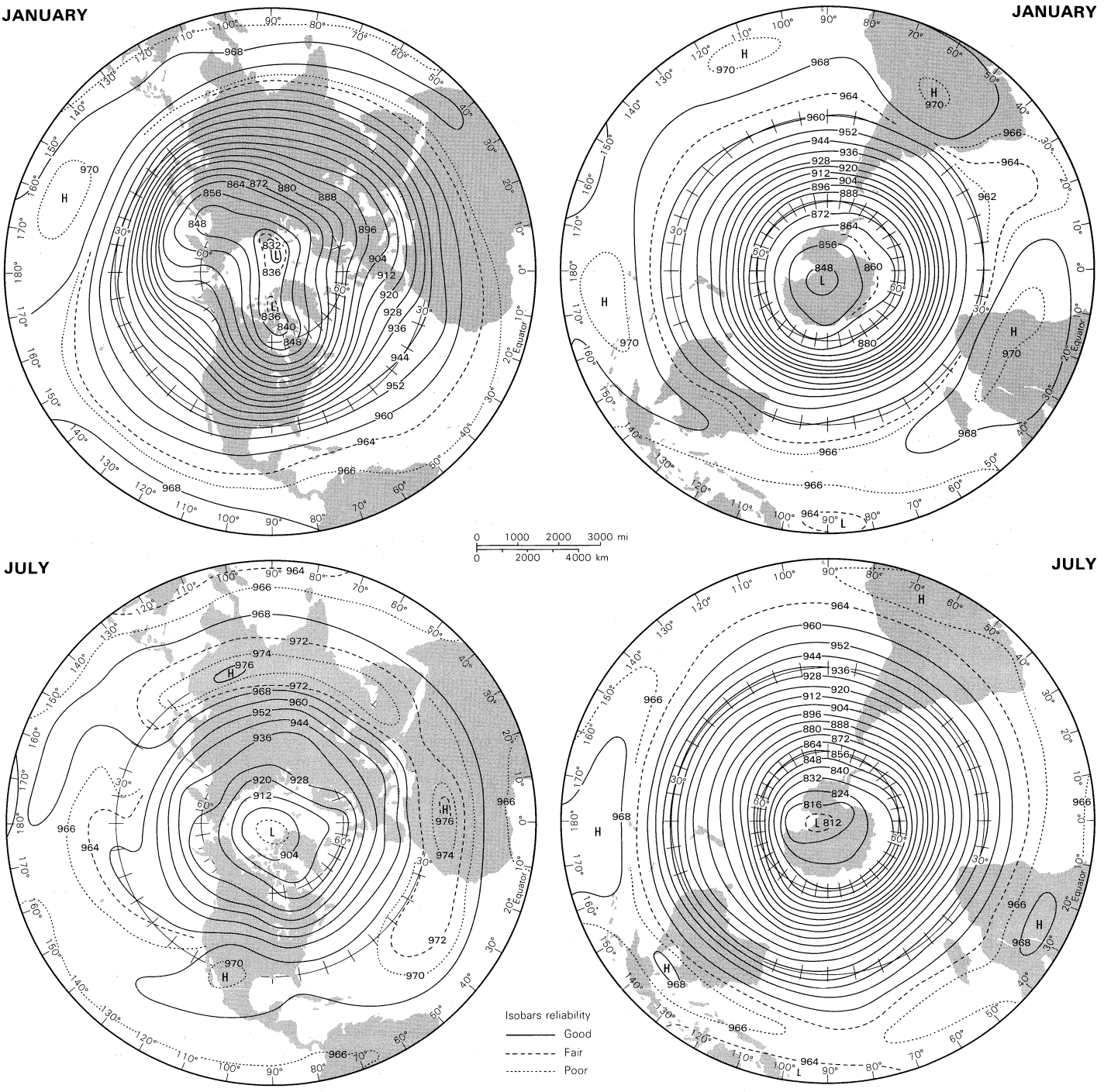



Figure 18: *Contours on the surface for which the pressure is 300 millibars.* Winds of the upper troposphere parallel these isobars. Note (left) the more marked seasonal change in pattern in the Northern Hemisphere.

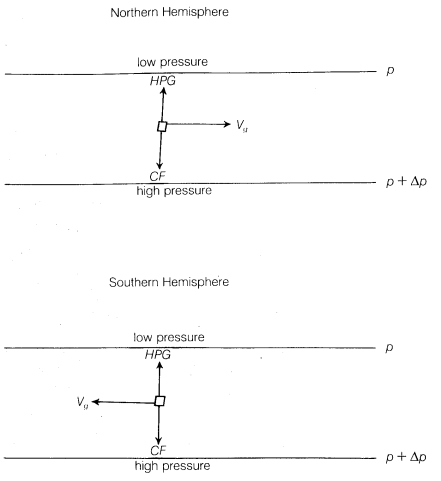
From H.L. Crutcher, R.L. Jenne, H. van Loon, and J.J. Taljaard, *Climate of the Upper Air*, part 1, *Southern Hemisphere*, NAVAIR 50-1C-55; U.S. Naval Weather Service Command

These storms are of smaller diameter than the extratropical cyclones, ranging from 100 to 500 kilometres in diameter, and are accompanied by winds that sometimes reach extreme violence. These storms are more fully described below in the section *Hurricanes and typhoons.*

**Extratropical cyclones.** Of the various examples of transient cyclones, extratropical cyclones are the most abundant and exert influence on the broadest scale, affecting the largest percentage of the Earth's surface. Furthermore, this class of cyclones is the principle cause of day-to-day weather changes experienced in middle and high latitudes and thus is the focal point of much of modern weather forecasting. The seeds for many current ideas concerning extratropical cyclones were sown between 1912 and 1930 by a group of Scandinavian meteorologists working in Bergen, Nor. This so-called Bergen school, founded by Vilhelm Bjerknes, formulated a model for a cyclone that forms as a disturbance along a zone of strong temperature contrast known as a front, which in turn constitutes a boundary between two contrasting air masses.

The model of the Bergen school

In this model proposed by the Bergen school, the masses of polar and mid-latitude air around the globe are separated by the polar front. The area bounding this region of strong temperature gradient has a reservoir of potential energy that can be readily tapped and converted into the kinetic energy associated with extratropical cyclones. For this reservoir to be tapped, a cyclone (called a wave, or frontal, cyclone) must develop much in the way shown in Figure 22. The feature that is of primary importance prior to cyclone development (cyclogenesis) is a front,




*HPG* = horizontal pressure gradient force
*CF* = Coriolis force
$V_g$ = geostrophic wind

Figure 19: *Balance of forces and resulting geostrophic wind.* A gradient is the rate of decrease of a quantity per unit distance over which that decrease occurs. It should be noted that *HPG* and *CF* are of equal magnitude but act in opposite directions (balanced forces) and that $V_g$ is perpendicular to both forces. Moreover, *HPG* acts from high to low pressure (*p*), and *CF* acts to the right of $V_g$ in the Northern Hemisphere and to the left in the Southern Hemisphere.

represented in Figure 22A as a heavy black line with alternating triangles or semicircles attached to it. This stationary or very slow-moving front forms a boundary between cold and warm air and so is a zone of strong horizontal temperature gradient (sometimes referred to as a baroclinic zone). Cyclone development is initiated as a disturbance along the front, which distorts the front into the wavelike configuration (22B). As the pressure within the disturbance continues to decrease, the disturbance assumes the appearance of a small cyclone and forces northward and southward movements of warm and cold air, respectively, which are represented by mobile frontal boundaries. As depicted in 22C, the front that signals the advancing cold air (cold front) is indicated by the triangles, while the front corresponding to the advancing warm air (warm front) is indicated by the semicircles. As the cyclone continues to intensify (22D), the cold, dense air streams rapidly southward, yielding a cold front with a typical slope of one to 50 and a propagation speed of eight to 15 metres per second (about 15 to 30 knots). At the same time, the warm, less dense air moving in a northerly direction flows up over the cold air east of the cyclone to produce a warm front with a typical slope of one to 200 and a much slower propagation speed of 2.5 to eight metres per second. This difference in propagation speeds allows the cold front to overtake the warm front to produce yet another, more complicated frontal structure known as an occluded front (22E, with alternating triangles and semicircles on the same side). This occlusion process may be followed by further storm intensification. The separation of the cyclone from the warm air to the south, however, eventually leads to the storm's decay (cyclolysis).

Occluded front

The life cycle of such an event is typically several days, during which the cyclone may travel from several hundred to a few thousand kilometres. In its path and wake occur dramatic weather changes. A typical sequence of weather that might be experienced as a cyclone and its fronts approach and pass through an area is depicted in Figure 23. Shown here is a cross section of the clouds and precipitation that usually occur along line *ab* in Figure 22D. Warm frontal weather is characterized by stratiform clouds, which descend as the front approaches and which eventually yield rain or snow. The passing of a warm front brings a rise in air temperature and clearing skies. The warmer air, however, may also harbour the ingredients for rain shower or thunderstorm formation, a condition that is enhanced as the cold front approaches. The passage of the cold front is marked by the influx of colder air, formation of stratocumulus cloud with some lingering rain or snow showers, and then eventual clearing. While this is an oft-repeated scenario, it is important to recognize that many other weather sequences can occur. For example, the stratiform clouds of a warm front may have imbedded cumulus formations and thunderstorms; the warm sector might be quite dry and yield few or no clouds; the pre-cold-front weather may closely resemble that found ahead of the warm front; or the post-cold-front air may be completely cloud-free. Cloud patterns oriented along fronts and spiraling around the cyclone vortex are consistently revealed in satellite pictures of the Earth (Figure 24).

The actual formation of any area of low pressure requires that mass in the column of air lying above the Earth's surface be reduced. This loss of mass then reduces the surface pressure. In the late 1930s and early 1940s, three members of the Bergen school, Jacob Bjerknes, J. Holmboe, and Carl-Gustaf Rossby, recognized that transient surfaces disturbances were accompanied by complementary wave features in the flow in the middle and higher atmospheric layers. These wave features (see below) are accompanied by regions of mass divergence and convergence that support the growth of surface-pressure fields and direct their movement.

While extratropical cyclones often form, intensify, or both in association with fronts, there are some that appear in the middle of a single air mass. A notable example is a class of cyclones, generally smaller than the frontal variety, that form in polar air streams in the wake of a frontal cyclone. These so-called polar lows are most prominent in subpolar marine environments and are thought to be caused by the transfer of heat and moisture from the warmer water surface into the overlying polar air and by supporting middle-tropospheric circulation features. Other cyclones form on the lee side of mountain barriers as the general westerly flow is disturbed by the mountain.

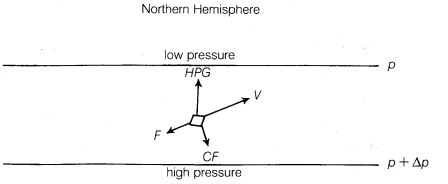
Polar lows

**Anticyclones.** While cyclones are typically regions of inclement weather induced by upward motions of five to 10 centimetres per second, anticyclones are usually meteorologically quiet regions. Generally larger than cyclones, anticyclones exhibit downward motions of one to three centimetres per second and yield dry stable air that may extend horizontally many hundreds of kilometres.

In most cases, an actively developing anticyclone forms in the region of cold air behind a cyclone as it moves away and before the next cyclone advances. Such an anticyclone is known as a cold anticyclone. A result of the downward air motion in an anticyclone, however, is compression of the descending air, and as a consequence of this compression the air is warmed. Thus after a few days the air composing the anticyclone at levels two to five kilometres above the ground tends to increase in temperature, and the anticyclone is transformed into a warm anticyclone. The air near the ground cannot descend and is not much affected by this process.

Types of anticyclones

Warm anticyclones move slowly, and cyclones are diverted around their periphery. During their transformation from cold to warm anticyclones, they usually move out of the main belt followed by cyclones in middle latitudes, often amalgamating with the quasi-permanent band of relatively high pressure that is found in both hemispheres around latitude 20° to 30°—the so-called subtropical anticyclones. On some occasions the warm anticyclones remain in the belt normally occupied by the mid-latitude westerly winds. The normal tracks of cyclones are then




*F* = friction
*V* = real wind speed
*CF* = Coriolis force
*HPG* = horizontal pressure gradient force

Figure 20: *Balance of forces with friction (F) included.* It can be seen that *V* has a component toward lower pressure and that *CF* remains perpendicular to *V* but is no longer equal and opposite to *HPG*.

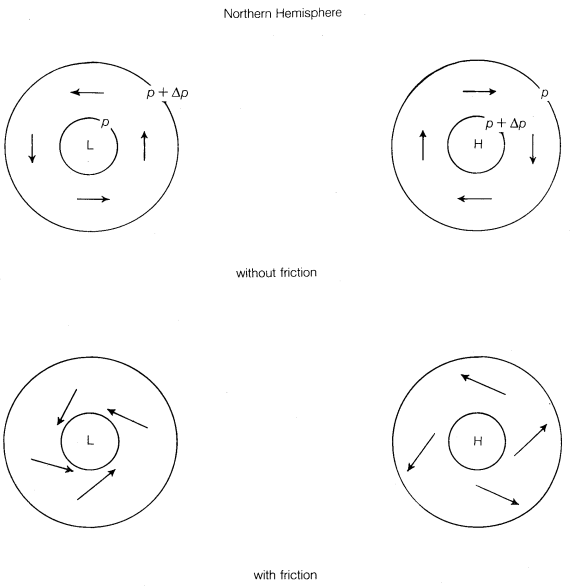



Figure 21: *Idealized depiction of flow around low (L) and high (H) pressure centres in the absence and presence of friction.* The solid lines are isobars. The arrows indicate the direction of flow in the Northern Hemisphere.

considerably modified; depressions are either blocked in their eastward progress or diverted to the north or south of the anticyclone. Anticyclones that interrupt the normal circulation of the westerly wind in this way are blocking anticyclones. They frequently persist for a week or more, and the occurrence of a few such blocking anticyclones may dominate the character of a season. Blocking anticyclones are particularly common over Europe, the eastern Atlantic, and the Alaskan area.

The descent and warming of the air in an anticyclone might be expected to lead to the dissolution of clouds and the absence of rain. This occurs often, particularly in summer. Since the air near the ground cannot take part in the descent, it may become stagnant. In winter the ground cools and the lower layers of the atmosphere also become cold. Fog may be formed as the air is cooled to its dew point. Under other circumstances the air trapped in the first kilometre above the surface may pick up moisture from the sea or moist surfaces, and layers of cloud may form from near the ground to a height of about one kilometre. Such layers of cloud can be persistent in anticyclones (except over the continents in summer), but they rarely grow thick enough to produce rain. If precipitation occurs, it is usually drizzle or light snow.

Anticyclones are thus regions of clear skies and warm sunny weather in summer; at other times of the year cloudy and foggy weather may be more typical. Winter anticyclones produce colder than normal temperatures at the surface, particularly if the skies remain clear. Anticyclones are responsible for periods of little or no rain, and such periods may be prolonged in association with blocking highs.

Cyclone tracks

**Cyclone and anticyclone climatology.** Migrating cyclones and anticyclones tend to be distributed around certain preferred regions, known as tracks, that emanate from preferred cyclogenetic regions. The contrast between winter and summer can be seen in Figures 16 and 17, which show cyclone tracks for January and July. Favoured cyclogenetic regions in the Northern Hemisphere are found on the lee side of mountains and off the east coasts of continents. Cyclones then track east or southeast before eventually turning northeastward and decaying. The tracks are also displaced farther northward in July, reflecting the more northward position of the polar front in summer. Continental cyclones usually intensify at a rate of 0.5 millibar per hour or less, although more dramatic examples can be found. Marine cyclones, on the other hand, often experience explosive development in excess of one millibar per hour, particularly in winter.

Anticyclones tend to migrate southeastward out of the cold air mass regions and then eastward before decaying or merging with a warm anticyclone, or they form in the wake of extratropical cyclones. These tracks also migrate northward with the warm season. Anticyclone tracks over North America and the surrounding ocean environment for January and July are shown in Figure 25.

In the Southern Hemisphere, where most of the Earth's surface is covered by oceans, the cyclones are distributed fairly uniformly through the various longitudes. Typically, cyclones form initially in latitudes 30° to 40° S and move in a generally southeastward direction, reaching maturity in latitudes around 60°. Thus the Antarctic continent is usually ringed by a number of mature or decaying cyclones, and the belt of ocean from 40° to 60° S is a region of persistent, strong westerly winds that form part of the circulation to the north of the main cyclone centres. These are the "roaring forties," where the westerly winds are interrupted only at intervals by the passage southeastward of developing cyclones.

Anticyclonic development seldom occurs over the southern oceans. Anticyclones thus are transient features of the weather regime in the Southern Hemisphere.

## LOCAL WINDS

**Scale classes.** Organized wind systems occur with dimensions ranging from tens of metres to thousands of kilometres and possess residence times that vary from seconds to weeks. The typical size and/or lifetime of a phenomenon is known as its scale. Since the atmosphere exhibits such a large variety of both spatial and temporal scales, efforts have been made to group various phenomena into scale classes. The class describing the largest and longest-lived of these phenomena is known as the planetary scale. Such phenomena are typically a few thousand kilometres in size and have lifetimes ranging from several days to several weeks. Examples are the semipermanent pressure centres discussed above and certain globe-encircling upper-air waves (see below *Upper-air waves*). Still another class is known as the synoptic scale. Spanning smaller distances, a few hundred to a few thousand kilometres, and possessing shorter lifetimes, a few to several days, this class contains the migrating cyclones and anticyclones that control day-to-day weather changes. Sometimes the planetary and synoptic scales are combined into a single

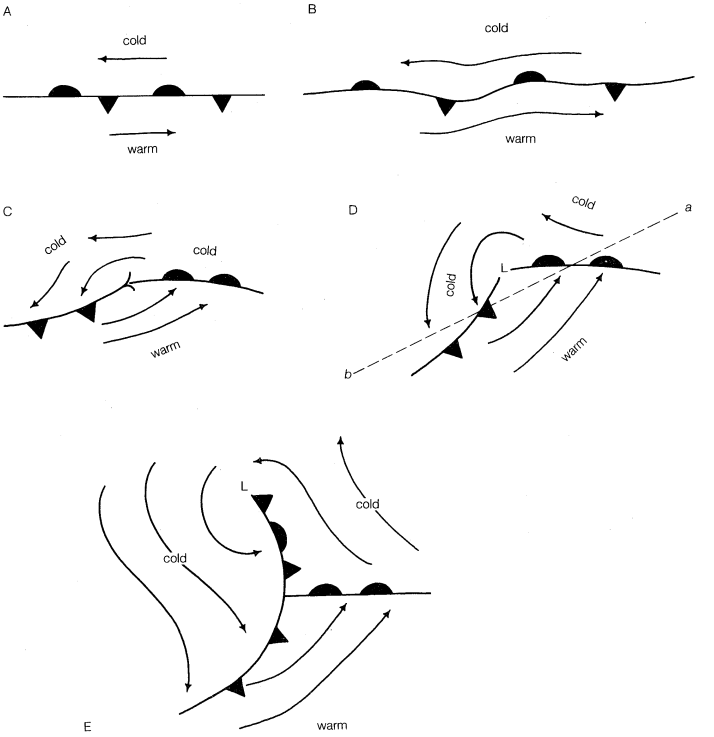



Figure 22: *Evolution of a wave (frontal) cyclone.* Dashed line *ab* in D is the cross-sectional line depicted in Figure 23 (see text).

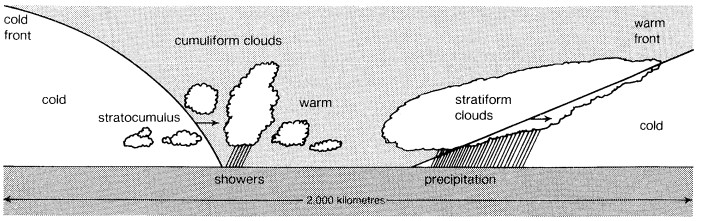



Figure 23: Cross section of clouds and precipitation often found along the cross-sectional line in Figure 22D. The direction of frontal movement is indicated by the arrows.

classification termed the large-scale, or macroscale. Large-scale wind systems are distinguished by the predominance of horizontal motions over vertical motions and by the preeminent importance of the Coriolis force in influencing wind characteristics.

There is a class of phenomena of even smaller size and shorter lifetime, whose vertical motions may be as significant as their horizontal motions and in which the Coriolis force often plays a less important role. Known as the mesoscale, this class is characterized by dimensions of tens to a few hundred kilometres and lifetimes of a day or less. Because of the shorter time scale and because the other forces may be much larger, the effect of the Coriolis force is sometimes neglected.

Two of the best-known examples of mesoscale phenomena are the thunderstorm and its devastating by-product, the tornado (see below *Thunderstorms* and *Tornadoes, whirlwinds, and waterspouts*). The present discussion focuses on less intense, though nevertheless commonly observed, wind systems that are found in rather specific geographic locations and thus are often referred to as local wind systems.

Sea/land breeze

**Local wind systems.** The first of these is the so-called sea/land breeze. This local wind system is typically encountered along coastlines adjacent to large bodies of water and is induced by differences that occur between the heating or cooling of the water surface and the adjacent land surface. Because water has a higher heat capacity (*i.e.*, more units of heat are required to produce a given temperature change) than do the materials in the land surface, daytime solar radiation heats the land surface more than the water surface. The higher-temperature land surface in turn transfers more heat to its overlying air mass and in so doing induces a circulation cell much like that depicted in Figure 26. It should be noted that the surface flow is from the water toward the land and thus is called a sea breeze. The landmass also cools more rapidly than does the water at night. Consequently, at night the cooler landmass yields a cooler overlying air mass and a circulation cell with air motions opposite to those found during the day (see Figure 26). This flow from land to water is known as a land breeze.

Sea/land breezes occur along the coastal regions of oceans or large lakes in the absence of a strong, large-scale wind system during periods of strong daytime heating or nighttime cooling. Those who live within 10 to 20 kilometres of the coastline will experience the cooler 19- to 37-kilometre-per-hour winds of the sea breeze on a sunny afternoon only to find it turning into a sultry land breeze late at night. One of the features of the sea/land breeze is a region of low-level air convergence in the termination region of the surface flow. Such convergence often induces local upward motions and cloud formations. Thus, in sea/land breeze regions, it is not uncommon to see clouds lying off the coast at night; these clouds are then dissipated by the daytime sea breeze, which forms new clouds, perhaps with precipitation occurring over land in the afternoon.

Another group of local winds is induced by the presence of mountain/valley features on the Earth's surface. One subset of such winds, known as mountain winds or breezes, is induced by differential heating or cooling along mountain slopes. During the day, solar heating of the sunlit slopes causes the overlying air to move upslope. At night, as the slopes cool, the motion is reversed and cool downslope motion occurs. Such winds may be relatively gentle or may occur in strong gusts, depending on the topographic configuration. In an enclosed valley the cool

By courtesy of Satellite Data Services Division, National Environmental Satellite, Data, and Information Service, National Oceanic and Atmospheric Administration

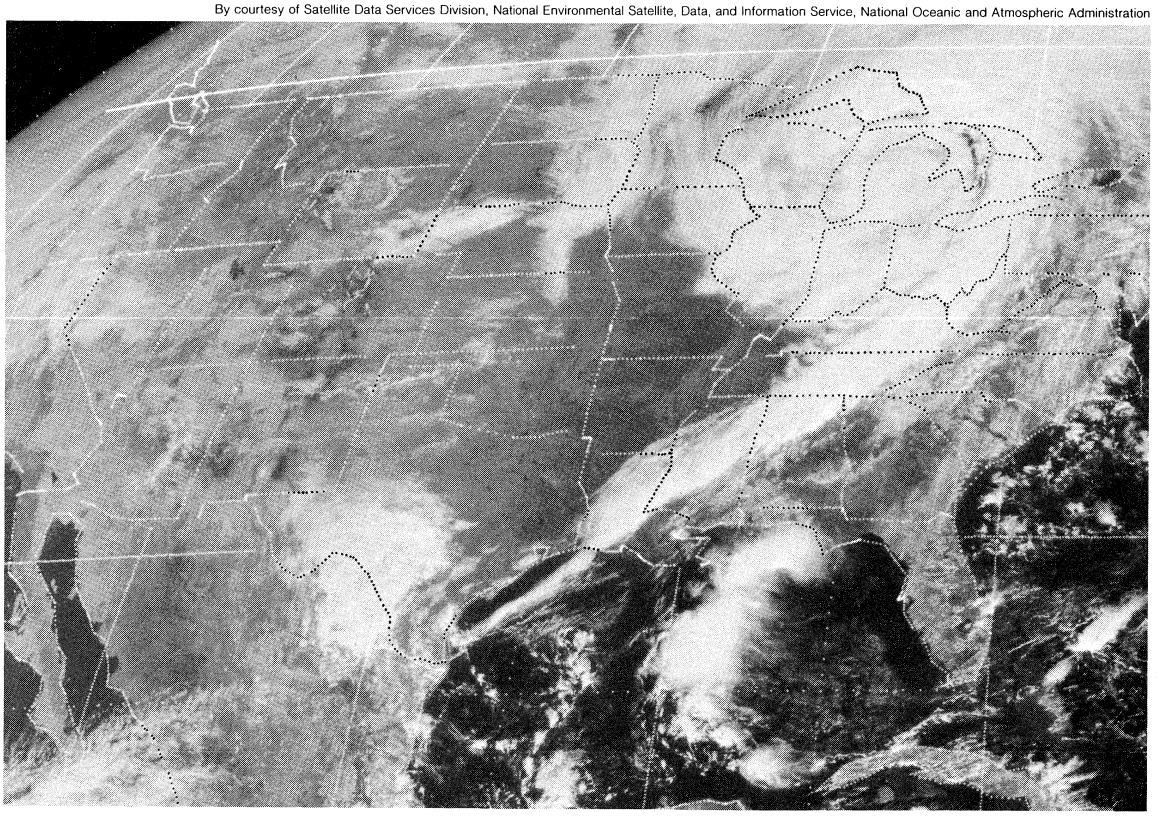


Figure 24: Image of cloud masses taken over the United States at 1701 Greenwich Mean Time on October 20, 1982, from the Geostationary Operational Environmental Satellite (GOES) EAST. At the right, clouds spiral around a cyclone over the Great Lakes and extend southwestward to the Gulf of Mexico along a cold front.

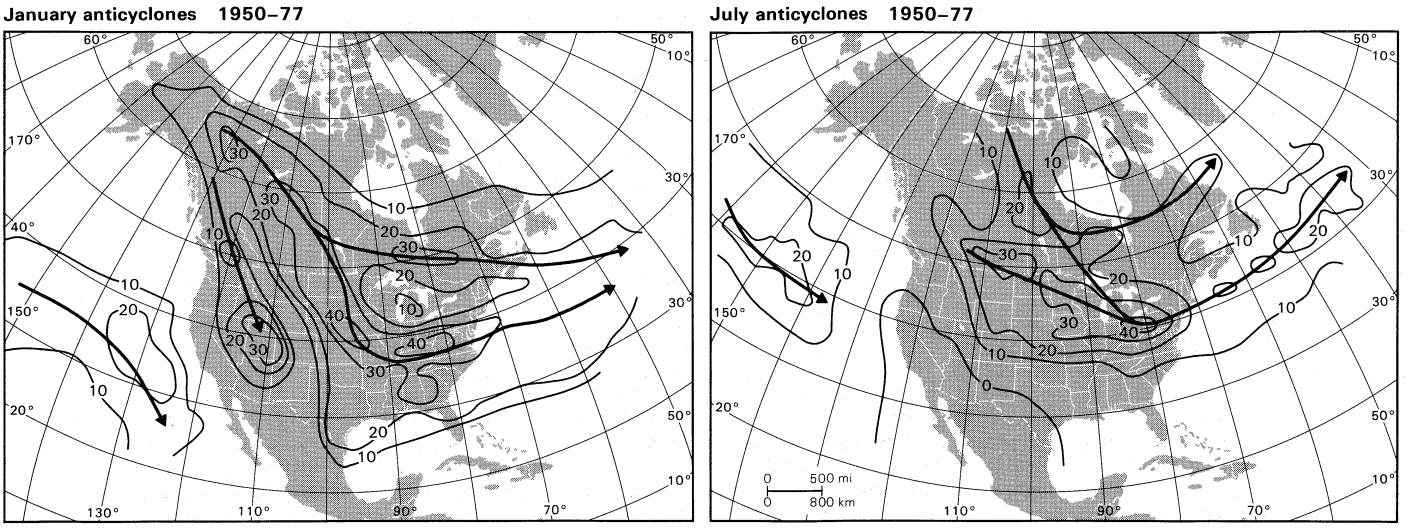



Figure 25: *January (left) and July (right) anticyclone tracks over North America and the surrounding ocean environs.*
The contours represent the total occurrences of anticyclones over a 28-year period (1950–77) in the area shown.
Adapted from Kennan M. Ziska and Phillip J. Smith, *Monthly Weather Review* (April 1980); American Meteorological Society

air that drains into the valley may give rise to a thick fog condition, which persists until daytime heating reverses the circulation and creates clouds ahead of the upslope motion at the mountain top.

Foehn winds

Another subset, called foehn winds (also known as chinook winds east of the Rocky Mountains or a Santa Ana in southern California), is induced by adiabatic temperature changes as air flows over a mountain. Adiabatic temperature changes are those that occur without the addition or subtraction of heat; they occur in the atmosphere when bundles of air are moved vertically. When air is lifted, it enters a region of lower pressure and expands. This expansion is accompanied by a reduction of temperature (adiabatic cooling). When air subsides, it contracts and experiences adiabatic warming. As air ascends on the windward side of the mountain, its cooling rate may be moderated by the heat that is released during the formation of precipitation. However, having lost much of its moisture, the descending air on the leeward side warms up adiabatically faster than it is cooled on the windward ascent. Thus, the effect of this wind, if it reaches the surface, is to produce warm, dry conditions. Usually such winds are gentle and produce a slow warming. On occasion, however, they may exceed 185 kilometres per hour and/or produce temperature increases of tens of degrees within only a few hours.

Still another local wind condition induced by terrain features is the katabatic wind. In this case the geography is characterized by a cold plateau adjacent to a relatively warm region of lower elevation. Such conditions are satisfied in areas in which major ice sheets or cold, elevated land surfaces border warmer, large bodies of water. Air over the cold plateau cools and forms a large dome of cold, dense air. Unless held back by background wind conditions, this cold air will spill over into the lower elevations with speeds that vary from gentle (a few kilometres per hour) to intense (93 to 185 kilometres per hour), depending on the incline of the slope of the terrain and the distribution of the background pressure field. Two special varieties of katabatic wind are well known in Europe. One is the bora, which blows from the highlands of Yugoslavia to the Adriatic Sea, while the other is the mistral, which blows out of central and southern France to the Mediterranean Sea.

## ZONAL SURFACE WINDS

Figures 16 and 17 reveal that on the average certain geographic locations can expect to experience winds that emanate from one prevailing direction largely dictated by the presence of major semipermanent pressure systems. Such prevailing winds have long been known in marine environments because of their influence on the great sailing ships.

Tropical and subtropical regions are characterized by a general band of low pressure lying near the Equator bounded by centres of high pressure that may extend poleward into the middle latitudes. Between these low- and high-pressure regions is the region of the tropical winds. Of these, the most extensive are the trade winds. So named because of their favourable influence on trade ships traveling across the subtropical North Atlantic, these winds flow westward and somewhat in the direction of the Equator on the equatorward side of the subtropical high-pressure centres. The "root of the trades" at the eastern side of the high centre is characterized by subsiding air. This produces the very warm, dry conditions with abundant sunshine found in the eastern extremes of the subtropical Atlantic and Pacific ocean basins. As the trade winds progress westward, however, subsidence abates, the air mass becomes more humid, and scattered showers appear, particularly on islands with elevated terrain features that interrupt the flow of the warm, moist air. The equatorward flow of the trade winds of the Northern and Southern hemispheres often results in a convergence of

Trade winds

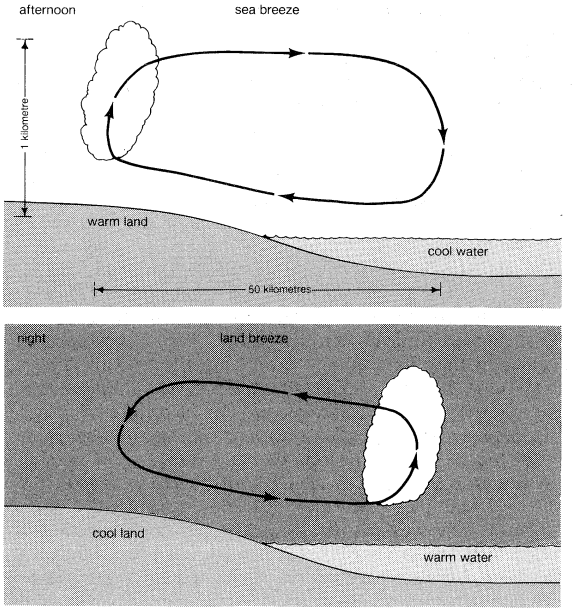



Figure 26: Typical sea-breeze (afternoon) and land-breeze (night) circulations with associated cloud formations.

the two air streams in a region known as the intertropical convergence zone (ITCZ; see Figure 34). Deep convective clouds, showers, and thunderstorms occur along the ITCZ.

When the air reaches the western extreme of the high, it turns poleward and then eventually returns eastward in the middle latitudes. The poleward-moving air is now warm, moist maritime tropical air, and it gives rise to the warm, humid, showery climate characteristic of the Caribbean region, eastern South America, and the western Pacific island chains. The westerlies are associated with the changeable weather common to the middle latitudes. Migrating extratropical cyclones and anticyclones associated with contrasting warm, moist air moving poleward from the tropics and cold, dry air moving equatorward from polar latitudes yield periods of rain, sometimes with violent thunderstorms, snow, sleet, or freezing rain interrupted by periods of dry, sunny, and sometimes bitterly cold conditions. Furthermore, these patterns are seasonally dependent, with more intense cyclones and colder air prevailing in winter but with a higher incidence of thunderstorms common in spring and summer. In addition, these migrations and the associated climate are complicated by the presence of landmasses and major mountain features, particularly in the Northern Hemisphere.

The mid-latitude westerlies

The westerlies lie on the equatorward side of the semipermanent, subpolar centres of low pressure. Poleward of these centres the surface winds turn westward again over significant portions of the subpolar latitudes. As in the middle latitudes, the presence of major landmasses, notably in the Northern Hemisphere, results in significant variations in these polar easterlies. In addition, the wind systems and the associated climate are seasonally dependent. During the short summer season, the wind systems of the polar latitudes are greatly weakened. During the long winter months, these systems strengthen and alternate periods of snow with long intervals of dry, cold air characteristic of continental polar or continental arctic air masses.

These major regions of surface circulation and their associated pressure fields are related to mean meridional circulation patterns as well (see Figure 33). Although their presence is discernible in long-term mean statistics accumulated over a hemisphere, such cells are often difficult to detect on a daily basis at any given longitude.

(P.J.S.)

## MONSOONS

A close examination of Figures 16 and 17 reveals particularly strong seasonal pressure variations over continents. Such seasonal fluctuations, commonly called monsoons, are more pronounced over land surfaces because these surfaces are subject to more significant seasonal temperature variations.

Monsoons blow for approximately six months from the northeast and six months from the southwest, principally in southern Asia and parts of Africa. Summer monsoons have a dominant westerly component and a strong tendency to converge, rise, and produce rain. Winter monsoons have a dominant easterly component and a strong tendency to diverge, subside, and cause drought. Both are the result of differences in annual temperature trends over land and sea.

**The Indian monsoon.** At the Equator, the area near India is unique in that dominant or frequent westerly winds occur at the surface almost constantly throughout the year; the surface easterlies reach only to 20° N in February, and even then they have a very strong northerly component. They soon retreat northward, and drastic changes take place in the upper-air circulation. This is a time of transition between the end of one monsoon and the beginning of the next. Late in March the high-sun season reaches the Equator and moves farther north. With it go atmospheric instability, convectional (rising, turbulent) clouds, and rain. The westerly subtropical jet stream (see below) still controls the flow of air across northern India and the surface winds are northeasterlies. As the high-sun season moves northward during April, India becomes particularly prone to rapid heating because the highlands to the north protect it from any incursions of cold air. In May the southwesterly monsoon is well established over Sri Lanka. There are three distinct regions of relative upper tropospheric warmth—namely (1) above the southern Bay of Bengal, (2) above the highlands of Tibet, and (3) across the still, dry trunks of the peninsulas. The relatively warm area above the southern Bay of Bengal occurs mostly at the 500–100 millibar level. It does not appear at a lower level and is probably caused by the release of condensation heat (associated with the change from water vapour to liquid water) at the top of towering cumulonimbus clouds along the advancing intertropical convergence.

Atmospheric instability and rapid heating

In May the dry surface of Tibet (above 4,000 metres) absorbs and radiates heat that is readily transmitted to the air immediately above. At about 6,000 metres an anticyclonic cell arises, causing a strong easterly flow in the upper troposphere above northern India. The subtropical jet stream suddenly changes its course to the north of the anticyclonic ridge and the highlands, though it may occasionally reappear southward of them for very brief periods. This change of the upper tropospheric circulation above northern India from westerly jet to easterly flow coincides with a reversal of the vertical temperature and pressure gradients between 600 and 300 millibars. On many occasions the easterly aloft assumes jet force. It anticipates by a few days the "burst," or onset, of the surface southwesterly monsoon some 1,500 kilometres farther south, with a definite sequential relationship, although the exact cause is not known. Because of India's inverted triangular shape, the land is heated progressively as the Sun moves northward. This accelerated spread of heating, combined with the general direction of heat being transported by winds, results in a greater initial monsoonal activity over the Arabian Sea (at late spring time), where a real frontal situation often occurs, than over the Bay of Bengal. The relative humidity of coastal districts in the Indian region rises above 70 percent and some rain occurs. Above the heated land the air below 1,500 metres becomes unstable, but it is held down by the overriding easterly flow. This does not prevent frequent thunderstorms in late May.

During June the easterly jet becomes firmly established at 150 to 100 millibars. It reaches its greatest speed at its normal position to the south of the anticyclonic ridge, at about 15° N from China through India.

In Arabia, it decelerates and descends to the middle troposphere (3,000 metres). A stratospheric belt of very cold air, analogous to the one normally found above the intertropical convergence near the Equator, occurs above the anticyclonic ridge, across southern Asia at 30°–40° N and above the 6,000-metre (500-millibar) level. These upper air features that arise so far away from the Equator are associated with the surface monsoon and are absent when there is no monsoonal flow. The position of the easterly jet controls the location of monsoonal rains, which occur ahead and to the left of the strongest winds and behind them to the right. The surface flow, however, is a strong, southwesterly, humid, and unstable wind that brings humidities of more than 80 percent and heavy, squally showers that are the "burst" of the monsoon. The overall pattern of the advance follows a frontal alignment, but local episodes may differ considerably. The amount of rain is variable from year to year and place to place. Most spectacular clouds and rain occur against the Western Ghāts, where the early monsoonal airstream piles up against the steep slopes, then recedes, and piles up again to a greater height. Each time it pushes thicker clouds upward until wind and clouds roll over the barrier and, after a few brief spells of absorption by the dry inland air, cascade toward the interior. The windward slopes receive from 2,000 to 5,000 millimetres of rain in the monsoon season. Various factors, and especially topography, combine to make up a complex regional pattern. Oceanic air flowing toward India below 6,000 metres is deflected in accordance with the Coriolis effect. The converging, moist oncoming stream becomes unstable over the hot land and is subject to convectional turmoil. Towering cumulonimbus clouds rise thousands of metres, producing violent thunderstorms and releasing latent heat in the surrounding air. As a result, the upper tropospheric warm belt migrates northwestward from the ocean to the land. The

Precipitation patterns

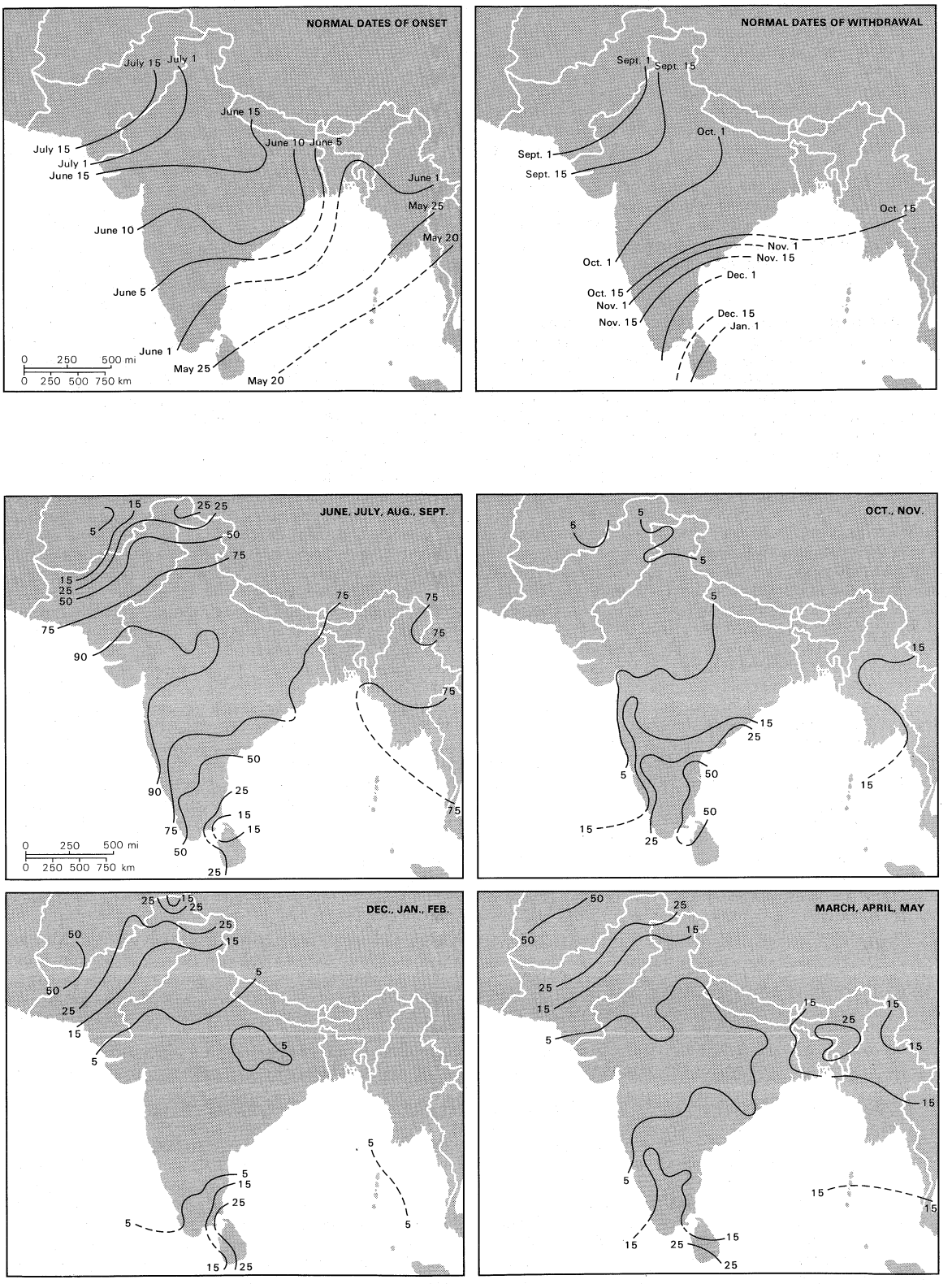



Figure 27: *The monsoonal system in the Indian subcontinent.*
(Top) Normal dates of onset and withdrawal of southwest monsoon. (Bottom) Seasonal distribution of rainfall as percentage of the annual.

From R. Ananthakrishnan and P. Rajagopalachari, India Meteorological Department; with permission from the Director General of Observatories, New Delhi, India

main body of air above 9,000 metres maintains a strong easterly flow.

Later, in June and July, the monsoon is strong and well-established to a height of 6,000 metres (less in the far north), with occasional thickening to 9,000 metres (Figure 27, top). Weather conditions are cloudy, warm, and moist all over India. Rainfall varies between 400 and 500 millimetres, but topography introduces some extraordinary differences. On the southern slopes of the Khāsi Hills at only 1,300 metres, where the moist airstreams are lifted and overturned, Cherrapunji has an average rainfall of 2,730 millimetres in July, with record totals of 897 millimetres in 24 hours in July 1915, more than 9,000 millimetres in July 1861, and 16,305 millimetres in the monsoon season of 1899. Over the Ganges Valley the monsoon, deflected by the Himalayan barrier, becomes a

southeasterly air flow. By then the upper tropospheric belt of warmth from condensation has moved above northern India, with an oblique bias. The lowest pressures prevail at the surface.

It is mainly in July and August that waves of low pressure appear in the body of monsoonal air. Fully developed depressions appear once or twice a month. They travel from east to west more or less concurrently with high-level easterly waves and bursts of speed in the easterly jet, causing local strengthening of the low-level monsoonal flow. The rainfall consequently increases and is much more evenly distributed than it was in June (Figure 27, bottom). Some of the deeper depressions become tropical cyclones before they reach the land, and these bring torrential rains and disastrous floods.

Associated dry spells

A totally different development arises when the easterly jet moves farther north than usual because the monsoonal wind rising over the southern slopes of the Himalayas brings heavy rains and local floods. The weather over the central and southern districts, however, becomes suddenly drier and remains so for as long as the abnormal shift lasts. The opposite shift is also possible, with mid-latitude upper air flowing along the south face of the Himalayas and bringing drought to the northern districts. Such dry spells are known as "breaks" of the monsoon. Those affecting the south are similar to those experienced on the Guinea coast during extreme northward shifts of the wind belts (as later discussed), whereas those affecting the north are due to an interaction of the middle and low latitudes. The southwest monsoon over the lower Indus Plain is only 500 metres thick and does not hold enough moisture to bring rain. On the other hand, the upper tropospheric easterlies become stronger and constitute a true easterly jet stream. Western Pakistan, Iran, and Arabia remain dry (probably because of divergence in this jet) and thus become the new source of surface heat.

By August the intensity and duration of sunshine have decreased, temperatures begin to fall, and the surge of southwesterly air diminishes spasmodically almost to a standstill in the northwest. Cherrapunji still receives over 2,000 millimetres of rainfall at this time, however. In September dry, cool, northerly air begins to circle the west side of the highlands and spread over northwestern India. The easterly jet weakens and the upper tropospheric easterlies move much farther south. Because the moist southwesterlies at lower levels are much weaker and variable, they are soon pushed back. The rainfall becomes extremely variable over most of the region, but showers are still frequent in the southeastern areas and over the Bay of Bengal.

By early October variable winds are very frequent everywhere. At the end of the month the entire Indian region is covered by northerly air and the winter monsoon takes shape. The surface flow is deflected by the Coriolis force and becomes a northeasterly flow. This causes an October–December rainy season for the extreme southeast of the Deccan (including the Madras coast) and eastern Sri Lanka, which cannot be explained by topography alone because it extends well out over the sea. Tropical depressions and cyclones are important contributing factors.

Most of India thus begins a sunny, dry, and dusty season. The driest period comes in November in the Punjab; December in Central India, Bengal, and Assam; January in the northern Deccan; and February in the southern Deccan. Conversely, the western slopes of the Karakoram and Himalayas are then reached by the mid-latitude frontal depressions that come from the Atlantic and the Mediterranean. The winter rains they receive, moderate as they are, place them clearly outside the monsoonal realm.

Because crops and water supplies depend entirely on monsoonal rains, it became imperative that quantitative, long-range weather forecasts be available. For a forecast to be released at the beginning of June, it is necessary to use, in April, South American pressure data and Indian upper-wind conditions (positive correlation) and, in May, rainfall in Zimbabwe and Java and easterly winds above Calcutta (negative correlation).

**The Malaysian-Australian monsoon.** Southeast Asia and northern Australia are combined in one monsoonal system that differs from others because of the peculiar and somewhat symmetrical distribution of landmasses on both sides of the Equator. In this respect, the northwest monsoon of Australia is unique. The substantial masses of water between Asia and Australia have a moderating effect on tropospheric temperatures, weakening the summer monsoon. The many islands (*e.g.,* Philippines and Indonesia) provide an infinite variety of topographic effects. Typhoons that develop within the monsoonal air bring additional complications.

Effects of water areas and topography

It would be possible to exclude North China, Korea, and Japan from the monsoonal domain because their seasonal rhythm follows the normal mid-latitude pattern—a predominant outflow of cold continental air in winter, and frontal depressions and rain alternating with fine, dry anticyclonic weather in the warm season. On the other hand, the seasonal reversal of wind direction in this area is almost as persistent as that in India. The winter winds are much stronger because of the relative proximity of the Siberian anticyclone. The tropical ridge of high pressure is the natural boundary between these nonmonsoonal areas and the monsoonal lands farther south.

The northern limit of the typical monsoon may be set at about latitude 25° N. Farther north, the summer monsoon is not strong enough to overcome the effect of the traveling anticyclones normally typical of the subtropics. As a result, monsoonal rains occur in June and also in late August and September, separated by a mild anticyclonic drought in July. In South China and the Philippines the trade winds prevail in the October–April (winter) period, strengthened by the regional, often gusty, outflow of air from the stationary Siberian anticyclone. Their disappearance and replacement by opposite (southwesterly) winds in the May–September (summer) period is the essence of the monsoon. In any case, these monsoonal streams are quite shallow, about 1,500 metres in winter and 2,000 metres in summer. They bring rain only when subject to considerable cooling, as happens to the windward anywhere on the steep slopes of the Philippines and Taiwan. On the larger islands there are contrasting effects, the slopes facing west receiving most of the rain from May to October and a drought from December to April, whereas the slopes facing east receive orographic rains (those produced when moist air is forced to rise by topography) from September to April and mainly convectional rains from May to October.

Vietnam, Thailand, and Indonesia

In Vietnam and Thailand the summer monsoon is more strongly developed because of the wider expanses of overheated land. The southwesterly stream flows from May to October, reaching a thickness of four to five kilometres; it brings plentiful but not extraordinary rainfall. November to February is the cool, dry season, and March to April the hot, dry one; in the far south the coolness is but relative. Along the east coast and on the eastward slopes more rain is brought by the winter monsoon. In the summer, somewhere between Thailand and Kampuchea in the interior, there may be a faint line of convergence between the southwesterly Indian-Burmese monsoon and the southeasterly Malaysian monsoon.

Monsoonal winds are weak over Indonesia because of the expanses of water and the low latitude, but their seasonal reversal is definite. From April to October the Australian southeasterly air flows, whereas north of the Equator it becomes a southwesterly. It generally maintains its dryness over the islands closer to Australia, but farther north it carries increasing amounts of moisture. The northeasterly flow from Asia, which becomes northwesterly south of the Equator, is laden with moisture when it reaches Indonesia, bringing cloudy and rainy weather between November and May. The wettest months are December in most of Sumatra and January elsewhere, but rainfall patterns are highly localized. In Java, for instance, at sea level alone there are two major regions, an "equatorial" west with no dry season and a "monsoonal" east with extreme drought in August and September.

Australian monsoon

Because of its relatively small size and compact shape, Australia shows relatively simple monsoonal patterns. The north shore is subject to a clear-cut wind reversal between summer (November–April, northwesterly) and win-

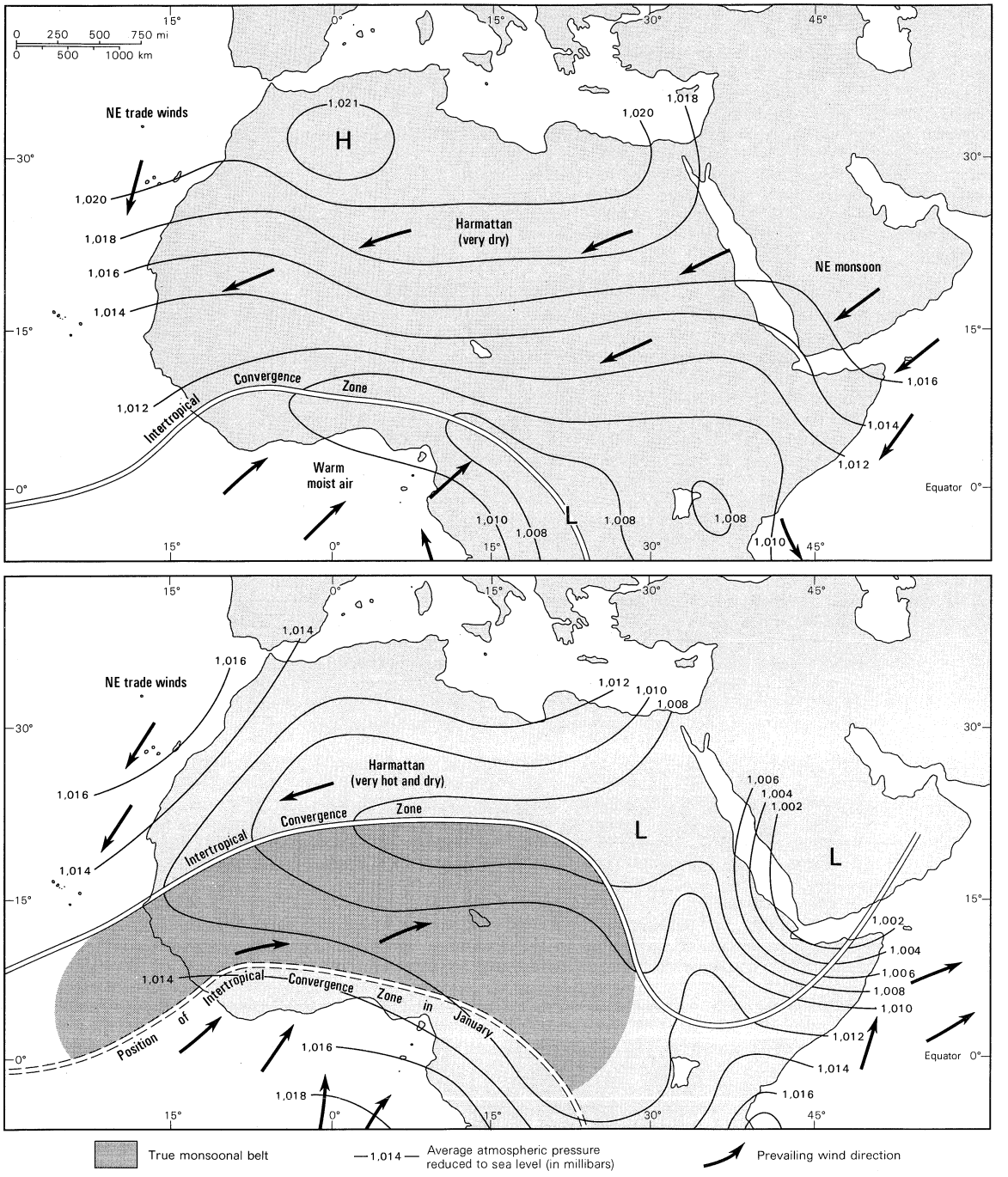



Figure 28: African monsoonal regimes in (top) January and (bottom) July.
From L. Welter and P. Suraud, *Monographie No. 8 de la Meteorologie Nationale* (November 1957)

ter (May–September, southeasterly), but with two definite limitations: first, the northwesterly, rain-bearing monsoonal wind is often held offshore and is most likely to override the land to any depth during January and February; and second, even in summer there often are prolonged spells of southeasterly trade winds issuing from traveling anticyclones, separating the brief monsoonal incursions. The Australian summer monsoon is thus typical in direction and weather type but quite imperfect in frequency and persistence. Its thickness is usually less than 1,500 metres over the sea and 2,000–2,500 metres over the land.

Much less typical are the marginal monsoonal manifestations. On the northwest coast there frequently is a northwesterly air flow in the summer (December–March), as opposed to the winter southeasterlies, but this stream is very shallow and does not bring any rain—*i.e.*, its weather is not monsoonal even though its direction is so. On the northeast coast the onshore air is humid and brings rain, but its direction is only partly modified in summer. It comes in mostly as a northeasterly, while at other times it is mostly southeasterly.

**The West African monsoon.** The main characteristics of the West African seasons were known over two centuries ago. The southwest winter monsoon flows as a shallow (less than 2,000 metres) humid layer of surface air, overlain by the primary northeast trade wind, which blows from the Sahara and the Sahel as a deep stream of dry, often dusty air. As a surface northeasterly, it is generally known as the harmattan, gusty and dry in the extreme—cool at night and scorchingly hot by day. As in a thorough monsoonal development, upper tropospheric anticyclones occur at about 20° N, while the easterly jet stream may occur at about 10° N, much closer to the Equator than they are in the Indian region.

Alternation with the harmattan

The West African monsoon is the alternation of the southwesterly wind and the harmattan at the surface. As seen in Figure 28, such alternation is normally found between latitudes 9° and 20° N. There are northeasterlies

farther north all the time, but only southwesterlies farther south. The whole year is more or less dry at 20° N except for erratic rains in the high-sun season. The drought becomes shorter and less complete farther south. At 12° N it lasts about half the year, and at 8° N it disappears completely. Farther south a different, lighter drought begins to appear in the high-sun months when the monsoonal southwesterly is strongest. It is due to the arrival of dry surface air issuing from anticyclones formed beyond the Equator in the Southern Hemisphere and is thus similar to the monsoonal drought in Java. Like the "break" of the monsoon in southern India, however, it occurs beyond the Equator.

The moist southwesterly stream, particularly frequent between 5° and 10° N, can reach much farther north, bringing warm, humid nights and moderately hot but still humid days. The harmattan brings cooler nights, but the extreme daily heating causes a thermal range of 10° to 12° C. Even in the daytime, the harmattan may give a sensation of coolness to the human skin as it evaporates moisture from it. The alternation of the two winds is seasonal on the basis of overall frequency, but in fact it varies considerably with the synoptic pressure patterns. The harmattan comes in spells that mostly last from a few days to more than a week.

The advancing fringe of the southwest monsoon is too shallow (under 1,000 metres) for many thunderstorms and disturbances. They usually occur 200–300 kilometres in arrears, where the moist air is deeper (1,000–2,000 metres) but the ground is still hot enough to make it very unstable. The tops of the cumulonimbus may reach 12,000 metres, well above freezing level (4,200–4,500 metres). The disturbances are usually along a given longitude line that is slightly curved and may in fact form one long line squall. They also reach 12,000 metres or more, traveling steadily westward at 37 to 56 kilometres per hour. This suggests that they originate in the primary trade wind aloft and as in India are probably related to the tropical easterly jet stream. The southwest monsoon dominates the weather, and clouds and rain abound. The rain is due primarily to coalescence of droplets, with most of the clouds located below 3,500 metres. The humidity is very high, and the daily range of temperature remains around 4° C.

Seasonal changes

If it were not for the change in wind direction when the southeast trades have crossed the Equator, the monsoon system of West Africa could not be distinguished from the weather system, caused by the seasonal shift in the latitude of the intertropical convergence, as experienced over most of Central Africa. There is a rainy season (in this case, the monsoon season), which on the west coast lasts two to three months at latitude 16° N, three to four months at 14° N, six to seven months at 10° N, and eight months at 8° N.

On the south coast, which is at latitude 4° N to 6° N, the southwest monsoon (as the intertropical convergence) may occur at any time, but the results are quite atypical for various reasons. In the low-sun season (December–February) the southwesterly is rare and ineffective, and the weather is cloudy but dry. From April to June the midday Sun is at its highest, and insolation (radiation received at the Earth's surface) is most intense. Because the southwest wind occurs most frequently, the consequent building up of clouds leads to the main rainy season. During July and August (the short drought) cloudy conditions prevail, but the air issues direct from anticyclones farther south and is dry in spite of the fact that its direction of flow does not change. Although cloudiness decreases after the second high-sun season in September and October, there is a period of occasional rains just sufficient to constitute a secondary maximum.

Toward the north, conditions are more distinctly monsoonal: by latitude 8° N the two wet seasons have merged into one long "wet" with two subdued peaks, which last approximately seven to eight months (March–October). The "dry," which is controlled by northeast winds, lasts from November to early March. There is one rainfall maximum (in August or September) only a short distance farther north, although the wet season is only a few weeks shorter. These changes are shown in Figure 29.

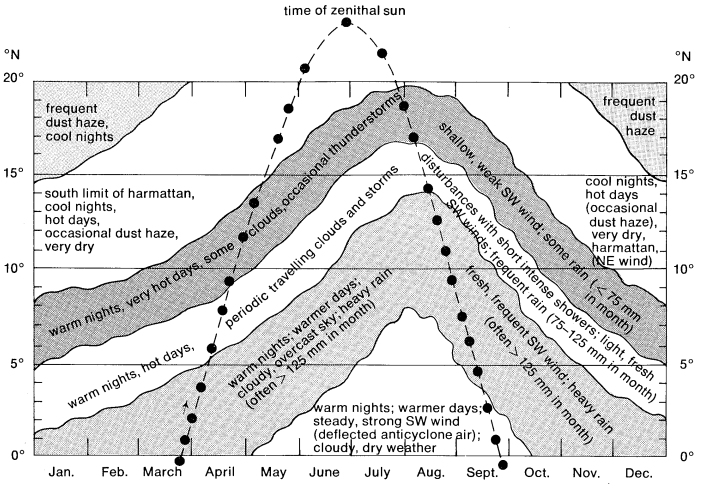



Figure 29: Distribution of climatic characteristics associated with monsoons in West Africa.

**Monsoonal tendencies in Europe and North America.** In central Europe, where the average wind direction in summer differs some 30° to 40° from that of the Atlantic, there are monsoonal tendencies that occur not as a continuous flow but rather intermittently within frontal depressions, bringing cool, cloudy weather, rain, and thunderstorms. Some see in this climatic pattern a true monsoon, but it is obvious that it is only an "embryo monsoon" that results in weather singularities. The latitude is too high for a true monsoon to arise.

In North America the relatively low latitude and the orientation of the land–sea boundary on the Gulf of Mexico are quite favourable to monsoonal developments. During the summer, low atmospheric pressure is frequent over the heated land; the northeasterly trade winds are consequently deflected to become easterly, southeasterly, or even southerly winds. Texas and the Gulf country generally may be completely overrun by a shallow sheet of oceanic air, which may continue for a long distance inland. The rainfall regime does not reveal any marked monsoonal pattern. There are mostly two, three, or even four minor peaks in the sequence of monthly rainfall totals. In the winter there often occur northers, which are offshore winds caused by the general anticyclonic flow of air from the cold land. Neither the summer onshore wind nor the winter offshore wind is persistent enough to constitute a monsoonal sequence, even though monsoonal tendencies are quite evident.

Central American cycle

In Central America a true monsoonal cycle occurs over a small area facing the Pacific Ocean between 5° and 12° N. Not only is there a complete seasonal reversal of the wind, but the rainfall regime is typically monsoonal. The winter period (from November to January and from March to April according to latitude and other factors) is very dry. The rainy season begins earlier (May) in the south and progressively later farther north, coming at the end of June in southern Mexico. It concludes at the end of September in the north and as late as early November in the south. The result is a rainy season that lasts three months in southern Mexico, increasing in duration with decreasing latitude, to six to seven months in Costa Rica. Latitude for latitude, this is a subdued replica of the monsoon of India. (J.G./P.J.S.)

## UPPER-AIR WAVES

**Characteristics.** Surface pressure and wind systems have already been shown to occur in response to surface heating and cooling, movements of warm and cold air masses, and the presence of various orographic features. Still to be discussed is the role played by pressure and wind systems located well above the surface. Known as upper-air circulations, their most prominent characteristic is the wavelike undulations exhibited in both the pressure and wind fields. These upper-air waves are often coupled with surface cyclones and anticyclones, dictating the severity of the associated weather, the intensity changes of the sur-

face-pressure systems, and the direction of propagation of both the weather and the systems.

Thermal-wind equation

A well-known property of upper-air waves is the presence of winds that are consistently stronger than those found at the surface. This is a consequence of the thermal-wind equation, a relationship between vertical changes in the vector wind and the horizontal temperature gradient. The thermal wind is simply the difference between the vector geostrophic wind at two different levels. The rate of this change per unit of vertical distance is given, for the $u$ and $v$ wind components, by the thermal-wind equation

$$\frac{\partial u}{\partial z} = -\frac{g}{fT}\frac{\partial T}{\partial y} \text{ and } \frac{\partial v}{\partial z} = \frac{g}{fT}\frac{\partial T}{\partial x}, \quad (2)$$

where $T$ is the air temperature; $g$ is the gravitational acceleration (9.8 metres per seconds squared); and $\partial T/\partial x$, $\partial T/\partial y$ are the components of the horizontal temperature gradients in the zonal and meridional directions, respectively. Using the Northern Hemisphere as an example, westerly wind speeds ($u > 0$) will increase with height if air is colder to the north (positive $y$ direction), and southerly wind speeds ($v > 0$) will increase with height if air is colder to the west (negative $x$ direction). Since these are precisely the circumstances often found in the atmosphere, increasing wind speeds with height are the rule. Furthermore, these increases are greatest in regions of strong temperature gradients. Therefore, it is common to find the strongest upper-air winds above low-level fronts.

Another property of upper-air circulations is their geographic displacement with height toward regions of warm or cold air. This so-called tilt of the waves with height is the consequence of an atmospheric property known as hydrostatic balance, and it is easiest to see by focusing attention on the lines of relatively low (trough) and high (ridge) pressure within the wave. This balance is an expression of the fact that a volume of air is subjected to a gravitational force acting toward the Earth's surface which is almost exactly balanced by a pressure force acting vertically in the opposite direction, represented mathematically by

$$g = -\frac{1}{\rho}\frac{\partial p}{\partial z}. \quad (3)$$

Remembering that $g$ is constant, if $\rho$ is larger, $-\partial p/\partial z$ also must be larger. Thus, equation (3) shows that in colder, denser air (larger $\rho$), pressure decreases more rapidly with height (larger $-\partial p/\partial z$). This leads to several wave-tilt scenarios. First, if at lower levels cold air (larger $-\partial p/\partial z$) is centred in the low-pressure region and warm air (smaller $-\partial p/\partial z$) is centred in the high-pressure region, the two regions will exhibit greater pressure differences at higher levels. Thus, the meteorologist says that "cold-core" lows and "warm-core" highs become more intense with increasing height. Such systems are often slow-moving circulations that influence local weather for several days. For example, warm-core highs, which are common in summer and fall, represent deep layers of warm, dry, stable air. Since they are very stable and have no precipitation, these stagnant air masses permit the accumulation of dust and air pollution, which can reach annoying or even hazardous levels. A second scenario is the reverse of the first. In this case, warm-core lows and cold-core highs decrease in intensity with increasing height. The tropical cyclone is an example of a warm-core low.

Cold-core lows and warm-core highs

Because the cold or warm air is centred on the high- or low-pressure region, neither of these first two cases exhibits any tilt with height. A third very important scenario, however, can exhibit marked tilt. This occurs when the cold or warm air is centred between the lower-level pressure regions. In this case, the most rapid decrease of pressure with height occurs in the cold air between the low and high. Thus, the position of the low (or trough) is displaced toward the cold air with increasing height. Similarly, the pressure decreases less rapidly with height in the warm air, and the high (or ridge) is displaced toward the warm air. In general, the cold air is located upstream from the surface low. This results in a typical tilted-wave configuration in which the upper-air troughs and ridges are located upstream from their corresponding low-level features. Since in middle latitudes (where tilted systems are most prominent) the flow is generally west to east, upstream usually means to the west. Such patterns, represented schematically in Figure 30, are typical of the weather-producing circulation systems of the middle and high latitudes.

**Propagation and development.** The propagation and development of upper-air waves are largely determined by the horizontal transport (referred to as advection) of cold or warm air and a fluid property known as vorticity. The vorticity of interest in the context of this discussion can be defined as the rotational motion experienced by air parcels moving in the horizontal plane. Mathematically, the vorticity of horizontal motions relative to the Earth (relative vorticity) is given by

Advection and vorticity

$$\delta = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}. \quad (4)$$

If the rotation imposed by the rotating Earth is added, a quantity known as the absolute vorticity results:

$$\delta_a = \delta + f \quad (5)$$

where $f$ is the Coriolis parameter (see equation [1]). The absolute vorticity is largest in troughs and smallest in ridges. Thus, positive advections (transport of larger values) tend to occur downstream from upper-air troughs. The resulting increasing in vorticity yields an increase in cyclonic circulation (*i.e.*, the circulation associated with cyclones or troughs). As the air in this region attempts to maintain geostrophic balance in the presence of this increased cyclonic circulation, the pressure is forced to decrease. Downstream from upper-air ridges, advections of negative vorticity occur, giving rise to increased anticyclonic circulations and corresponding pressure increases. The coupled decrease of pressure downstream from the trough and increase of pressure downstream from the ridge in an upper-air wave are precisely the combination of pressure changes required to propagate the wave. Thus, upper-air wave propagation can be largely explained by vorticity advection. This propagation is usually in the direction of the general background flow (west to east in middle latitudes) and is generally faster for shorter waves. In some cases, very long waves (several thousand kilometres) are nearly stationary or actually move in a direction opposite to that of the background flow (referred to as a retrograding wave).

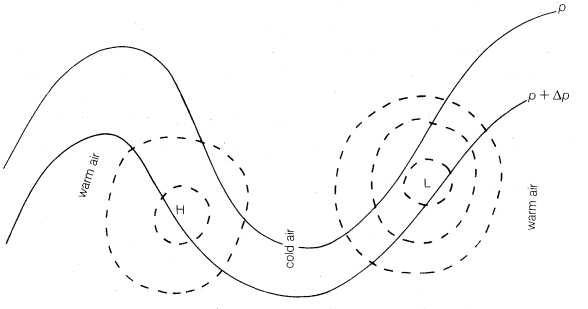



Figure 30: *Middle-latitude tilted-wave pattern.* The pressure field at a height of six kilometres is shown by solid isobars and at sea level by dashed isobars.

The advection of warm or cold air (known as thermal advection) plays quite a different role in upper-air wave behaviour. In order for such advections to occur, the temperature wave must be out of phase with the pressure wave, a so-called baroclinic atmosphere. If the two waves were in phase (barotropic atmosphere), the winds would blow along the isotherms, therefore yielding zero thermal advection. If the two waves are out of phase, it is typical to find the cold air upstream from the trough and the warm air upstream from the ridge, producing cold-air advection in the trough and warm-air advection in the ridge. The resulting colder temperatures in the trough and warmer temperatures in the ridge force a further reduction of

Role of thermal advection

pressure in the trough and increase in the ridge—namely, a wave amplification. Thus, wave development, or amplification, is largely explained by thermal advections and is more likely in strongly baroclinic atmospheres.

Alternately, upper-air wave development has been described as an example of fluid instability. In one theory, originally proposed by the Chinese-American meteorologist Hsiao-lan Kuo in 1949, the waves develop from small disturbances on an atmosphere made unstable by an appropriate horizontal wind shear (barotropic instability), and the energy of the growing wave is drawn from the kinetic energy of the background flow. In another theory, originally published by the American meteorologist Jule G. Charney in 1947, the waves develop in an atmosphere made unstable by an appropriate horizontal temperature gradient (baroclinic instability) and derive their energy from the potential energy of the disturbed atmosphere.

**Relationships to surface features.** Upper-air circulations play a key role in the development and propagation of surface pressure systems. As noted above, the surface pressure is simply the weight per unit area of the air above the surface. Thus, in order to change the surface pressure, it is necessary to change the mass content of the overriding air. Because large amounts of mass are transferred by upper-air flows, such flows are capable of forcing significant changes in surface pressure.

Generation of surface-pressure changes

The transfer of mass by upper-air circulations occurs in a very nonuniform fashion. As a result, mass tends to build up (mass convergence) in some regions and be depleted (mass divergence) in others. If these convergences or divergences occurred uniformly at all levels above a given surface location, it would be easy to anticipate the resulting surface-pressure changes. Such uniformity, however, is not observed. Rather, divergences in one portion of a column of air are largely compensated by convergences in another. Thus, changes in surface pressure may result from the subtle differences between these conflicting mass divergence/convergence patterns.

Examples of divergence/convergence cross sections for non-tilting and tilting wave systems are presented in Figure 31. Also included are the directions of the vertical air motions required to satisfy mass conservation principles. Such vertical motions are typically very small compared to horizontal motions yet are important because they are largely responsible for the formation or dissipation of clouds. Clouds form in regions of upward motion if sufficient moisture is present and dissipate in those of downward motion. In tilting wave systems, this accounts for the occurrence of inclement weather with low pressure and fair weather with high pressure.

In a non-tilting system, which is characteristic of an occluded cyclone (see above), the bulk of the upper-air divergence and convergence occurs between surface low (L) and high (H) centres, with little happening immediately above the centres. This signifies a propagating wave—a very slow one in some cases—with little intensification of the surface centre. In fact, decay (filling) of the surface centre might occur because of frictionally induced mass convergence in the low and divergence from the high, which is not compensated for by corresponding upper-air mass divergence and convergence. In the tilting system, much of the upper-air divergence and convergence lies above the surface centres. In this case, opportunities are much better for decreasing or increasing the mass above the cyclone or anticyclone, respectively, and thus for intensifying the surface centre. Such a wave also will propagate in the general direction of the mean tropospheric flow.

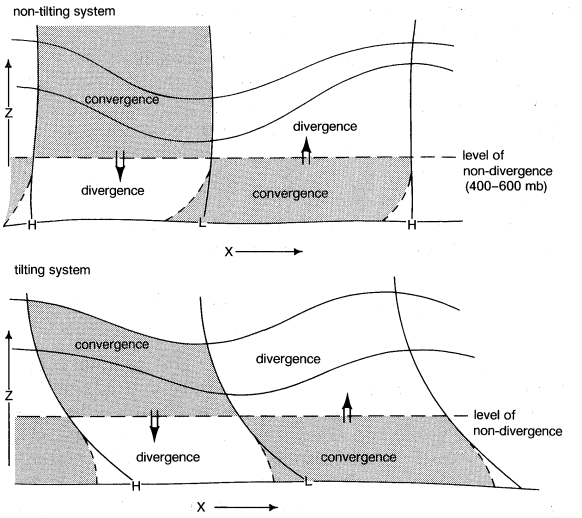



Figure 31: *Vertical cross sections through a wave system depicting typical divergence/convergence distributions for non-tilting and tilting systems.*
Double arrows represent direction of corresponding vertical air motions. X is the horizontal coordinate with arrow pointed in the direction of wave propagation, and Z is the vertical coordinate with arrow pointed upward. L and H are the positions of surface low (cyclone) and high (anticyclone) centres, respectively. The solid lines extending upward from L and H are trough and ridge axes, respectively. The small areas of convergence or divergence near the surface and upstream from the centre positions (dashed lines) are caused by frictional inflow into the low and frictional outflow from the high (see text).

## JET STREAMS

Among the more fascinating features of upper-air circulations are discontinuous bands of relatively strong winds (usually in excess of 30 metres per second) called jet streams. As with other wind fields that increase with increasing height, jet streams can be explained as an application of the thermal-wind equation. They are located above areas of particularly strong temperature gradients—*e.g.*, frontal zones. In such areas, the pressure gradients and the resulting wind speeds increase with increasing height so long as the temperature gradients persist in the same direction. In general, this will extend to the tropopause, after which the temperature gradient reverses direction and the wind speeds diminish. Thus, jet streams are usually found in the upper troposphere (*i.e.*, at levels of nine to 18 kilometres).

Because regions of strong temperature gradients can be created in different ways, there are several classes of jet streams. Perhaps the most familiar is the polar-front jet stream. As noted earlier, the polar front is the boundary between polar and mid-latitude air. In winter this boundary may extend equatorward to 30°, while in summer it retreats to 50°–60°. Winter fronts also are distinguished by stronger temperature contrasts than summer fronts. Thus, jet streams are located more equatorward in winter and are more intense during that time. Such streams are generally strongest east of major continents, sometimes exceeding 75 metres per second.

Polar-front jet stream

The polar-front jet is important for several reasons. First, because it is a region of maximum upper-tropospheric flow, it is also one of concentrated upper-air divergence and convergence. Divergence is favoured in the downstream, poleward sector of the jet core as well as in its upstream, equatorward sector (Figure 32). This means that such regions are favoured, though not exclusive, regions for extratropical cyclone development and inclement weather. Second, the polar-front jets, which move west to east but meander with the general upper-air waves, often "steer" the movement of major low-level air masses. This steering is simply a reflection of the very strong mass transport associated with jet streams. Identification of changes in jet-stream flow can often assist in predicting major changes in air mass—hence in temperature and weather—over a

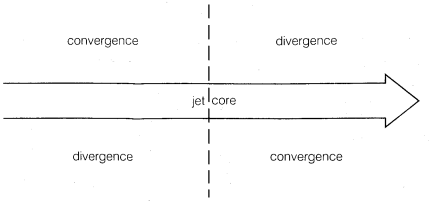



Figure 32: Typical horizontal divergence/convergence patterns associated with a jet stream.

region. Third, jet streams are an important factor in high-altitude flight. Military and civilian jet aircraft depend heavily on reliable information about upper-air winds, which is used for planning the duration of flights and corresponding fuel consumption. Furthermore, jet streams generate a great deal of turbulence because of their strong wind shears. Airlines seek to avoid such "clear-air turbulence" in order to maximize passenger safety and comfort.

Subtropical jet stream

A second jet stream is located at the poleward limit of the equatorial tropical air above the transition zone between tropical and mid-latitude air. This subtropical jet stream is usually found at latitudes 30°to 40°in general westerly flow. This jet may not be marked by pronounced surface temperature contrasts but rather by relatively strong temperature gradients in the mid-troposphere. Moreover, when the polar-front jet penetrates to subtropical latitudes, it may merge with the subtropical jet to form a single band. The subtropical jet, which is generally weaker and located at higher elevations than its higher-latitude cousin, is less commonly associated with migrating cyclones; instead, it may be accompanied by sporadic periods of strong convection, along with heavy rain showers. Also peculiar to the tropical latitudes of the Northern Hemisphere is a high-level jet called the tropical easterly jet stream. Such jets are located about 15° N over continental regions due to the latitudinal heating contrasts over tropical landmasses that are not found over the tropical oceans.

The general location of all three jet streams in relation to other mean meridional circulation features is shown in Figure 33. Jet streams occur in both hemispheres. Those in the Southern Hemisphere resemble the northern hemispheric systems, though they exhibit less day-to-day variability due to the presence of smaller landmasses.

(P.J.S.)

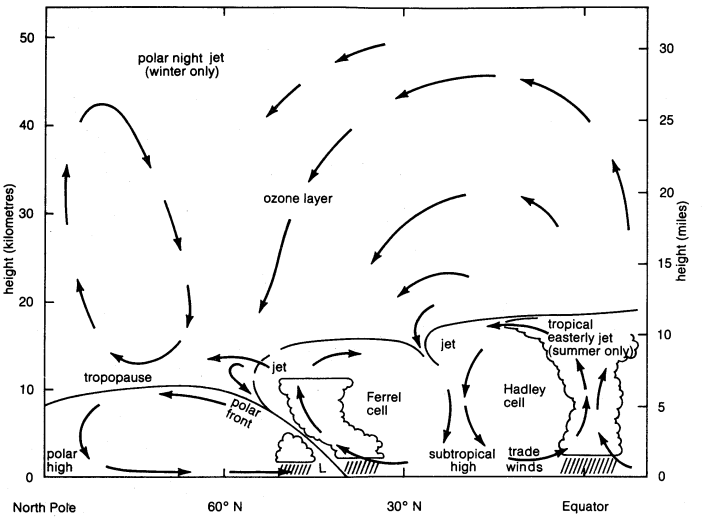



Figure 33: Positions of jet streams in the atmosphere. Arrows indicate directions of mean motions in a meridional plane.

## STRATOSPHERIC AND MESOSPHERIC WIND SYSTEMS

The zonal currents and disturbances so far described have their roots and energy sources in the troposphere, the lowest region of the Earth's atmosphere. A second group of wind systems occurs within the overlying stratosphere and mesosphere and extends upward into the ionosphere, where it interacts with the very different systems of the upper atmosphere. The stratosphere and mesosphere are dry, and even the strongest disturbances of the higher level cannot show themselves by means of the clouds so abundant in tropospheric storms. Only at about 25 kilometres does one occasionally see (in winter) small patches of mother-of-pearl cloud. The so-called noctilucent clouds, also rare, occur much higher up—at about 80 kilometres.

Early evidence of stratospheric wind patterns

Before the organization of a worldwide radiosonde network that systematically extended its range above 20 kilometres, it was possible only to infer from acoustic evidence that a unique and different wind regime existed at these levels. It was deduced during the 1930s from the observed travel of the noise of explosions that the winds of the stratosphere apparently reversed annually, in a sort of stratospheric monsoon (though in every other sense this was a misnomer). Below 20 kilometres the basal stratosphere was dominated by the upper parts of the Ferrel westerlies. It was generally believed that the stratosphere above this level was a layer of quiet, undisturbed flow because of the increase of temperature with height.

It is now known—after several decades of additional observation to 35 kilometres and a rather shorter period of rocket-sonde observation to above 60 kilometres—that this impression was false, though the earlier inference of an annual reversal in mid-latitudes has been vindicated.

**Polar-night westerlies.** Figure 34 sketches the observed circulation. In middle and high latitudes there is indeed a reversing vortex in both hemispheres. In the winter hemisphere the mesospheric westerlies (so called because their

By courtesy of World Meteorological Organization

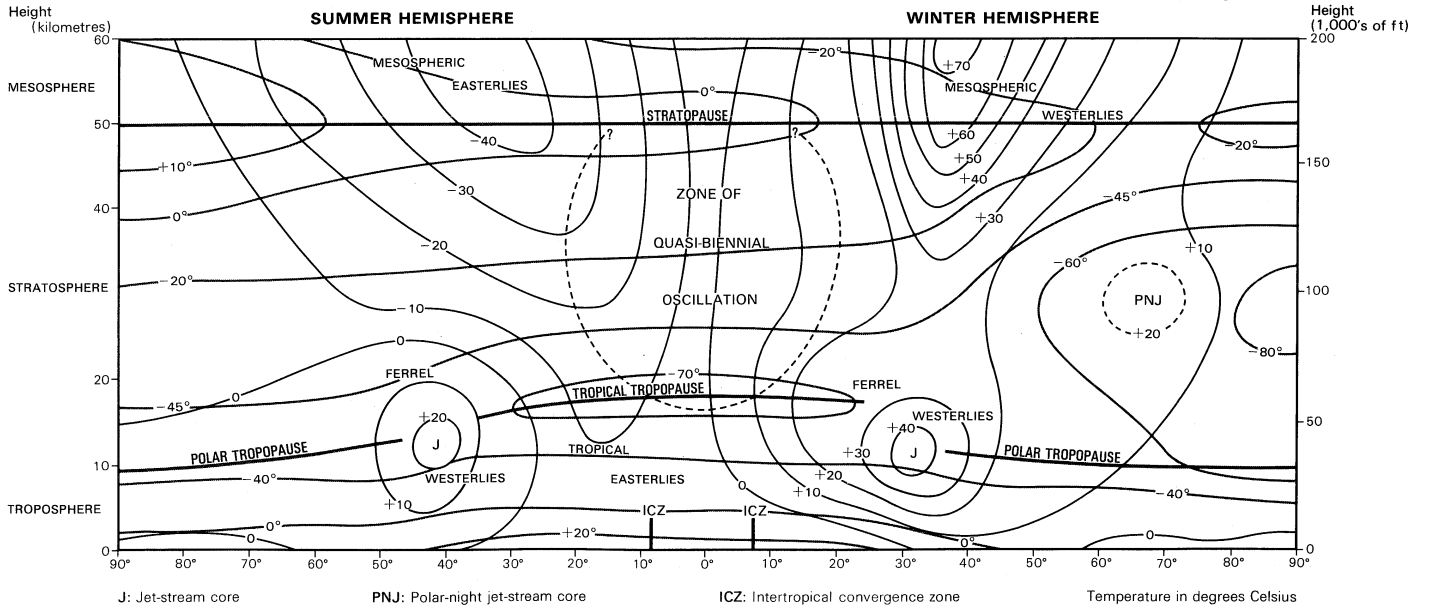



Figure 34: Meridional cross section of the atmosphere to a height of 60 kilometres in summer and winter hemispheres, showing seasonal changes. Numerical values for wind are in units of metres per second and are typical of the Northern Hemisphere, but the structure is much the same in the Southern Hemisphere. Positive and negative signs indicate winds of opposite directional sense (see text).

jet cores lie in the mesosphere) are believed strongest at about 65 kilometres in mid-latitudes, where winds average more than 80 metres per second. In Figure 34 a possible subsidiary-jet maximum near the edge of the polar night is indicated (the polar-night jet). The polar-night area is bitterly cold (down to below −80° C in the 25–35-kilometre layer), and strong westerlies are usually present in the intensely baroclinic zone around this cold core. These westerlies, which are often visible on charts as low as 20 kilometres, have become famous since 1952, when the first explosive warming in them was observed over Berlin. Their relation to the main mesospheric jet is not clear, and they may be distinct.

Deformation by standing waves

The polar-night westerlies of the Northern Hemisphere are deformed by standing waves (mostly of the two-wave sort), with a ridge over and west of Alaska and troughs over eastern Siberia and eastern North America, which are often continuous with similar waves in the Ferrel westerlies below. The southern polar-night system lies over the ocean for much of the winter and is less well known, but it is probably more nearly a pure zonal westerly. Both systems are subject to transient disturbances involving strong vertical motion, as is the whole depth of the mesospheric westerlies. These disturbances show themselves by sudden warmings, sometimes more than 30° C in a day, through substantial depths of the stratosphere. The warmings, once thought to result from some kind of extraterrestrial impulse, are almost certainly due to subsidence on the order of one or two kilometres per day. They temporarily destroy or even reverse the westerly circulation. The last of these warmings brings winter to an abrupt and spectacular end, usually at or before the spring equinox. These invisible events also transfer much ozone downward into the lower stratosphere from the layers of rapid formation higher up, giving the high-latitude belts the curious phenomenon of a spring ozone maximum.

**Summer easterlies of the mesosphere and stratosphere.** Summer easterlies at high levels, established after some weeks of further vacillation, blow continuously until September (March in the Southern Hemisphere). They are probably strongest at about 70 kilometres in about 50° latitude, where they attain 60 metres per second. As Figure 34 shows, they probably tilt equatorward as one descends, and below 40 kilometres the strongest winds are in about 15° latitude. These easterlies lack the spectacular disturbances of winter but contain rather feeble westward-drifting troughs and ridges.

Between these vast zonal currents, in broadly equatorial latitudes, mean wind speeds are low. In the layer above about 15 kilometres and up to at least 40 kilometres within the 30° parallels, there is a remarkable regime in which the zonal wind reverses on a unique 26-month cycle (the quasi-biennial oscillation), with easterlies giving way to westerlies all around the Earth and then reappearing 26 months later. It has been shown that a 26-month component of variation occurs in many atmospheric phenomena, but in the equatorial stratosphere it dominates the motion of the winds. (F.K.H./P.J.S.)

# MAJOR FORMS OF WEATHER DISTURBANCES

The Earth's wind systems, as noted above, are subject to transient but often severe disturbances. In most cases these storms constitute centres of large-scale air ascent. As such, they act as important generators of clouds and precipitation. The principal types of storms include thunderstorms, tornadoes, and hurricanes and typhoons.

## Thunderstorms

A thunderstorm is a short-lived storm that is produced by clouds of great vertical extent and that is always accompanied by lightning and thunder. A thunderstorm frequently produces strong, gusty winds, heavy rain, and occasionally hail. Almost invariably, thunderstorms are associated with cumulonimbus clouds, which are dense rain clouds with exceptional vertical development that look like mountains or enormous towers. In an advanced stage, their summits have a smooth, fibrous appearance and occasionally resemble a huge anvil (Figure 35). The base of a cumulonimbus cloud is usually dark because it has great depth. Sometimes, particularly in the mountainous areas of the western United States, water drops from a thunderstorm are small, and they evaporate before reaching the ground as precipitation.

Association with cumulonimbus clouds

### VISUAL AND AURAL PHENOMENA

Within a cumulonimbus cloud there are strong updrafts that carry cloud particles and raindrops to the cold upper parts of the clouds. The interaction of water and ice particles causes the separation of electric charge. In general, positive charge is concentrated in the upper regions of the cloud and negative charges near the central regions. When the accumulated electric charge becomes sufficiently large, a lightning discharge occurs. A South African authority, D.J. Malan, has proposed a cloud charge distribution that can schematically represent the distribution in a typical mid-latitude cumulonimbus cloud (Figure 35). A net positive charge is located at an altitude of 10 kilometres (where the temperature is about −45° C) and a net negative charge at five kilometres (−15° C); both charges are of the same magnitude, usually 40 coulombs (an absolute coulomb = 0.1 electromagnetic unit). A smaller positive charge is located at an altitude of two kilometres (+5° C). The electrical potential between the cloud and ground is on the order of $10^8$ volts. At this time the charge resides on ice particles or water drops or both. Discharges can occur between any of the oppositely charged regions or between the cloud and the surrounding air.

**Cloud-to-ground lightning.** The flash of cloud-to-ground lightning is initiated by electrical breakdown between the small positive charge region and the negative charge region. On a time scale measured in fractions of a second, high-speed cameras can record luminous events in the flash (Figure 36). A faint luminous process in regular distinct steps, typically of 50-metre length, at time intervals of 50 microseconds (μsec), descends in a downward branching pattern toward the ground. Carrying currents on the order of hundreds of amperes, this stepped leader or initial stroke propagates at a typical velocity of $1.5 \times 10^5$ metres per second, or about one two-thousandth the speed of light. Diameter estimates for the stepped leader range from a few centimetres to a few metres. The current carrying core is on the order of one or two centimetres, and photographic measurements indicate that a sheath of

Branching pattern

From M.A. Uman, *Lightning* (1969); McGraw-Hill Book Co.

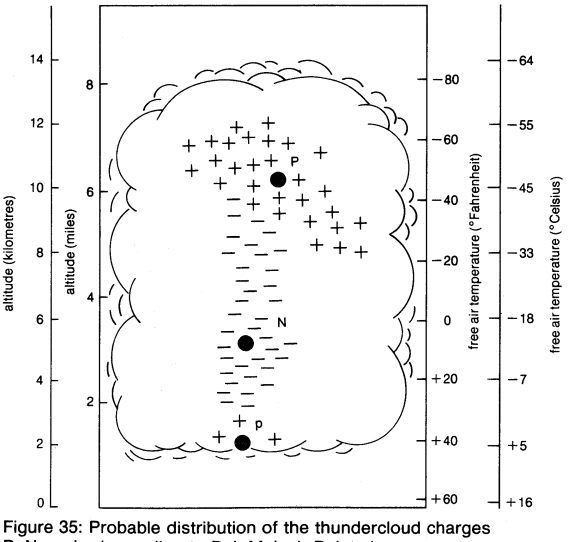



Figure 35: Probable distribution of the thundercloud charges P, N, and p (according to D.J. Malan). Point charges are indicated by solid black circles.

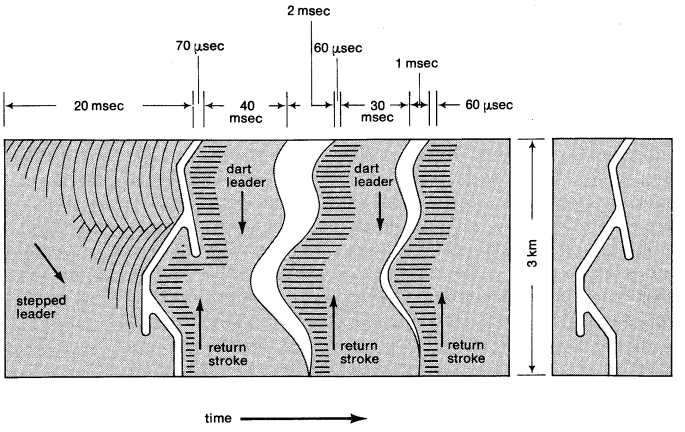



Figure 36: *Time-resolved lightning flash between cloud and ground.*
(Left) Flash of lightning as it would be recorded by moving film. (Right) The same flash recorded by stationary film.
From M.A. Uman, *Lightning* (1969); McGraw-Hill Book Co.

electric charge (corona sheath) with a diameter of one to 10 metres exists around the core.

As the branching process nears the ground, approximately five coulombs of charge have been deposited on the channel, inducing an opposite charge on the ground and increasing the electric field between the leader and the point to be struck. An upward discharge occurs from the ground, church steeple, house, or other object to meet the stepped leader about 50 metres above the surface. At this moment of junction the cloud is short-circuited to the ground and a highly luminous return stroke of high-current-occurs. Stepped leaders that have not reached the ground become the branches of the return stroke, and charge on the branches flows into the main channel. The five coulombs of charge typically deposited on the stepped leader flow to ground in a few hundred microseconds to produce peak currents that are usually on the order of 20,000 amperes but may range from a few thousand to 200,000 amperes. Temperatures in the channel are on the order of 30,000 K (50,000° F). Due to the junction process occurring near the ground, the time to peak current measured at the ground is typically 10 microseconds. As the charge avalanches toward the ground, the luminosity (return stroke) propagates toward the cloud base at $5 \times 10^7$ metres per second, or approximately one-tenth the speed of light, and the high-current-carrying core expands to a diameter of a few centimetres. Laboratory experiments suggest that when pressure equilibrium is attained between the channel and the surrounding air, the channel approximates a high-current arc characterized by a current density of 1,000 amperes per square centimetre.

In the rapid passage from ground to cloud, the luminous return stroke is observed to pause at points where the branches join the main channel (Figure 37), and the channel is observed to brighten as charge from the branch flows into the channel. The stroke then continues its upward propagation, reaching the cloud base in about 70 microseconds; the downward propagating stepped leader traverses the same distance in 20 milliseconds (msec). At this moment there is a pause for tens of milliseconds, and the channel cools to a few thousand kelvins. If a second stroke occurs, it begins with the appearance of a dart of light, perhaps 50 metres in length, propagating down the channel of the return stroke. At a speed of $2 \times 10^6$ metres per second (about one one-hundredth the speed of light), the dart leader carries a current on the order of 1,000 amperes toward the ground. Once again a charge centre in the cloud is short-circuited to the ground, and the return stroke occurs. This sequence of dart leader-return stroke typically occurs three to four times, although a flash to the ground that had 26 strokes and lasted two seconds has been reported. When a flash does have more than one stroke, the subsequent return strokes draw charge from successively higher charge centres in the cumulonimbus clouds.

Strokes and energy dissipation

During the return-stroke stage, approximately $10^5$ joules of energy per metre are dissipated within the lightning channel. This energy is divided among the dissociation, ionization, excitation, and kinetic energy of the particles, the energy of expansion of the channel, and radiation. Spectroscopic measurements reveal that the air particles, principally nitrogen, oxygen, and water molecules, are split into their respective atoms and that on the average one electron is removed from each atom. The conversion from air molecules to a singly ionized plasma occurs in a few microseconds. At this time the temperature of the plasma is at least 30,000 K and the pressure is greater than 1,000 kilopascals (10 atmospheres). The high pressure greatly exceeds the ambient (surrounding) pressure. The return-stroke channel expands at a supersonic rate, and the shock wave decays to a sound wave that is eventually heard as thunder. Because it is estimated that 1 percent of the input energy is stored in the particles and less than 1 percent emitted as radiation in the visible and infrared region (4,000 to 11,000 angstroms [Å], where $Å = 10^{-8}$ centimetre), it is probable that most of the energy dissipated goes into the energy of channel expansion, a process requiring no more than 10 to 20 microseconds.

A small percentage of discharges between the cloud and ground are actually initiated at the ground and propagate upward to a charged region in the cloud. These discharges often are initiated (or triggered) by tall structures or by

By courtesy of (top) the University of Arizona, Tucson; photograph (bottom), Icelandair

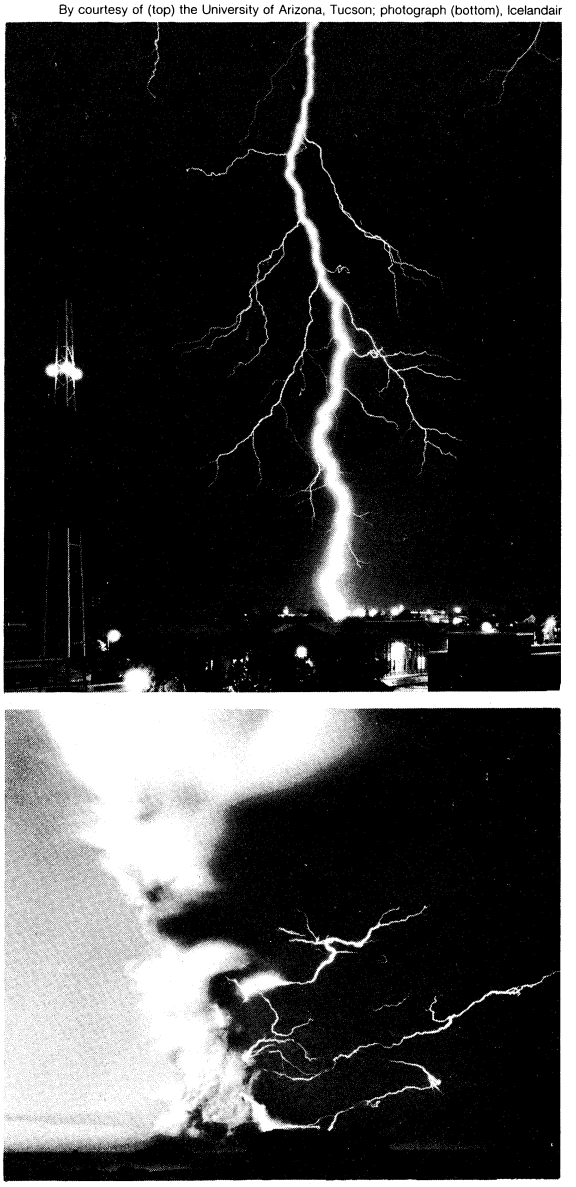


Figure 37: *Natural lightning occurrences.*
(Top) Cloud-to-ground lightning flash. (Bottom) Lightning in a volcano cloud on Surtsey, an island off the Icelandic coast.

**Table 3: Velocities of Lightning Components and Related Data**

| | minimum* | representative | maximum* |
|---|---|---|---|
| **Stepped leader** | | | |
| Length of step (m) | 3 | 50 | 200 |
| Time interval between steps (μsec) | 30 | 50 | 125 |
| Average velocity of propagation of stepped leader (m/sec)† | $1.0 \times 10^5$ | $1.5 \times 10^5$ | $2.6 \times 10^6$ |
| Charge deposited on stepped-leader channel (coul) | 3 | 5 | 20 |
| **Dart leader** | | | |
| Velocity of propagation (m/sec)† | $1.0 \times 10^6$ | $2.0 \times 10^6$ | $2.1 \times 10^7$ |
| Charge deposited on dart-leader channel (coul) | 0.2 | 1 | 6 |
| **Return stroke‡** | | | |
| Velocity of propagation (m/sec)† | $2.0 \times 10^7$ | $5.0 \times 10^7$ | $1.4 \times 10^8$ |
| Current rate of increase (ka/μsec)§ | <1 | 10 | >80 |
| Time to peak current (μsec)§ | <1 | 2 | 30 |
| Peak current (ka)§ | — | 10–20 | 110 |
| Time to half of peak current (μsec) | 10 | 40 | 250 |
| Charge transferred excluding continuing current (coul) | 0.2 | 2.5 | 20 |
| Channel length (km) | 2 | 5 | 14 |
| **Lightning flash** | | | |
| Number of strokes per flash | 1 | 3–4 | 26 |
| Time interval between strokes in absence of continuing current (msec) | 3 | 40 | 100 |
| Time duration of flash (sec) | $10^{-2}$ | 0.2 | 2 |
| Charge transferred including continuing current (coul) | 3 | 25 | 90 |

*The words maximum and minimum are used in the sense that most measured values fall between these limits. †Velocities of propagation are generally determined from photographic data and thus represent "two-dimensional" velocities. Since many lightning flashes are not vertical, values are probably slight underestimates of actual values. ‡First return strokes have slower average velocities of propagation, slower current rates of increase, longer times to current peak, and generally larger charge transfer than subsequent return strokes in a flash. §Current measurements are made at the ground.
Source: Uman, *Lightning*, 1969.

towers on hilltops (Figure 38, top right). The upward branching of such discharges makes them visually distinguishable from their "right-side-up" counterparts, giving the impression of a cloud-to-ground lightning flash that is upside down.

Some physical properties of cloud-to-ground lightning are summarized in Table 3.

Occurrence, duration, and current flow of intracloud lightning

**Cloud-to-cloud and intracloud lightning.** True cloud-to-cloud lightning is rare because most lightning flashes occur within a cloud. The first lightning flash in a thundercloud is typically an intracloud discharge. When an intracloud discharge occurs, the cloud becomes luminous for approximately 0.2 second. The discharge is initiated by a leader that propagates between charge centres. The 0.2-second illuminosity is continuous and has several pulses of higher luminosity of one-millisecond duration superimposed upon it. This situation suggests minor return strokes as the leader contacts pockets of charge, but the similarity ends there. During the 0.2 second, the amount of the charge transfer is probably similar to the amount involved in a ground discharge: 20 coulombs, with a range from 0.3 to 100 coulombs. The mean velocity of propagation of the intracloud flash is $1 \times 10^4$ to $2 \times 10^4$ metres per second. Electric currents associated with the luminous brightening are probably in the range of 1,000 to 4,000 amperes. Strikes to aircraft indicate peak currents of only a few thousand amperes, an order of magnitude less than currents in ground flashes. Rise times to peak currents are measured in milliseconds, three orders slower than rise times in return strokes. The energy of intracloud flashes is unknown.

**Thunder.** When air is crossed by a spark or a lightning flash, the air is heated rapidly and the cylindrical column expands at supersonic speed. Within a metre or two the shock wave decays to a sound wave. The sound heard as thunder comes from the entire channel length and is modified by the intervening medium. The result is a series of sounds that are variously described as peals, claps, rolls, and rumbles. These can conveniently be condensed into claps, which are sudden loud sounds, and rumbles, which are all other sounds of thunder. At short distances of a few hundred metres, the thunder begins with a sudden clap followed by a long rumble. At extended distances the thunder begins with a rumble. Because light travels at 299,300 kilometres per second and sound travels at about

By courtesy of (left and top right) Richard E. Orville, (bottom right) William S. Bickel, University of Arizona, Tucson

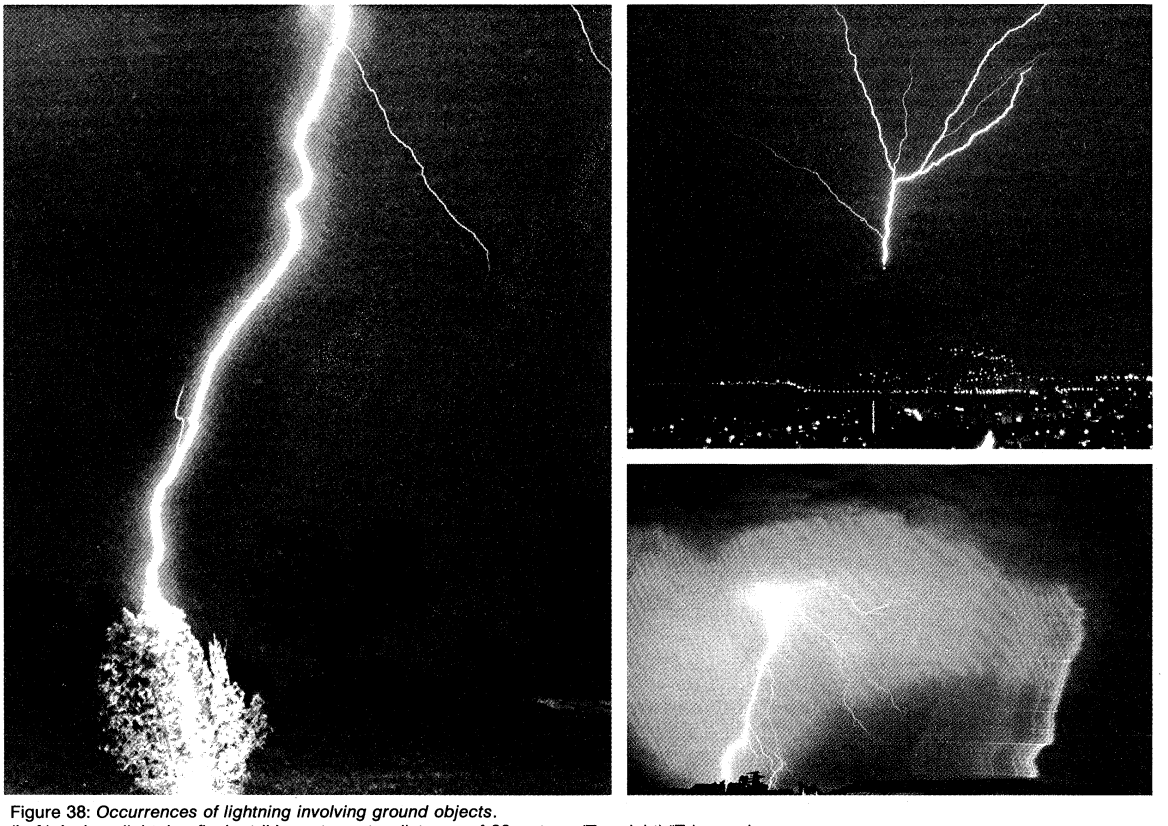


Figure 38: *Occurrences of lightning involving ground objects.*
(Left) A close lightning flash striking a tree at a distance of 60 metres. (Top right) "Triggered lightning"; the discharge is triggered by the presence of the tall tower atop Mount San Salvatore, near Lugano, Switz. (Bottom right) Spectrum of a lightning flash.

335 metres per second, observation of the time of arrival of thunder permits calculation of the distance to the flash; the sound will travel 1.6 kilometres in five seconds. For close lightning the elapsed time until the beginning of the thunderclap gives the minimum distance to the flash, and the time duration to the rumble is a minimum estimate of the channel length. A thunderclap heard two seconds after the flash, followed by a rumble of 20 seconds, for example, indicates a distance to the flash of approximately 610 metres and a channel length at least 6.4 kilometres. For reasons that include the atmospheric temperature lapse rate, wind shear, and terrain features, thunder is rarely heard at distances greater than 24 kilometres. The energy spectrum of thunder has been recorded and shows that most of the energy is in the low-frequency audible range with a maximum at 50 hertz (cycles per second).

(R.E.Or./L.J.B./Ed.)

## CAUSES OF THUNDERSTORMS

### Instability and convection

Thunderstorms occur when the atmosphere exhibits instability. In essence, when a weather system is unstable, a small displacement leads to a larger one. When the atmosphere is unstable, a volume of air having an upward displacement for any reason will continue moving upward at an accelerating rate. If a sufficiently large mass of air begins rising, an updraft will result. If the air is moist as well as unstable, a cloud will form. Further growth can lead to precipitation and lightning.

The atmosphere becomes unstable when the temperature decreases rapidly with height, this change resulting most often from solar heating of the Earth's surface and of the air in the lowest layers of the atmosphere. Some of the heat is transported upward by means of molecular motion and small-scale turbulent eddies of air. If the incoming energy exceeds that being transported upward, the temperatures in the lowest layers rise, causing the temperature lapse rate (the rate of temperature decrease with height) to increase. When the lapse rate exceeds certain specific values depending on the humidity properties of the air, the atmosphere becomes unstable. If upward air motion is initiated by a mountain ridge, for example, a rising convection current is set in motion. The ascending stream of air will be warmer and more buoyant than the surrounding air. Heat and moisture are carried upward. Countercurrents transport cooler and drier air downward. The associated puffy clouds that start as cumulus and may become cumulonimbus are called convective clouds.

Such clouds may be initiated in a variety of ways. Some early studies indicated that convective clouds and thunderstorms were initiated as a result of excessive warming of certain surface areas. The temperature differences were attributed to differences in the colour and hence the absorption properties of various soils and rocks. Sometimes thermals may be traced to such hot spots, but evidence indicates that most thunderstorms are set off as a result of organized lifting of the air.

As noted previously, mountain ridges are important in generating thunderstorms. Not only do the ridges serve as a barrier that forces the air to rise but they also become what are sometimes called high-level heat sources. Solar heating of the high terrain causes the air close to them to be warmer than the air at the same altitudes over the adjacent valleys. Because the warm air is less dense, it rises through the surrounding cooler and heavier air. The importance of this effect is shown by the fact that, even in mountainous areas, most thunderstorms occur during the afternoon and early evening hours when it is warmest.

Thunderstorms are often initiated when a cold front advances toward moist, unstable air (see above), which is forced to rise over the frontal surface. Sometimes lines of thunderstorms occur above and nearly parallel to a cold frontal surface. More often, particularly over the central United States, long lines or zones of thunderstorms develop in the warm air many tens of kilometres, sometimes several hundred kilometres, ahead of the cold front.

### Prefrontal squall lines

The tendency of these so-called prefrontal squall lines to be more or less parallel to the front has suggested that they have been initiated by a dynamic mechanism related to the front.

In the tropics, weather fronts of the kinds so commonly noted at higher latitudes are almost never observed. Air masses of contrasting temperature do not come into contact, as is the case at the polar front, when air from the polar regions encounters air from the tropics. Nevertheless, the regions of maximum thunderstorm frequency occur in equatorial and tropical regions (Figure 39). These storms are triggered by air rising as a result of converging wind systems. When the northeast trade winds meet the

By courtesy of the World Meteorological Organization

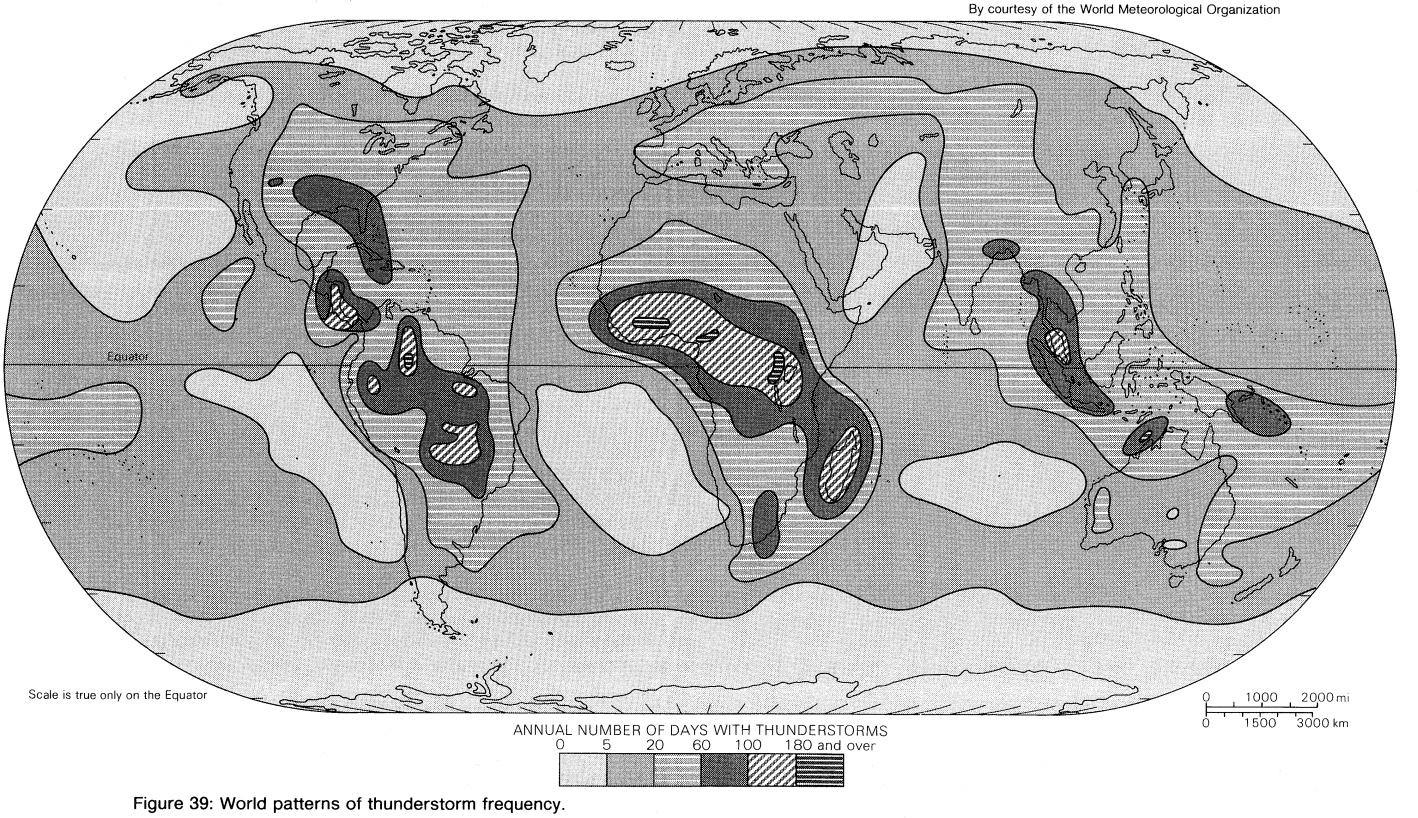



Figure 39: World patterns of thunderstorm frequency.

southeast trades, the air at the zone of confrontation must rise. Because it is both moist and unstable, showers and thunderstorms are produced in abundance.

Disturbances in the low-altitude wind field may lead to convergence and ascending air. Such disturbances occur often both in equatorial areas and at higher latitudes and give rise to cumulonimbus clouds and thunderstorms.

## TYPES OF THUNDERSTORMS

At one time it was common to classify thunderstorms according to where they occurred; for example, as airmass, frontal, or orographic (mountain-related) thunderstorms. Since the early 1960s, it has been found more meaningful to classify them according to the chief characteristics of the storms themselves. Such characteristics depend largely on the properties of the environment in which a given thunderstorm develops.

Isolated thunderstorms, especially those that occur under atmospheric conditions such that the wind velocity does not change markedly with height, are sometimes called air-mass, or local, thunderstorms. They are mostly vertical in structure, relatively short-lived, and usually do not produce violent weather at the ground level. Extensive flight measurements indicate that such thunderstorms are composed of one or more individual cells, each of which passes through a definable life cycle. Early in the growth of the cloud, air motions are mostly upward not as a steady, uniform stream but as one that is composed of a series of rising eddies. The precipitation particles grow as the cloud grows; when the accumulated water and ice becomes excessive, a downdraft is started. At maturity a local storm may be composed of intense updrafts and downdrafts side by side (Figure 40). In its later stages the updraft spreads throughout the cell and diminishes in intensity as the precipitation falls out.

Organized storms

Violent weather at ground level usually results from a type of thunderstorm that is sometimes referred to as an organized storm. Over the central United States, it commonly occurs when the winds in the middle layer of the atmosphere are from the southwesterly quadrant and are strong. Although there is still considerable debate about the structure of such storms, many authorities are convinced that in such thunderstorms the updraft is tilted from the vertical. Storms of this sort may last for many hours, remaining in a constant state of regeneration because moist, unstable air is drawn into the lower part of the cloud. Long swaths of hail can be produced by these storms. Strong, gusty winds are very common, and tornadoes sometimes are generated.

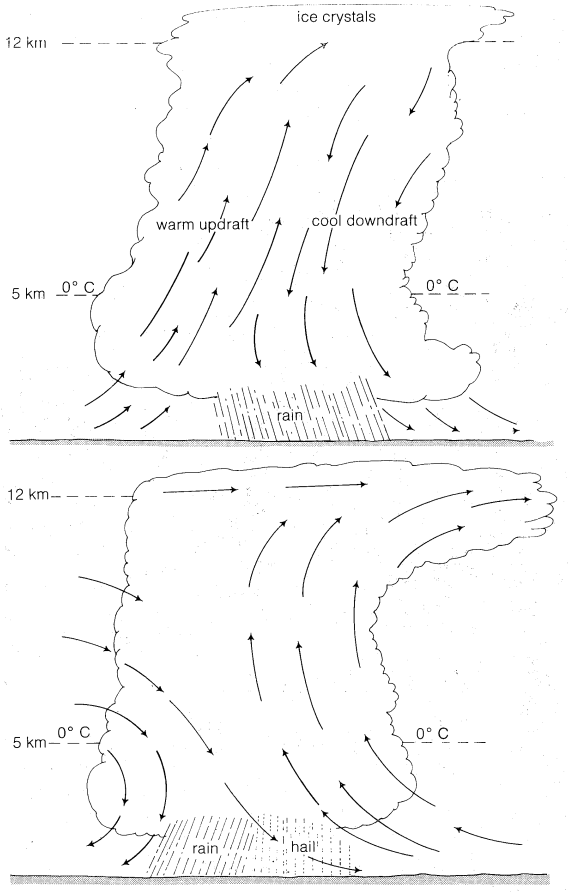



Figure 40: *Thunderstorm models.*
(Top) Mature stage of an isolated thunderstorm cell.
(Bottom) Cross section through a hail-producing organized thunderstorm.

## ENERGY OF THUNDERSTORMS

The energy to drive thunderstorms is supplied by a variety of sources, but most of it comes from the heat released when water vapour condenses to form cloud droplets. For every gram of water condensed, about 600 calories of heat are made available. When the water freezes in the upper parts of the cloud, another 80 calories of heat per gram of water are released. This energy goes to increase the temperature of the updraft and, in part, is converted to kinetic energy of upward and downward air movement. If the quantity of water that is condensed in and subsequently precipitated from a cloud is known, then the total energy of a thunderstorm can be calculated. In an average thunderstorm, the energy released amounts to about $10^7$ kilowatt-hours, which is equivalent to a 20-kiloton nuclear warhead. A large, severe thunderstorm might be 10 to 100 times more energetic.

## PHYSICAL CHARACTERISTICS

An individual thunderstorm cell may be three kilometres in diameter, extend to an altitude of eight to 10 kilometres, and last less than 30 minutes. A larger local thunderstorm would be composed of many cells in different stages in their life history. Such a storm can be tens of kilometres in diameter, its summit may extend to altitudes exceeding 18 kilometres, and its duration may be many hours.

When referring to the diameter of a thunderstorm, it is necessary to specify what is being measured. The quantities cited above refer to the diameter of the cloud at middle altitudes, the distance observed by an airplane flying through the cloud at an altitude of perhaps five kilometres. In such a penetration, the aircraft might encounter several updrafts and downdrafts. The most extensive measurements of thunderstorm properties by means of airplanes were made by the Thunderstorm Project conducted in 1946 and 1947 by the U.S. Weather Bureau in collaboration with the U.S. Navy and Army air forces.

**Updrafts and downdrafts.** Although updrafts and downdrafts encountered in Ohio and Florida thunderstorms sometimes had diameters exceeding four kilometres, most often they were between about 500 and 2,500 metres at altitudes from three to eight kilometres. At flight levels near 1,500 metres, the drafts tended to have a greater diameter and be weaker than at greater heights in the clouds.

Measured updraft speeds

Updraft speeds measured by the Thunderstorm Project ranged to a maximum of 26 metres per second (m/sec) but usually were less than half that much. They increased with height and averaged 5.0, 7.3, and 8.4 m/sec at altitudes of 1.5, 4.5, and 7.6 kilometres, respectively. At the same altitudes, mean downdrafts were 4.6, 6.1, and 7.3 m/sec, respectively. At greater altitudes, higher updraft speeds have been measured by aircraft and radar and have been inferred on theoretical grounds. Updrafts exceeding 20 m/sec are not considered unusual in the upper parts of large thunderstorms. Airplanes flying through them at altitudes about 10 kilometres have measured updrafts exceeding 30 m/sec. The strongest updrafts have been observed in severe organized thunderstorms, which often are many tens of kilometres in diameter. Lines or zones of such storms sometimes extend for many hundreds of kilometres.

**Vertical extent of thunderstorms.** The heights of thunderstorms have been measured by radar and by aircraft flying over their tops. They commonly extend to altitudes of more than 11 kilometres and sometimes to more than 20 kilometres. On very unstable days when the atmosphere is moist, the upper limit of the thunderstorm cloud

is determined by the depth of the troposphere, that region of the atmosphere lying below the stratosphere. On such occasions the rising updraft air would be warmer than its environment until it penetrates through the base of the stratosphere. Since this layer is characterized by air temperatures that are nearly constant or may increase with height, it has great stability. The momentum of an updraft would carry it into the stratosphere, but after a short penetration the rising air would be cooler and heavier than the surrounding air. The resulting downward force would stop further upward motion.

The base of the stratosphere varies with both latitude and season of the year. In general, it ranges from about 10 to 20 kilometres with the higher found in the summer at lower latitudes. When the rising air of the cloud encounters the stable stratosphere, it spreads outward and forms the anvil cloud so characteristic of thunderstorms (see above Figure 11, top centre). If winds are light, the anvil may extend in all directions around the cloud. Most of the winds at the anvil altitudes carry cloud material downwind.

Thunderstorm tops do not usually reach the stratosphere. More often their vertical extent is determined by a stable layer in the troposphere. For a number of reasons, layers may be formed where the temperature is constant or increases with height. Sometimes the air temperature decreases only slowly with height. Regions where the temperature decreases slowly or increases with height are stable layers, and a rising volume of air encountering them often cannot penetrate them. As a result, the stable layer determines the maximum vertical extent of the cloud and the altitude where the cloud spreads out to form the anvil.

**Turbulence.** An airplane flying through a thunderstorm is commonly carried upward and downward by the drafts in the storm. In addition, it often is buffeted from side to side and up and down. This motion is called turbulence, which is caused by large changes in air motions over distances comparable to the dimensions of the airplane. Turbulence not only causes discomfort for the crew and passengers but also subjects the aircraft to undesirable stresses.

Turbulence may be expressed in various units. Commonly, the g unit, equal to the acceleration of gravity of the Earth, is used. A gust of 1 g will cause severe aircraft turbulence. In the upper part of violent thunderstorms, vertical accelerations of about 3 g have been reported.

**Movement of thunderstorms.** The motion of a thunderstorm is determined largely by the average wind velocity in the layer of the atmosphere in which the storm develops. The average speed of 120 thunderstorms observed in Florida and Ohio in the United States was 20 kilometres per hour, but some storms moved much faster. In extreme circumstances thunderstorms may move at 65 to 80 kilometres per hour.

When considering the movement of a thunderstorm, the storm's dynamic character must be taken into account. Most storms are in a constant state of change, with new cells developing while old ones dissipate. When wind speeds are light, an individual cell may move very little, less than two kilometres, during its lifetime; however, new cells forming downwind may give the illusion of rapid cloud motion. This type of behaviour is often observed in mountainous areas. The first storms of the day generally form over the ridges. As the day progresses, new clouds develop near the existing ones but closer to the valleys. An observer who sees the nearest edge of the clouds coming closer may readily assume that the original clouds are moving over the valley.

Although new cloud developments generally occur downwind of an existing storm, this is not necessarily the case. Sometimes they grow on the flanks, and, as a result, visual observations may lead to the notion that there is little relation between the winds and thunderstorm motion. As would be expected, this view is most likely to arise when wind speeds are low.

## WEATHER UNDER THUNDERSTORMS

The air in a thunderstorm downdraft descends from altitudes where the temperature is lower than at the ground. Moreover, the downdraft is maintained at a cooler temperature than its environment by the evaporation of water and ice particles. The sinking air is not only heavier than the surrounding air but its so-called horizontal momentum also differs from that of the surrounding air. If the descending air originated at an altitude of 10 kilometres, for example, it would reach the ground with a horizontal velocity that somewhat resembles the wind velocity at its level of origin. When the air strikes the ground, it usually moves outward ahead of the storm at a higher speed than the speed of the storm itself.

Wind velocities and temperature

In extreme circumstances, the outrushing, cool air may reach velocities of 100 kilometres per hour or more and do extensive damage to property and vegetation. This severe wind most often occurs when organized lines of severe thunderstorms form in an environment where the middle level winds are very strong. When serious wind damage is produced by such a storm, the victims may suspect that it was caused by a tornado. Of course, if a funnel cloud is observed, then the nature of a storm would be obvious. If a funnel cloud is not observed, the character of the damage can be revealing. Tornadoes blow debris in a tight circular pattern, whereas the outflowing air from a thunderstorm pushes it mostly in one direction.

An observer on the ground watching a thunderstorm approach can feel the gusty, cool air before the storm passes overhead. The outspreading, downdraft air forms a pool some 500 to 2,000 metres deep. Often there is a very distinct boundary between the cool air and the warm, humid air in which the storm formed in the first place. The passage of such a boundary is easily recognized as wind speeds increase and the air temperature suddenly drops. Over a five-minute period, a cooling of more than 5° C is not unusual, and cooling that is twice as great is not unknown.

Precipitation from storms

By the time the cool air begins spreading over the ground, rain usually is reaching the surface. Sometimes all the raindrops evaporate while falling, and the result is a dry thunderstorm. At the other extreme, thunderstorms can produce torrential rain and hail. Under a cumulonimbus cloud, impending rain usually can be predicted when the cloud base darkens, indicating a deep layer of cloud obscuring the rays from the Sun. Often, but not usually, a flash of lightning occurs and the precipitation follows several minutes later. There is still disagreement over the interpretation of this sequence of events. Once the precipitation commences, it builds up rapidly, commonly reaching its maximum intensity between five to 10 minutes after the start. In small Ohio thunderstorms, for example, the maximum five-minute rate was 122 millimetres per hour, but most often was less than one-tenth of this amount. The average thunderstorm yielded $2 \times 10^8$ kilograms (220,000 short tons) of rain, but some large storms produced 10 times more. Large organized storms can generate $10^{10}$ to $10^{12}$ kilograms of rain.

Table 4 lists certain extreme rainfall quantities in various parts of the world. They represent measurements at a single point. As the duration of the rainfall increased, the average rates decreased but were extreme in every case shown. Such rapid accumulations of water invariably cause flooding; greater damage is, of course, associated with longer durations and greater quantities.

## OCCURRENCE OF THUNDERSTORMS

Thunderstorms are most likely to occur in a moist, unstable atmosphere. Air masses having these properties are commonly formed over equatorial and tropical areas. When a body of air moves over low-latitude oceans, it is humidified by evaporation from the underlying water surface. Heat is transferred from the warm ocean water to the air. In addition, the nearly direct rays from the Sun warm the moist, lowest layers of the atmosphere. As a consequence of these processes, moist, tropical air masses develop suitable properties for thunderstorm formation.

Disturbances in the wind field caused by pressure perturbations or changes in topography leading to an upward displacement of the air initiate the development of thunderstorms. All the conditions necessary for their occurrence are most often met over the land areas in the equatorial zone between 10 degrees north and south latitudes where

**Table 4: Some Extreme Rainfall Rates***

| date | place | duration | amount (inches) | amount (cm) | average rate (mm/hr) |
|---|---|---|---|---|---|
| April 5, 1926 | Opids Camp, Calif. | 1 min | 0.65 | 1.65 | 990 |
| Sept. 30, 1925 | Haughton Grove, Jamaica | 3 min | 1.6 | 4.07 | 814 |
| Nov. 29, 1911 | Porto Bello, Panama | 5 min | 2.5 | 6.37 | 765 |
| May 12, 1916 | Plumb Point, Jamaica | 15 min | 8.0 | 20.3 | 812 |
| June 22, 1947 | Holt, Mo. | 42 min | 12.0 | 30.5 | 436 |
| July 14–15, 1947 | Baguio, Phil. | 24 hr | 46.0 | 117.0 | 48.7 |

*Probably from thunderstorms, but not documented.

Frequency of storms

the average number of days with thunderstorms exceeds 100 per year. In certain places in equatorial Africa and South America, there are more than 180 days of thunderstorms in an average year (Figure 39).

At higher latitudes, the frequency of thunderstorms depends on the character of the topography and the frequency of invasions of moist, tropical air. Since this happens most often in the spring and summer, the Northern Hemisphere experiences most of its thunderstorms between May and September, whereas thunderstorm maximums are found some six months later in the Southern Hemisphere.

Over the large continental area of North America, thunderstorms occasionally occur, even in the winter, in the states along the coast of the Gulf of Mexico. As summer approaches, moist, tropical air gradually extends to higher latitudes and thunderstorms do likewise. By July and August they occur on the average on more than six days per month over the flatlands of central Canada.

During midsummer, areas of high thunderstorm frequency are centred over the Florida peninsula and northern New Mexico where 18 to 20 thunderstorm days per month are experienced. The frequent thunderstorms in the New Mexico–Colorado area are initiated as warm, humid air is forced to rise on the east slope of the Rocky Mountains. On a yearly basis there are about 90 thunderstorm days in central Florida.

Association with summer monsoons

Thunderstorms constitute a common feature of the summer monsoons (see above) in many parts of the world, particularly over southern Asia. As solar radiation warms the continent, an ocean-to-land air current is established. Moist, unstable air from the Indian Ocean is carried inland and is forced to rise up the steep slopes of the Himalayas. As a result, showers and thunderstorms are produced in great abundance. Record high rainfalls occur.

In cold regions, poleward of about 60° latitude, thunderstorms are scarce or nonexistent because of the cold air near the ground and the stable atmospheric conditions. There also are few thunderstorms over those parts of the Earth dominated by semipermanent high-pressure centres. In these regions, which correspond to the great deserts of the world, the air generally descends. This causes a drying of the air and a stabilization of the atmosphere. As a result, thunderstorm development is inhibited.

#### OBSERVATIONS OF THUNDERSTORMS

Most of the existing data on thunderstorms have been accumulated by weather observers. Unfortunately, by visual observations, it is difficult and often impossible to obtain quantitative information about the characteristics of the clouds that produce the thunder. When widespread thunderstorms occur, the sky commonly is overcast with many intermingled layers of clouds of various types. Cloud summits are frequently obscured.

Advances in aviation, and in particular the development of airplanes of high structural integrity, have made it possible to fly through thunderstorms to measure their internal properties. High-flying airplanes have yielded accurate data on the vertical extent of thunderstorms.

Radar observation of storms

The development of radar during World War II contributed much to scientific knowledge about thunderstorms. Most of the available statistics on thunderstorm dimensions and movements have been obtained by means of radar. Conventional radar sets measure the location and reflectivity of water and ice particle concentrations. By noting changes in the intensity, size, or position of radar echoes, it has been possible to infer the nature of the motions in the thunderstorms.

Networks of radar sets are employed by the civil and military weather services in several countries to detect and track thunderstorms. All commercial airliners are now equipped with radar to allow the pilots to observe thunderstorms and take evasive action and thereby reduce flight hazards and increase passenger comfort.

Since about 1960, atmospheric scientists have used pulsed-Doppler radar sets. Such equipment not only can make the same observations as were made by earlier radars but can measure the instantaneous fields of motion of the precipitation particles. In certain circumstances, this information can be used to measure the speeds of updrafts and downdrafts. Furthermore, it is sometimes possible to infer the size of the precipitation particles.

Weather satellites photograph cloud patterns and telemeter the pictures back to Earth. Radiometers on board the satellites supply information on the temperatures of the clouds that are observed. On the basis of appearance and temperature, it is possible to infer when a cloud is a cumulonimbus and, hence, probably a thunderstorm source. At the present time, satellite data are revealing patterns of convective cloud and thunderstorm development on a scale never seen before. The information has been particularly valuable over the vast, low-latitude oceanic areas that were never adequately observed in the past. (L.J.B./Ed.)

## Tornadoes, waterspouts, and whirlwinds

Tornadoes, whirlwinds, and waterspouts are atmospheric vortices, or rotating funnel-cloud air masses of small diameter. They are differentiated from each other by the intensity of their rotation and by the surfaces that they traverse. Though tornadoes and whirlwinds both travel over landmasses, whirlwinds are atmospheric systems smaller than tornadoes. Waterspouts are tornadoes that form or pass over a water surface.

#### GENERAL CHARACTERISTICS

**Tornadoes.** The name tornado comes from the Spanish *tronada* ("thunderstorm"), which supposedly was derived from the Latin *tornare* ("to make round by turning"). The most violent of atmospheric storms, a tornado is a powerful vortex, or "twister," whose rotational speeds are estimated to be close to 480 kilometres per hour but may occasionally exceed 800 kilometres per hour. The direction of rotation in the Northern Hemisphere is usually, though not exclusively, counterclockwise.

The first visible indication of tornado development is usually a funnel cloud (Figure 41), which extends downward from the cumulonimbus cloud of a severe thunderstorm (see above). As this funnel dips earthward, it becomes darker because of the debris forced into its intensifying vortex. Some tornadoes give no visible warning until their destruction strikes the unsuspecting victims. Tornadoes often occur in groups, and several twisters sometimes descend from the same cloud base (Figure 41).

Speed of a tornado

The forward speed of an individual tornado is normally 48 to 64 kilometres per hour but may range from nearly zero to 112 kilometres per hour. The direction of motion is usually from the southwest to the northeast, although tornadoes associated with hurricanes may move from the east. The paths of twisters average only several hundred metres in width and 26 kilometres in length, but large deviations from these averages may be expected; *e.g.*, a devastating tornado that killed 689 persons in Missouri, Illinois, and Indiana in the midwestern United States on March 18, 1925, was at times 1.6 kilometres wide, and its path extended 352 kilometres.

In the short time that it takes to pass, a tornado causes fantastic destruction. There have been cases reported in which blades of straw were embedded in fence posts; a schoolhouse with 85 pupils inside was demolished and the pupils carried 137 metres with none killed; and five railway coaches, each weighing 70 tons, were lifted from their track and one coach was moved 24 metres.



Figure 41: *Tornadoes of the midwestern United States.*
(Left) A typical killer tornado photographed in Tracy, Minn., on June 13, 1968; it traveled 21 kilometres along the ground. (Right) Twin funnels, in one of the worst recorded tornado outbreaks, crossing a highway south of Elkhart, northern Indiana, on April 11, 1965; debris can be seen falling out of the left funnel.
(Left) Eric Lantz, (right) Paul Huffman

Although much remains to be learned about tornado formation and movement, remarkable advances have been made in the effectiveness of tornado detection and warning systems. These systems involve analyses of surface and upper-air weather, detection and tracking of atmosphere changes by radar, and spotting severe local storms.

**Waterspouts.** When a tornado forms or passes over a water surface, it is termed a waterspout. Like tornadoes, they may assume many shapes and often occur in series or families. Measurements of their forward speeds are scarce, but estimates vary from a few kilometres an hour to as high as 64 to 80 kilometres per hour. Contrary to popular opinion, a waterspout does not "suck up" water to great heights, though it may lift the water level a few metres. The main visible cloud consists mostly of freshwater clouds produced by condensation of water vapour; however, a sheath of spray often rotates around the lower portion of the vortex tube.

One of the largest and most famous waterspouts, observed near Massachusetts on Aug. 19, 1896, was witnessed by thousands of vacationers and several scientists. Its height was estimated to be 1,095 metres and its width 256 metres at the crest, 43 metres at centre, and 73 metres at the base. The spray surrounding the vortex tube near the water surface was about 200 metres wide and 120 metres high. The spout lasted 35 minutes, disappearing and reappearing three times. Most waterspouts are smaller, with much shorter lives. This exceptional spout is an example of one that apparently was spawned by thunderstorm-squall conditions, similar to those that produce tornadoes over land.

There are few authentic cases of large ships ever being destroyed by a spout, although spouts are a dangerous hazard to small vessels. A few intense waterspouts have caused deaths when they moved inland over populated areas. The belief that firing a cannonball or other projectile into a spout can "break it up" has no scientific foundation.

**Whirlwinds.** In the general sense, a whirlwind is any rotating mass of air or atmospheric vortex. The term is, however, commonly restricted to atmospheric systems smaller than a tornado but larger than eddies of microscale turbulence. A whirlwind is usually named after the visible phenomenon associated with it; thus there are dust whirls, or dust devils; sand whirls, or sand pillars; and fire, smoke, and even snow whirls, or spouts.

Differences from tornadoes

In contrast to the pendant form of the tornado funnel, a dust or sand devil develops from the ground upward, usually under hot, clear-sky conditions. The whirl shape is normally that of a cylindrical column or an inverted cone. The axis of rotation is usually vertical, but it may be inclined. The direction of rotation may be either clockwise or counterclockwise. Vortices with a horizontal axis of rotation are sometimes called rolls, or rotors. Dust and sand whirls are not nearly as violent as tornadoes, although jackrabbits have been lifted by the more intense vortices. The whirls measure in diameter from several centimetres to a few hundred metres, and visible heights from a few metres to at least 1,500 metres. This is probably not the upper limit, for sailplane pilots have used their spirally ascending currents to soar to above 4,570 metres. Like tornadoes and waterspouts, dust and sand devils often appear in groups or series. Eleven of them were simultaneously sighted in Ethiopia, and in the Mojave Desert in eastern California a series of smaller whirls followed in the wake of a larger primary vortex. Such secondaries are sometimes referred to in India as "dancing devils."

Fire whirlwinds are a problem to the forest rangers who must cope with them. A historical example of a great fire vortex, produced by war, is that which formed over Hamburg following a massive aerial bombing during the night of July 27–28, 1943. The fire fueled a storm of counterclockwise rotation 1.6 to three kilometres in diameter and nearly five kilometres high, with winds estimated to be more than 160 kilometres per hour in some sections.

## OCCURRENCE AND DISTRIBUTION

**Tornadoes.** Tornadoes strike in many areas of the world, but nowhere are they as frequent or as fierce as in the United States. More than 1,100, for example, were reported there during 1973 alone. Direct comparisons of relative tornado frequencies in various countries are biased because observational data are often lacking in sparsely settled regions. It appears, however, that Australia, where several hundred per year have been reported, has the dubious honour of second place. Other countries reporting tornadoes include, but are not limited to, Great Britain, Canada, China, France, Germany, The Netherlands, Hungary, India, Italy, Japan, the Soviet Union, and even Bermuda and the Fiji Islands.

Distribution and frequency of tornadoes

A vast "tornado belt" embracing the Great Plains of the United States and the southeastern portion of the country is threatened by tornadoes every year (Figure 42). Every state in the nation, including Alaska and Hawaii, has experienced twisters. The average frequency varies from more than 100 in Texas to fewer than three in the far western and most of the northeastern states.

A more important criterion for gauging tornado severity is the average number of tornadoes in a unit area, such as a square kilometre or a two-degree square on a map. Computations of this tornado density for a 45-year period show that the greatest concentration of tornadoes per unit area is found in the states of Oklahoma and Kansas. The area with the greatest potential for casualties is that which combines a high tornado incidence rate with a thick population concentration. Southwestern Oklahoma, for example, has the highest tornado incidence per unit

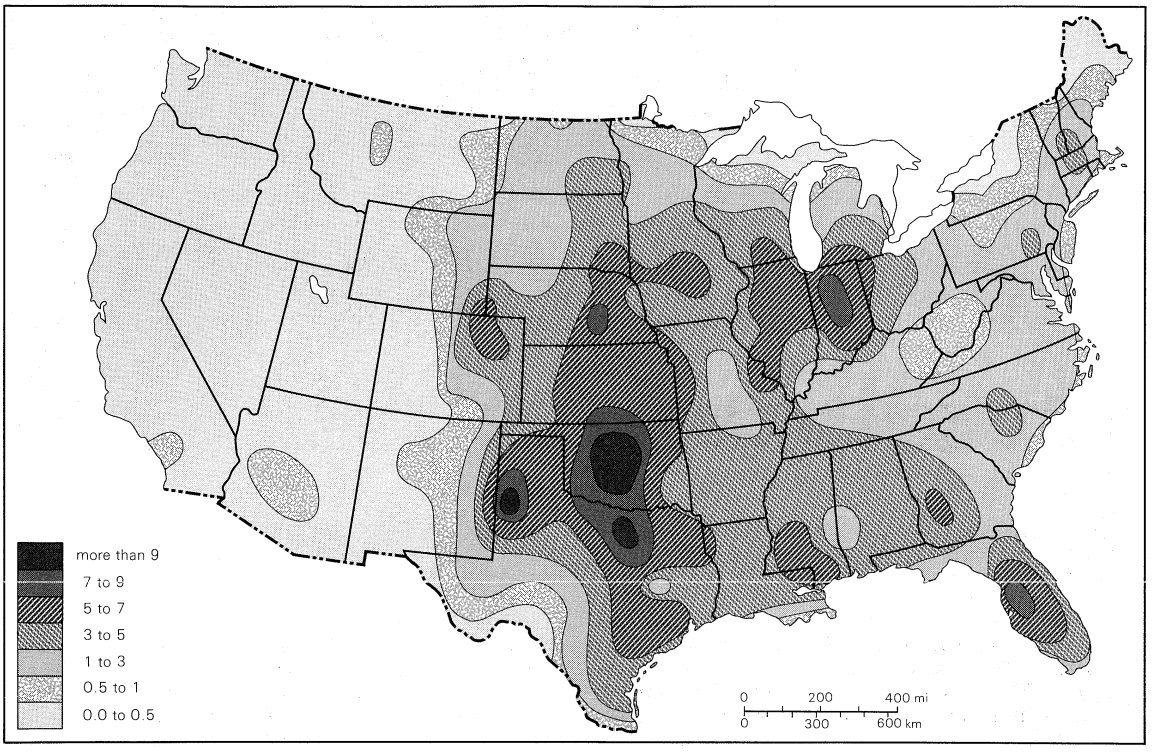



Figure 42: Annual frequency of tornadoes in the coterminous United States. (Based on 1953–80 data.)
NOAA

area, but because it is thinly populated as compared with the Chicago area, which has less than one-half its tornado incidence, casualty potential for Chicago is much greater. Although the probability that a specific locality will be struck by a tornado in any one year is very small, this low probability does not mean that such an event will not happen. An extreme example of an apparent defiance of probability statistics took place in Codell, western Kansas, a small town hit by tornadoes on the same date, May 20, in three successive years, 1916, 1917, and 1918.

Seasonal and diurnal variation

Tornadoes may occur during any month of the year. Normally, for the United States as a whole, the month with the most tornadoes is May; and more than half the year's total occurs during the three months of April, May, and June. The lowest frequency is in December and January. No season of the year is free from tornadoes, but in spring and summer they are five times as numerous as in winter and fall.

Tornadoes are generated from severe thunderstorms, which form readily between warm, moist air from the south or southwest and contrasting cool, dry air from the west or northwest. A squall line of severe local storms often develops along this boundary, and sometimes a family of tornadoes is spawned. In February, warm, moist air from the Gulf of Mexico begins to penetrate the Gulf states. As it continues its northern and eastern penetration, tornado frequencies reach their peaks during April over the southern Atlantic states, during May over the southern Plains states, and during June over the area extending from the northern plains to western New York state. In the summer and fall the decreasing contrast between the air masses results in a reduction of tornado incidence in most sections, although they may develop from the severe thunderstorms associated with unstable air masses, especially in Florida, where tornadoes are often associated with hurricanes. During December and January, cold air dominates the country, and the moisture-temperature relationships required for tornado genesis are not usually present.

Although tornadoes may strike at any hour of the day or night, they generally form during the middle or late afternoon, between 3:00 to 7:00 PM, the period most favourable for the development of severe thunderstorms from which they are bred.

**Waterspouts.** The worldwide distribution of waterspouts is difficult to determine because most of them occur over oceans and their detection depends on observations from coasts or from ships or airplanes. They are most frequent over tropical and subtropical waters; many of them have been observed over the Gulf of Mexico, off the coast of Florida and the Bahamas, and over the Gulf Stream. The spouts are common off the west coast of Africa near the Equator and off the coasts of China and Japan, but they may appear in unexpected places, such as the Grand Banks of Newfoundland and even near Seattle, Wash. As mentioned earlier, one of the largest observed waterspouts occurred off the coast of Massachusetts. A broad spout estimated to be 30 metres high and 210 metres thick appeared off the California coast in 1914, and one offshore from New South Wales, Australia, had a measured height of more than 1,500 metres.

The most common season for waterspouts in the Northern Hemisphere is between May and October, but they may appear at any time of the year or of the day or night.

**Whirlwinds.** Every summer hundreds of thousands of dust and sand devils travel across the arid and semiarid regions of the world. Only an extremely small fraction of such whirls are observed, however, because desert areas are sparsely settled. Observers report that whirlwinds are sighted almost daily during the hot season over the Sahara and the arid regions of Australia and the southwestern United States. They also are common over sections of India and the Middle East. An extensive scientific census of dust devils near Tucson, Ariz., was taken from June 23 to July 28, 1962. A total of 1,663 whirls were spotted, and the daily average was near 80.

Dust devils may, at times, develop in more temperate weather. They have even been observed as far north as the subarctic.

Diurnal frequency and duration

Dust and sand devils are usually most active over desert areas in the early afternoon, 12:30 to 2:00 PM, local standard time. These times precede by several hours those at which the maximum air temperature is recorded in a standard, shaded instrument shelter 1.5 metres above the ground. One reason for this apparent anachronism is that the much hotter sand surface attains its peak temperature earlier than the air at the 1.5-metre level.

The life of a dust or sand devil may be longer than visual observations indicate, because the vortex itself does not depend on the visible material forced into it. Durations vary from several seconds to about seven hours, but most dust devils probably last less than five minutes, and few persist for more than an hour. The larger, more vigorous whirls have a longer lifetime than the smaller vortices. One large dust devil, with a height of about 750 metres, lasted for seven hours as it traveled 64 kilometres on salt flats in western Utah. And in northwestern Mexico a large whirl formed at the end of an embankment and remained there for four hours.

## PHYSICAL CHARACTERISTICS OF THE VORTICES

Although the visible characteristics of tornadoes, waterspouts, and other atmospheric vortices are well known, details regarding the distribution of velocity, pressure, energy, and particulate matter within them are lacking.

**Tornadoes and waterspouts.** Tornadoes destroy all standard measuring instruments; hence, most values given for velocity, pressure, and energy distribution have depended on theory and engineering damage estimates. Since the 1940s, however, radar, instrumented aircraft, and photogrammetric techniques have provided some quantitative data for analysis of the vortex structure.

Velocity distribution

The first quantitative observation of the distribution of tangential and vertical speeds within a tornado was made from scaled movies of the debris and cloud fragments in a tornado in Dallas, Texas, on April 2, 1957. The tornado was fairly stable and relatively small, with a spotty damage path. The variation of the tangential (rotational) velocity with radius at the 300-metre level is shown in Figure 43, in which the tangential velocity is seen to rise steeply from zero at the centre of the vortex to 240 kilometres per hour at about 60 metres from the centre and then fall off more gradually. This type of velocity distribution closely resembles the Rankine (theoretical) combined vortex (dashed and dotted lines in the Figure), in which the inner core is assumed to rotate with a constant angular speed, designated by the Greek omega ($\Omega$) and corresponding to the rotation of a solid body. The tangential velocity is the product of the constant angular speed times the radius of the vortex. This is expressed in the equation $v_l = \Omega r_l$, in which $v_l$ indicates tangential velocity and $r_l$ the radius distance. Hence, for this particular tornado, the tangential velocity in the inner core at the 300-metre level was almost directly proportional to the radius until the maximum speed (indicated by $vm$) was reached at the corresponding radius ($rm$). At this point, $vm = \Omega rm$.

Adapted from H. Kuo, *Journal of the Atmospheric Sciences*, vol. 23, no. 1 (January 1966)

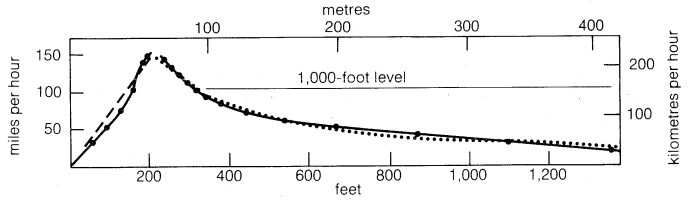



Figure 43: Variation of the tangential velocity with radius at the 300-metre level in the Dallas tornado on April 2, 1957. Dashed and dotted line represents the theoretical values; solid line represents observed values.

The tangential velocity distribution in the outer portion of this two-celled vortex follows closely the relation $v_2 r_2 = C = \text{constant}$. Hence, $v_2 = C/r_2$, or the velocity is inversely proportional to the radius. The velocity profiles for the 90- and 30-metre levels showed somewhat greater deviation from the Rankine (theoretical) vortex, possibly because of frictional effects near the ground. The peak tangential speed for this tornado was 274 kilometres per hour, at a radius of 40 metres and elevation of 69 metres.

Vertical speeds are measured by tracking debris in the air and correcting for free-fall velocities (which are not related to the tornado-derived speeds). Isotachs (lines of equal value of wind velocity) of these vertical speeds are depicted in Figure 44, in which distances are in metres and speeds in metres per second. A jet of strong vertical flow extends upward from a high-speed centre at 41 metres,

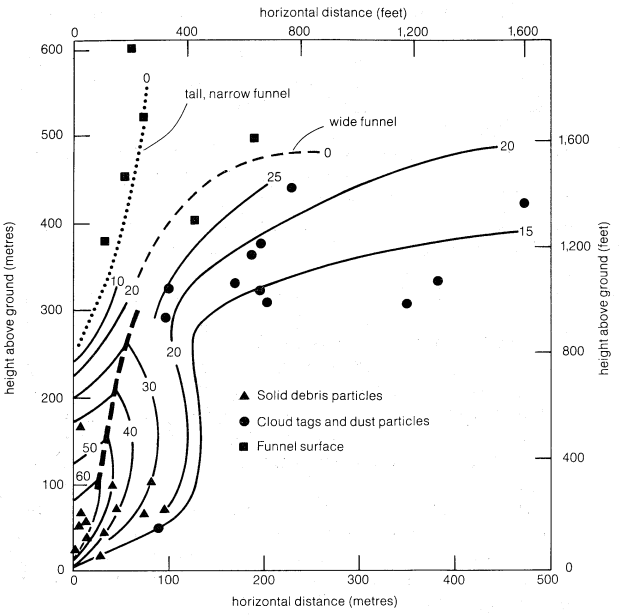



Figure 44: Distribution of the vertical wind velocities (in metres per second) in the Dallas tornado on April 2, 1957.

Adapted from H. Kuo, *Journal of the Atmospheric Sciences*, vol. 23, no. 1 (January 1966)

but the jet disappears near the 300-metre level. Above this altitude the line of zero velocity values at any one time appears to coincide with a zone of moisture condensation around the funnel. This indicates that the upper portion of the funnel probably does not act as a sink mechanism, in which air condenses at high levels and then sinks to lower levels. Whether the horizontal and vertical velocity distributions presented for the Dallas tornado are typical of most other tornadoes is uncertain.

Pressure, heat flux, and energy

No direct measurements have been made of pressures within a tornado, although barographs outside the zone of destructive winds have recorded the sudden drops of pressure associated with the tornado passage. By the use of theoretical relationships between wind and pressure and evaluation of damage patterns, however, some estimate of the distribution of pressure may be obtained.

Although standard barographs are not designed to register the rapid changes of pressure produced by a tornado, there have been cases in which barographs or aneroid barometers on the fringes of the zones of maximum winds have done so. Pressure drops of 100 millibars (or 10 kilopascals), about 10 percent of normal atmospheric pressure, are not uncommon, and a fall of 200 millibars has been reported. Because these changes may occur within an interval of less than 30 seconds, the normal pressure inside a building does not have time to adjust to the fast reduction of pressure outside, causing explosive or "suction" effects. Roofs are lifted and walls are blown outward.

Calculated values for the kinetic energy of a typical tornado vary from $10^{10}$ to $10^{13}$ joules ($10^4$ to $10^7$ kilowatt-hours). The energy input required to generate this organized motion is probably 10 to 100 times as great. For comparison, the total energy (including radiation) of the 20-kiloton nuclear bomb dropped upon Hiroshima was about $10^{13}$ joules ($10^7$ kilowatt-hours). Any theory of tornado development must explain how the potential energy of the environment is converted into the organized kinetic energy of the tornado vortex.

The lower portion of a tornado funnel often appears as a mass of dust and debris picked up by the vortex. The rim of the funnel is usually rendered visible by clouds produced by the condensation of water vapour. The inner core is almost cloudless. Heavy rains and hail are often associated with the severe thunderstorms from which tornadoes develop. Waterspouts sometimes draw fish and frogs into the vortex and then drop them onto land.

**Dust devils and "fair-weather" waterspouts.** Until quite recently few detailed measurements of horizontal and vertical velocities in dust devils had been made. Quantitative data on "fair-weather" waterspouts are still sparser, but it

Velocity distribution

is probable that the velocity distributions are similar to those of dust devils in many respects.

Recent measurements in dust devils indicate that the variation of the tangential velocity with radius corresponds closely to that of the Rankine vortex, as described earlier. Tangential speeds of 40 kilometres per hour have been measured and are probably common in moderately strong vortices. Velocities of more than 80 kilometres per hour probably occur in some of the larger, more vigorous dust devils. Sailplane pilots have measured vertical speeds of 16 kilometres per hour in moderate whirls and 32–48 kilometres per hour in stronger vortices.

Research indicates that a mature dust devil may be divided into three vertical regions. Region 1 is a shallow frictional layer near the ground. It is there that the air and dust are entrained into the core of the whirl. Above that shallow boundary layer is Region 2, a stable vortex in which the pressure force and centrifugal force are in balance. Region 3 begins where the top of the vortex becomes unstabilized and the turbulent air diffuses radially with height.

Pressure drops of a few millibars are typical in dust devils. Details regarding the pressure distribution are not known, but there is good evidence that pressure in the well-defined vortex (Region 2) follows a radial distribution similar to that of the Rankine vortex.

The main energy source for a dust or sand devil is the heat flux from the hot surface, although, under certain circumstances, dust or sand devils can be formed in a moderate surface temperature environment, provided that other necessary meteorologic conditions are fulfilled. A heat transport (flux) of 0.7 calories per square centimetre per minute was measured over a desert dry lake in California. For a circular area with a 10-metre diameter, this flux would correspond to a net upward heat transport of about 50,000 watts. A dust devil moving over a heated surface utilizes this energy to maintain itself. When dissipative forces, such as surface friction and eddy interaction with the environment, exceed the available energy, the whirlwind is destroyed.

## THEORIES OF VORTEX FORMATION: THE GENERATION OF TORNADOES, WATERSPOUTS, AND WHIRLWINDS

At present, meteorologists have not generally agreed on any particular quantitative theory for complete explanation of the formation and maintenance of tornadoes, waterspouts, and other vortices. Some of the more notable theories that seem to fit atmospheric observations and laboratory simulation experiments are reviewed in this section. Any theory of vortex formation must explain, quantitatively, how the potential energy of the environment is converted into the organized rotational motion of the vortex.

**Downward development theory.** One kind of tornado formation theory assumes that the tornado vortex develops downward from the parent thunderstorm cloud. It is known that the bases of cumulonimbus clouds associated with severe thunderstorms have stronger rotation than those of ordinary thunderstorms. Further, the visible funnel appears to work downward toward the ground. This visible aspect does not necessarily mean, however, that the tornado vortex itself must originate at upper levels, since spirally rising air currents could produce a cloud sheath resulting from condensation of water vapour in the rising air.

**Vortex-contraction theory.** During the latter half of the 20th century radar has been extensively employed in the detailed analysis of severe storms. The use of radar in this area is based on the fact that clouds and precipitation "backscatter" the emitted radiation, thus producing an "echo" that can be photographed on a radarscope. Areas of relatively clear air ("echo-free") are delineated by contrast. Radar meteorologists have contributed significantly to both theories of tornado development and the methods of warning.

The "hook" echo

A photograph of a historically significant echo associated with an Illinois tornado is reproduced in Figure 45. A projection from the mother thunderstorm lengthened and curved in a counterclockwise manner to form an echo resembling a "figure 6" or "hook." The hook itself is not the tornado, but tornadoes are often associated with echoes of this shape. An analysis of echo alignments and motions and of surface pressure and wind distribution revealed that this distribution corresponded to that of a "tornado cyclone" detected in 1948. This kind of cyclone is a small low-pressure area of about eight to 40 kilometres in cross section; the 1948 tornado funnel was 1.6 to 3.2 kilometres south of the cyclone centre, and it has been suggested that the contraction of this cyclone would result in an increase in angular momentum and a large upward motion, producing condensation within the tornado cyclone.

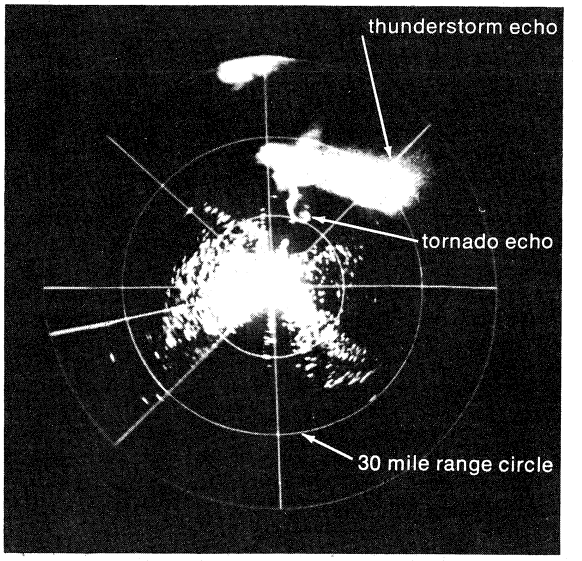



Figure 45: Hook echo of a tornado in Champaign, Ill., photographed on a radar scope on April 9, 1953. This was the first occasion on which the hook echo, an important clue in the tornado warning system, was recognized and photographed.

By courtesy of the Illinois State Water Survey, Urbana, Illinois; photograph, Donald W. Staggs

**Severe local-storm theory.** A model for severe local storms that may produce tornadoes is shown in Figure 46. It includes a strong shearing of wind (scissorlike action) veering (turning, in clockwise sense) with height. The essential features of the model are the following:

1. An updraft of warm, moist air, in which most of the streamlines (lines following flow paths) turn counterclockwise (cyclonically) through about 270 degrees.
2. A downdraft within the precipitation areas ahead of the updraft.
3. An extensive overhanging radar echo ahead of the storm.
4. A region of low reflectivity or "vault" in which there is a persistent strong updraft. The largest hail usually falls in the area of intense echo surrounding the vault.

Adapted from K.A. Browning, *Journal of the Atmospheric Sciences*, vol. 21, no. 6 (November, 1964)

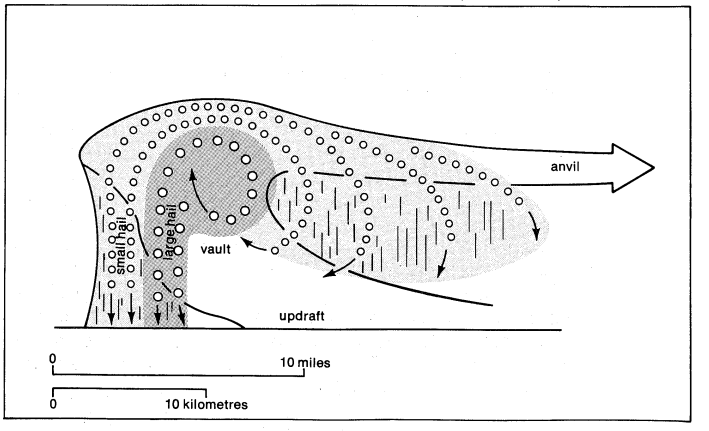



Figure 46: Vertical section illustrating precipitation trajectories in different parts of a severe storm traveling to the right of the prevailing winds in the middle troposphere.

5. A hook-shaped echo that has a definite counterclockwise circulation. When a tornado forms, it is usually near the forward edges of the hook. It is believed that the hook-shaped echo results from rotation within the strong updraft. Large hail, frequently associated with tornado-breeding storms, requires a vigorous updraft for its formation. The development of a core of rotation within the updraft appears to be required for the genesis of severe storms moving to the right of the environmental winds in the troposphere. It has been suggested that the necessary core of rotation within the updraft could develop only when the updraft was strong enough to generate the growth of giant hail.

The lightning hypothesis

**Electricity theory.** It is known that lightning, when it accompanies a tornado, has peculiar characteristics. The high-voltage discharges have been described as brighter and bluer than those from other storms. A farmer who had looked up into the core of a tornado later described the walls as constantly illuminated by lightning flashes that zigzagged from side to side. So it is not strange that lightning has been hypothesized as the source of energy for the genesis or maintenance of tornadoes. Electrical theories for tornado formation are not new, but have been revived in modern studies.

Electrical energy is assumed to be transferred to the vortex or the incipient vortex by mechanisms such as the electromagnetic acceleration of ions or heat released from lightning strokes. These electrical transfer mechanisms have been analyzed on the basis of theory and the use of simulated tornado vortex experiments. It has been concluded that the electrodynamic acceleration of ions is unimportant for either the genesis or maintenance of a tornado. Although the heat released from lightning could initiate updrafts, this lightning energy would have to persist for several minutes in order to generate a tornado. Laboratory experiments indicate that electrical discharges in the centre of an intense vortex inhibit, rather than strengthen, its circulation.

These negative results do not mean that electrical phenomena might not produce important indirect effects: the rate of release of latent heat from condensation and precipitation could be influenced by the electrodynamic acceleration of charged particles in the vortex.

**General vortex theory.** Because of the complex interrelationships of the meteorological factors that must operate together to develop a tornado, a quantitative dynamic theory for generation and maintenance of these twisters is very difficult to formulate.

A theory, nevertheless, has been formulated for the dynamics of atmospheric vortices, which applies not only to tornadoes but also to milder twisters, such as dust devils and other whirlwinds. This model is based on the preexistence of two atmospheric conditions considered to be essential for the genesis and maintenance of the vortices: (1) a thick layer of unstable stratification, either with or without the influence of the latent heat of condensation; and (2) existence of vorticity. The first factor produces sinking because vertical instability aids strong updrafts and the rising air mass is replaced by horizontally converging air. The second factor provides a source of momentum for the development of a stronger tangential velocity as the air converges into a smaller radius. Thus tornadoes and waterspouts, because they are basically convective in nature, are similar to dust devils and fire whirlwinds. For all of these vortices the ratio between horizontal extent and vertical extent (ranging from $1/5$ to $1/50$) is small in comparison to that for hurricanes and extratropical cyclones (from about 10 to 200). Because only the general mechanisms that govern the vortices are of concern, differences depending on atmospheric structural details have been disregarded and it is assumed that all of these small-scale vortices are governed by the same simple model.

Atmospheric conditions essential to the development and maintenance of vortices

From a set of equations for atmospheric vortex motions, the effective radius of the innermost cell assumed to be developing into a vortex was found to be inversely proportional to the one-third power of the initial vorticity and directly proportional to the one-third power of the product of the eddy viscosity coefficient and the effective depth of the unstable layer. When this layer is shallow, only relatively small and weak whirlwinds, such as dust devils, can be produced.

Two-cell solution

For a steady state (equilibrium) circular vortex, two solutions have been derived. The first is a two-cell type, with descending motion in the centre and ascending motion in the outer part; the second is a one-cell type. The theoretical descending motion in the centre is compatible with the observed sparseness of particulate matter in the cores of tornadoes and dust whirls.

The two-cell solution also indicates that the pressure distribution in the outer portion of the vortex is the same as the flow in the outer part of the Rankine combined vortex described earlier.

Other solutions for a steady state vortex are possible, but it appears that all of them have the following common characteristics:

1. The maximum vertical speed is proportional to a stability factor and increases linearly with height in the unstable layer.
2. The maximum tangential speed is proportional to the angular momentum and the square root of the stability factor; it is inversely proportional to the square root of the coefficient of eddy viscosity, but it is almost independent of the height.
3. The radius of the maximum tangential wind is proportional to the square root of the eddy viscosity coefficient divided by the stability factor.
4. The tangential flow outside the radius of the maximum tangential wind closely approximates the relation in which the product of velocity and radius is a constant ($vr = C$).

The theoretical results of the dynamic-model theory agree quite well with observations of tornadoes and dust devils.

(N.R.W./Ed.)

## Hurricanes and typhoons

Hurricanes and typhoons are severe tropical storms characterized by wind speeds exceeding 32 metres per second. The winds spiral inward toward the storm centre (characteristic of cyclonic flow), clockwise in the Southern Hemisphere and counterclockwise in the Northern Hemisphere. At the centre of the storm there are light breezes, even the proverbial calm. Barometric pressure decreases rapidly toward the centre, where record low pressures have been recorded. Wind speed, humidity, and rainfall increase toward the central zone, termed the eye, where they suddenly decrease. If there is a strong downdraft in the eye, temperatures there may be some 8° to 10° C higher than in the storm's main body.

Such tropical cyclones are not common. Their total number in a given year may vary from 30 to 100, with about one-quarter occurring near Southeast Asia, one-seventh in the Caribbean and adjacent waters, and one-tenth in the southwest Pacific and Australian waters. They play a noticeable, if spasmodic, role in the general circulation of the atmosphere, transporting large amounts of warm, moist air from very low to middle latitudes. It is estimated that a mature hurricane may export more than 3,500,000,000 tons of air per hour, thus contributing greatly to redistribution within the troposphere. The development of a hurricane entails the release of large amounts of energy and the transfer of substantial quantities of water over several degrees of latitude.

### ORIGINS

A tropical cyclone is likely to occur whenever several of the following prerequisites occur simultaneously: (1) latitude sufficiently high (5°–6°) for the Coriolis force (see above) to be appreciable; (2) a warm-water surface (at least 27° C) of sufficient area to supply the overlying air with large amounts of vapour; in some cases a cyclone may form over water at 23° or 24° C if much colder air is present at higher altitude; (3) pronounced instability in the air column or relatively low pressure at the surface and often an anticyclone aloft; (4) little or no vertical wind shear (shearing effect produced by the movement of one air mass past another). These conditions are most likely to occur over the oceanic areas where the intertropical convergence zone moves 10° or more away from the Equator.

Wind flow and cyclonic triggering

The Coriolis force is proportional to both the latitude and the angular velocity (rotational speed) of the Earth. At these low latitudes, the value of the Coriolis force is minimal; this is why, in large and shallow tropical depressions with very weak winds, the pattern of air flow is indistinct. The general instability and enormous vapour load make the air most susceptible to any triggering factor and especially to convergence due to external wind flow. Convection may rapidly become tumultuous. With a strengthening of convection, the centripetal wind flow gains speed; soon the angular velocity component of the Coriolis force becomes sufficient to impart a definite cyclonic curvature to the air flow, and a cyclone becomes established. The input of warm, very moist air continues. Large-scale condensation of moisture occurs during the ascent, and enormous amounts of previously latent energy are released. This energy results in stronger winds, which in turn lead to the intake and uplift of larger amounts of humid air, with a further release of energy. The evolution from tranquil tropical depression to violent tropical cyclone takes four to eight days. Strong vertical wind shear (*e.g.*, by a jet stream overhead) would impede convection and prevent the development of the cyclone. Latent heat is the main source of energy in a tropical cyclone. Thus, a rapid inflow of dry air can reduce the cyclone to a much slower tropical depression.

**Development and classification.** The life span of a tropical cyclone may vary from a few hours to nearly three weeks; most cyclones last five to 10 days. All of them begin as tropical depressions over warm oceanic waters over an area 200 to 400 kilometres across. They may remain in this formative, rather shapeless stage, with bulging but not closed isobars (lines of equal pressure) and pressure above 1,000 millibars for many days. A substantial number of tropical depressions never evolve to a cyclonic stage.

The immature stage of a tropical storm begins when the wind gains strength (from the 65 to 87 kilometres per hour associated with weak storms to the gale-force winds of 89 to 118 kilometres per hour characteristic of cyclonic storms) and follows a spiraling path toward a distinct centre. The eye measures five to 15 kilometres across. Wind and pressure gradients are very steep around it. Some isobars are closed, and pressure falls below 1,000 millibars. The diameter decreases to 80 to 200 kilometres. Many storms remain at this stage throughout their life. In some there may be occasional gusts of hurricane force.

Many investigators hold that the term tropical cyclone (hurricane, typhoon) should be reserved for the mature stage of intense tropical storms, those in which wind speed exceeds 118 kilometres per hour usually over an area of at least 100 kilometres in diameter. Isobars are very nearly circular at first, and pressure falls well below 1,000 millibars at the centre, but gradients become less steep. The size of the cyclone as shown by the closed isobars may vary considerably from just over 100 to more than 2,000 kilometres in diameter at the surface, with a 20- to 100-kilometre eye.

The cyclone may reach a decaying stage in which the pressure rises and the wind slows down while the eye becomes wider and less distinct and a revitalized stage in which the inflow of cold, moist air brings new energy and a frontal structure arises.

Classification of storms

The international classification, in use since the 1950s, is based on wind speed. Tropical cyclones are evaluated according to several objective criteria: (1) minimum pressure; (2) wind speed that may be measured over five, three, or one minutes or in single gusts; (3) wind direction; (4) rainfall quantity and intensity; (5) area or diameter of the widest closed isobar or the isobar of 1,000 millibars, or area or diameter with winds more than 87 or 118 kilometres per hour; and (6) point of origin and characteristics of the track.

**Core structure.** A record low pressure of 870 millibars was measured in the eye of a typhoon in the Guam area in October 1979. The lowest pressure on land, 892 millibars, occurred in a 1935 Florida hurricane. The pressure gradient may reach three millimetres per kilometre; pressure has been observed to fall 38 millibars in 30 minutes and 16 millibars in five minutes.

Radar observations show the dynamic core of the cyclone as a near-circular, echoless area, also revealed by the inward spiraling rainbands, 10–20 kilometres wide and 100–150 kilometres long, which converge to form a central ring. The core (eye) is quite distinct from the barometric centre, which may be up to 80 kilometres away from it on the poleward side while the cyclone travels westward and later on to the equatorward side when the cyclone heads eastward. The position of the core near the ground also differs from its position as shown by radar, because the rain that causes the radar echoes occurs at about the 700-millibar level.

**Travel and modification.** A tropical cyclone traveling over a land surface loses much energy because of increased friction, and wind speeds are reduced. Most of the cyclones that reach the high Vietnamese and Malagasy escarpments die out immediately. On the other hand, the low Western Australian plateau is no obstacle, and cyclones have been known to travel 1,500 to 1,800 kilometres over it. Hurricane "Camille" (1969) traveled 1,800 kilometres along a broad arc from Louisiana to Virginia, losing much strength but preserving its structure. The intake of much dry air, as occurs in high-pressure situations on the west coasts of Asia, North America, and Australia, fails to supply any further latent energy and may lead to the rapid decay and dissipation of the cyclone, especially if it travels over cool water. Similar developments occur on the east coasts of these continents but are less frequent because coastal waters are usually warmer and lasting high-pressure situations are less common in the hurricane season.

Mid-latitude occurrences

A most significant development occurs when a tropical cyclone happens to enter a long north–south pressure trough. The cyclone may then reach the middle latitudes without too much change or loss of energy. The intake of moist air is still possible, especially on east coasts, where warm currents (Gulf Stream, Kuroshio) are present.

When a stream of mid-latitude air is drawn into the cyclone, frontogenesis (formation of frontal structure) occurs, generally with the appearance of an extraordinarily active warm front due to the high speed of the original cyclonic circulation. This results in the new extratropical frontal structure of the revitalized cyclone becoming apparent on surface maps, while at altitude (500-millibar level and higher) the nonfrontal tropical features remain almost unchanged. At this stage three possibilities arise, namely: (1) the cyclone meets a large anticyclone and is rapidly filled, eliminating its centre; (2) the cyclone continues on its course for some distance and disappears gradually by infilling without much further change; or (3) the cyclone continues on its course until it enters a westerly stream where wind speed exceeds the wave velocity, in which case the transformation to full-scale mid-latitude depression rapidly extends upward and becomes total. For some time, however, the cyclone may remain distinctly warm and humid and relatively homogeneous, and the inner end of the frontal surface remains eccentric. In the meantime a cold front is likely to arise on the equatorward side, usually also eccentrically. Considerable amounts of moisture will condense and fall as rain. The additional energy thus released, combined with the more pronounced Coriolis force due to the higher latitude, causes the cyclone to travel at greatly increased speed. During this phase much damage may occur in coastal areas of Japan and the Atlantic United States.

## PHYSICAL CHARACTERISTICS

The shape of the streamlines in a tropical cyclone is not a simple spiral, as it would be if there were no zonal (latitude belt) easterly winds around it. Because of the greater speed around the core, the zonal easterlies are slightly deflected outward before they begin their inward spiraling course toward the centre of cyclonic indraft. Also, downwind of the cyclonic spiral there is a hyperbolic (saddle-shaped) divide between the airstreams that flow into the cyclone and those that rejoin the zonal, easterly flow after having been deflected around the edge of the cyclone. Winds are weakest near this hyperbolic divide: there is no wind at all at the point where the air may be equally likely to flow into the cyclonic indraft or to flow away from it. This

is the stagnation point, and it is a necessary element of the cyclonic structure. The corridors of convergence that cross the stagnation point may be marked by a line of clouds, which are only small cumulus on either side of the stagnation point and build up to towering cumulonimbus both eastward and westward where the air streams crowd together.

A tropical cyclone is often preceded, at a distance of about 1,000 kilometres, by a short spell of fine weather due to divergence and subsidence in the prevailing (and preceding) easterly air flow. Pressure then begins to fall, but cloudiness does not appear until the cyclone is fewer than 300 to 500 kilometres away.

**Wind velocity.** Wind gusts fluctuate rapidly from 37 or 56 kilometres per hour to more than 185 kilometres per hour, even within two or three minutes, but there is no rhythm or regularity. Strong winds often are associated with rain bands. The highest gust velocities, recorded around the eye, may be close to 260 kilometres per hour. The strongest gust recorded (1966) reached 316 kilometres per hour. Within the eye, calms or light breezes prevail. Beyond the eye, the strongest gusts resume from the opposite direction.

The actual wind velocity is the result of two distinct movements—namely, the inward spiraling of the air streams and the progress of the cyclone as a whole along its path. The zonal wind strengthens or weakens the wind field, thereby extending or reducing the area of strong winds and making it asymmetrical. Before its recurvature, a cyclone lies in the trade-wind belt (zonal easterlies); its poleward side may already have hurricane winds 100 to 130 kilometres from the centre because the trade winds add momentum to that side of the spiraling flow. Conversely, the storm area is greatly reduced on the equatorward side. The speed and direction of travel of the cyclone and the strength and direction of the trade winds are, however, always minor elements in cyclonic wind patterns, the spiraling winds being paramount.

Onset of a tropical storm

**Clouds and precipitation.** The usual herald of a tropical storm is a broad trail of cirrus. Heavy showers of rain may fall on a coastline 500 to 600 kilometres ahead of the oncoming storm. Cirri depart in all directions at high altitude. A thin watery veil may pervade the air. Within a day or two, the storm appears as a distant wall of cumulonimbus. The middle and high clouds travel along divergent paths. The lower the clouds, the less direct the exit. Altocumulus may diverge by 20° from the tangents to the isobars. Low clouds, which may travel 50 metres above the ground, tend to follow a converging course. The wind freshens and rain begins to fall. Cloudiness suddenly becomes overwhelming. The wind becomes more violent and rain showers heavier. The heaviest rainfall occurs patchily just before the eye, in a radius of 20–50 kilometres, until one side of the wall passes over. There follows a 10–45-minute rainless interval with calm or light breezes while the eye passes overhead. Then follow the heaviest downpours and strongest gusts from the opposite direction, which occur beneath the steepest towering portion of the cloud wall. In coastal locations much sea spray is added to the rain. This second rainy spell is also likely to last longer, extending radially over a further 20–30 kilometres. The rain gradually decreases but may still last two or three days if there is enough moist air. The increase or decrease of rainfall intensity is normally decided by streamline convergence (confluence of air flow) or divergence, hence the spiraling bands of intense rain. Rain gauges may miss up to 50 percent of the rain because of the driving wind, and they may overflow before readings are taken. In some cases all instruments have been blown away. It is certain, however, that cyclones have brought the majority of the heaviest rainfalls recorded (*e.g.*, 1,583 millimetres in 24 hours in April 1958 to Aurère, on a mountain slope on Réunion Island, and 116.8 centimetres on a hillside at Baguio, Phil., in 1911).

When the hurricane's structure becomes revitalized, a renewal of precipitation occurs. The new rainfall zone is more distinctly asymmetrical with respect to the hurricane's track. The rain may reach 150 or even 200 millimetres in 24 hours at the most active point of the front, but in general it is 50–60 millimetres over some 20,000–30,000 square kilometres, enough to cause local flooding in some situations.

**Energy of storms.** The power of a typhoon $K$ in ergs per second may be derived from the empirical formula relating power to the lowest pressure of the storm and the radius of the largest circular isobar, namely:

$$K = 0.71 \times 10^{22}(1{,}010 - P_{min})\,(R/111)^2,$$

in which $P_{min}$ is the lowest pressure in the typhoon and $R$ is the radius of the largest near-circular isobar in kilometres. The energy that is associated with very severe storms is about $10^{25}$ ergs per second by this formulation.

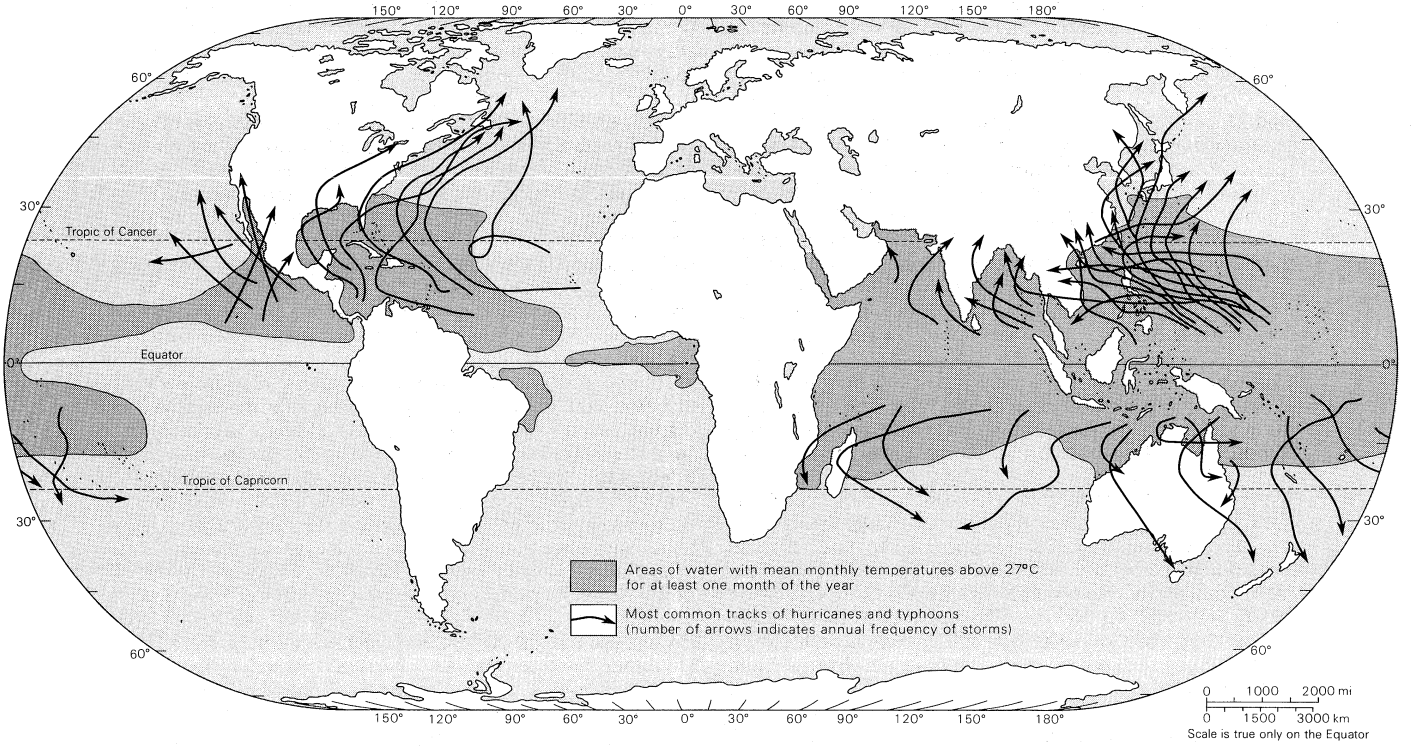



Figure 47: Major tracks and frequency of hurricanes and typhoons.

DISTRIBUTION

The average distribution and frequency of hurricanes and typhoons around the world are shown in Figure 47. More details appear in Table 5, which gives the frequency by area or region. Because of the cool surface waters, tropical cyclones do not originate in the eastern South Pacific and South Atlantic oceans (Figure 47).

**Table 5: Average Frequency of Cyclones by Area**

| area | average frequency per year | total |
|---|---|---|
| East and south Asia | | 29 |
| Northwest Pacific | 21 | |
| Bay of Bengal | 6 | |
| Arabian Sea | 2 | |
| North and Central America | | 14 |
| Caribbean and northwest Atlantic | 8 | |
| East Pacific and west Mexico | 6 | |
| Australia and Oceania | | 9 |
| Southwest Pacific and northeast Australia | 6 | |
| Timor Sea and west Australia | 3 | |
| Southwest Africa and islands | | 7 |
| Southwest Indian Ocean and Madagascar | 7 | |
| World | | 59 |

**Seasonality and frequency.** Before 1940 many cyclones went undetected. At present every disturbance is detected by satellite photography. There are four main areas of origin of North Atlantic hurricanes: south of the Cape Verde Islands, the open waters east of the Lesser Antilles, the western Caribbean Sea, and the Gulf of Mexico. The early summer hurricanes usually begin in the Gulf of Mexico and western Caribbean, where water temperatures rise early because of the shallows and the nearby heated land. Low pressure farther north often allows the hurricane to enter the Gulf area. As the broad expanses of Atlantic surface waters warm to 28° C or more in late July and August, more hurricanes form much farther east, as far as the Cape Verde Islands. They gather more energy as they travel west before recurving over the West Indies. By mid-September surface waters are cooler, but outbursts of cold air aloft are likely; the area of origin moves west again to the Caribbean and to the Gulf of Mexico. The high-pressure outflow of continental air often pushes the recurving hurricane along a more easterly path so that it may sweep over Florida or the Atlantic seaboard.

Migration of tropical storms

**Cyclonic tracks.** Whereas a tropical depression may remain almost stationary, a tropical cyclone becomes smaller and more intense and begins to travel along a westward path, slowly at first, then at about seven to 28 kilometres per hour. The path gradually recurves toward the tropic, and the speed of travel decreases, in some cases practically to a standstill. The westernmost point is reached at latitude 15°–30° when the path gradually recurves toward the east and the speed of travel increases to 28–46 kilometres per hour and more. Hurricane "Carol" of 1954 hit New England after having traveled 650 kilometres in little more than 12 hours. Although tracks have often been described as parabolic, this is sometimes true only of their tropical part before recurvature. After recurvature the track may be almost straight or broadly waving. Some tracks continue westward and never recurve. Loops in either sense are uncommon but may occur any time because of outside pressure influences. To the east of Australia where there is a regular procession of anticyclones, during a period of 14 years there were nine single cyclonic and four single anticyclonic loops, four double loops and one triple loop among 93 cyclone tracks, and one case (January 1956) of two cyclones merging.

Some of the mid-latitude depressions that reach the British Isles from across the Atlantic have their origin as hurricanes east of the Caribbean. The last and still fearsome winds of a former Caribbean hurricane were experienced by the Norwegian explorer Fridtjof Nansen on Sept. 5–6, 1888, on the Greenland ice cap. The Galveston, Texas, hurricane of 1900 was still a recognizable storm after it recurved across the Atlantic, over Europe, and into Siberia.

Hurricanes maintain their enormous strength along the coast of North America, probably because of the latent energy supplied by the warm current waters. Even within the same region their paths may vary: New York City suffered badly in September 1821 while Boston was immune, but Boston was damaged in 1869 while New York City was untouched.

Tropical cyclones passing within 700–800 kilometres of the Pacific coast of Mexico send violent southerly winds against the shore, where they are known as cordonazos. On the average there are six such cyclones a year, mainly in the summer and early autumn, with a peak of two in September. The average duration recorded is one to eight days per storm. Most of these cyclones travel parallel to the coast for some distance and then recurve onshore before reaching the tropic. A few (fewer than one a year) continue northward without recurving and may then affect Baja California or, very rarely, southern California north of San Diego. Their traveling speed varies from nearly 13 kilometres per hour in June to more than 20 kilometres per hour in August and falls rapidly by September. A high-pressure situation over the western United States tends to hold a tropical cyclone farther offshore, and the resulting greater pressure gradient causes strong, hot winds. A pressure trough over the western United States induces a more rapid recurvature.

Cordonazos

An index of relative cyclonicity has been devised, based on the formula relating cyclonicity to the number of closed isobars and number of cyclones, namely:

$$j = \frac{1}{n} \Sigma a_i^2,$$

in which $a_i$ is the number of closed isobars in the cyclone and $n$ the number of cyclones in a given latitudinal zone.

DAMAGE CAUSED BY TROPICAL CYCLONES

Tropical cyclones can cause immense damage, both directly (by wind, pressure, and rain) and indirectly (mainly through storm surges and floods).

The wind causes damage that generally increases in proportion to the square of its velocity, according to the basic formula $P = KV^2$, in which $P$ is the pressure exerted by the wind against vertical surfaces, $V$ is the wind velocity, and $K$ is a factor depending on air density and drag. A wind of 74–93 kilometres per hour strips leaves and small branches off trees; at 111–130 kilometres per hour it can topple shallow rooted trees or snap weaker trees outright, blow down thin walls, shift roofing materials, and occasionally lift a whole roof. A wind of this force may blow in large glass windows. At about 130 kilometres, lifting of roofs and snapping of trees is general. Hurricane winds may exert a pressure of more than 400 kilograms per square metre (82 pounds per square foot) on tall structures and can flatten weak buildings at first impact. Gusty winds, combined with a suitable period of vibration of a given structure, have a dynamic effect because they may produce resonance (vibrations in phase), leading to breaking point. Damage also is caused to roofs and windows by the suction produced by strong winds on the downwind side, amounting to 0.5–1 times the windward pressure. As with tornadoes, whirlwinds, and waterspouts (see above), much direct damage is caused by the rapid fall of external pressures. The pressure differential could amount to 300–400 kilograms per square metre for a sealed structure. Damage may be increased by the fact that the strongest gusts, in excess of 185 kilometres per hour, occur immediately after the transit of the eye and blow in the contrary direction to the preceding gusts, thus adding considerable stress to any structures exposed to them. Loose objects lifted by the wind become missiles that shatter glass, batter walls, and flatten roofs. Wind causes injuries and deaths by toppling structures and hurling loose or torn objects about with enormous force. Since the eye may take from a few minutes to an hour or more to pass over a given point, depending on (1) the central or eccentric position of that point, (2) the size of the eye itself, and (3) the traveling speed of the cyclone, many victims are struck after they have left their shelters in the belief that the storm was over. The kinetic energy of the whole cyclonic system is nearly proportional to its power, but the amount of damage and loss of life depends on many other factors, among which lack of warning and insufficient prepared-

Wind and pressure damage

ness can be extremely significant. Furthermore, the torrential rain brought by a typhoon may erode the soil, causing landslides in mountain country and making streams and reservoirs overflow.

Storm surges

Indirect damages are mostly due to a storm surge. This is a complex surface deformation of the sea induced by the cyclonic winds on coastal waters, which surge as a sudden tide against the coast, flooding the countryside and impeding the flow of rivers. The level of the sea is raised by up to three metres for a period that may last several hours, depending on the characteristics and relative position of the cyclone and of the coastline affected. The level of the sea is raised an additional 0.5–1 metre by the low atmospheric pressure. Extreme tides recorded on the Gulf Coast were all due to hurricanes. Hurricane "Hazel" (1954) brought a 3.5-metre tide to North Carolina. The worst storm surge in Tokyo Bay (October 1917) rose 2.3 metres above normal. Ōsaka Bay had nine storm surges in half a century, one of them (Sept. 21, 1934) reaching 3.1 metres above normal. Most coastal cities are less than three metres above sea level and may thus be extensively flooded. Industrial plants in coastal areas may be badly damaged by salt spray and seawater. Pounding waves are an additional cause of damage to coastal installations and structures. Coastal erosion may reach catastrophic proportions. Most significant may be the cumulative effect of a close succession of two or three cyclones, as happened in North America in 1954 and 1955. Lakes are affected in the same way as the sea but over shorter periods. Because of their smaller dimensions, they may develop storm-surge seiches (oscillations) of remarkable amplitude; *e.g.*, the disastrous 5.5-metre swelling of the southern end of Lake Okeechobee in Florida in 1926.

### WARNING AND TRACKING OF TROPICAL CYCLONES

Because of the small diurnal pressure changes common in the low latitudes, it is difficult to single out changes due to approaching cyclones. A local drop in pressure of more than three to four millibars in 24 hours (or over a distance of 500 kilometres or less) may be considered a danger sign. The development of markedly curved isobars moving toward the observer is a much more definite indication of an approaching cyclone, and the roughly circular shape of the isobars soon reveals the position of the cyclone's centre. A rapid increase in the height of any thermal inversion, as revealed by misty or hazy layers aloft, is often a sign of an approaching cyclone, as is reversal of the wind direction in altitude. Humidity is normally high and is no useful indicator, but cloud types and movements—*e.g.*, a sheet of cirrus or bright red sunsets—may provide a belated warning.

Indications of approaching cyclones

Clearer advance warning may be given by a long ocean swell with a slower frequency (two to four times the usual interval) than that of normal waves. This swell travels outward from the centre for hundreds of kilometres.

In Australia a cyclone warning station was set up on Willis Island, off the Queensland coast, in 1921. Special hurricane forecast centres have been maintained by the United States Weather Bureau since 1935.

If a ship reports strong winds or rapidly falling pressure, unusual squall activity, or even a wind flow unusual for the season, special observations at three-hour intervals are requested from all ships in the area, and warnings are issued. Because the shape of the cyclone's path mainly depends on the pressure pattern of the time (the cyclone tends to travel toward any locus of lower pressure and around or away from any locus of high pressure), a synoptic barometric map may allow a gross forecast of a cyclone's track. Pioneer flights into a hurricane took place in 1943. Regular aerial hurricane patrols began in 1945, reporting location, characteristics, and movements of any likely disturbance. Flights may be made into, through, and above the storm. The easiest access to the eye of a cyclone is from the side of the stagnation point, where the hyperbolic divide is located. Because this is essentially an area of very light winds, exact location is very difficult without the use of drift smoke signals. The smoke plumes soon reveal the direction of the air flow. The plume extending toward the lower pressure will follow the indraft-accosting spiral, which leads in the shortest time (but not as the shortest route) to the centre of indraft. A distinct, clear break along a line of growing cumulus to cumulonimbus clouds is likely to be due to the prevailing stillness at the stagnation point and may be taken as an indicator.

Tracking activities

Radar units provide warning of any storm within range. The characteristic spiral banding of dense clouds and rain makes cyclonic storms easy to identify. Satellites transmit photographs of any part of the Earth and its cloud systems. They give the most reliable and comprehensive coverage of cloud patterns and reveal storm systems from remote areas, where other methods of detection may not always penetrate. (J.G./Ed.)

# CLIMATIC VARIATIONS AND CHANGE

The meteorologist's concept of climate is a dynamic one, including day-to-day changes of weather, the seasonal cycle, and small-scale variations of atmospheric conditions measurable over periods of two to 25 years. Long series of observations show something like a "beat" phenomenon, with an unevenly varying wavelength. This built-in variability of the atmosphere results from the varying time lag between any single cause and effect, from the interaction of multiple factors, and from mechanisms set in progress by one or more variables operating over different time scales. Climate is never stable but is subject to continuing oscillations, such as the waxing and waning of many atmospheric components over a variable period of 23 to 29 months. For these reasons it is necessary to define climatic variations of different wavelength and amplitude and to decide which qualify as short- or long-term fluctuations and which are part of the built-in, year-to-year variability and thus deserving of the rank of oscillations. (K.W.B.)

## Seasonal cycle

Each year the Earth experiences changes in its radiation budget, and in the distribution of heat input into the atmosphere, in the course of the seasonal migration of the zenith Sun between 23° N and 23° S. These are far greater than any changes associated with the differences of one year from another and also are greater than the differences between different climatic epochs, except those resulting from the presence of extensive ice sheets during the ice ages. The seasonal changes are accompanied by a general northward and southward movement of the subtropical anticyclone belts and other main features of the atmospheric circulation, including expansion of the circumpolar vortex of upper winds over the hemisphere where it is winter and contraction of the vortex over the summer hemisphere. The seasonal migration of these circulation features, however, is on the average no more than 10° of latitude (and generally less than this in the Southern Hemisphere); *i.e.*, much less than that of the zenith Sun.

As a result of the seasonal migration north and south of the respective belts of cyclonic activity over the middle and high latitudes of each hemisphere and of the equatorial rain zone, many places quite regularly experience a wet season and a dry season each year. Near the Equator there are two wet seasons, as the rain belt associated with the zone of convergence between the surface winds from the two hemispheres moves north and south between its extreme positions. Over the continents the seasonal migration of this intertropical convergence zone proceeds farther into the summer hemisphere than it does over the oceans, presumably because the heating of the continents in summer weakens the subtropical anticyclone development, and the trade winds are correspondingly weak at that season over the continental sectors. The extreme case is Asia, where, in summer, the southwest monsoon winds, supplied from the Southern Hemisphere (where it is win-

Wet and dry seasons

ter), drive what is essentially the equatorial rain belt to latitude 30° N and the foot of the Himalayas. In the northern winter the Indian subcontinent is largely dominated by the northeasterly trade winds, and dry weather on the whole prevails except where these winds blow against the coast and mountains.

The highest and lowest temperatures of the year generally occur away from the main belts of disturbed weather and in conditions that allow periods of clear sky.

The seasonal migration of the wind zones and belts of cloud, rainfall, and cyclonic activity is not a smooth or continuous progression but occurs in stages. It is the end product of an alternating sequence of pulselike advances and retreats. Occasionally the successive pulses occur at nearly uniform time intervals, which presumably correspond to the common duration of some cyclic process in the atmosphere or involving atmosphere and ocean. The physical nature and origins of these cycles are not yet adequately accounted for in spite of much research. Some may represent no more than the normal life history of an individual anticyclone cell from its first appearance to its decay or rejuvenation. Others seem to correspond to a repeating cycle affecting much of one hemisphere, with intense development of the mid-latitudes westerlies, at one phase, followed by breakdown and the appearance of blocking or extended meridional (north–south) circulation cells. It is unusual for any of these sequences, however, to recur more than a few times before their regularity is lost. Their occurrence is prominent enough, however, to produce many statistical traces of preferred periods of around five, seven, and (especially) 30 days.

The 30-day oscillation

The 30-day oscillation is often expressed in a recurrent tendency toward periods of low barometric pressure, cyclonic influence, and stormy weather spreading in over Europe from the Atlantic about the end of each month, particularly from October to March, and toward anticyclonic influence culminating about the 15th to the 20th of each month. Calendars that are based on weather phenomena related to such cyclic oscillations have been produced through the years. The tendency for precision (or regularity) of the dates of these phenomena seems greatest in high latitudes in the sectors of the Earth affected and may indicate an origin associated with the onset of the polar winter darkness.

One other period length, the very curious one of about 26 months, is strongly marked in a monsoonal alternation between prevailing west and east winds in the equatorial stratosphere, and traces of it are found in the statistics of so many surface weather conditions in both low and high latitudes that it must be accepted as affecting also the lower atmosphere over the whole Earth. This also appears to be the time scale of the so-called Southern Oscillation. Like the other circulation cycles here mentioned, it is not altogether regular in its performance. The cycle length varies at least from 22 to 35 months, and as every third or fourth cycle is more prominent than the others, it also gives rise to an appearance of some roughly seven-year repetitions. At other times, it may cease to be discernible at all. Its occurrences presumably bear witness to an incomplete return of the oceanic and atmospheric heat economy to its initial condition in certain individual years. (H.H.L.)

## Climatic change

Researchers are able to approach the study of climatic change with some confidence because of two main technological changes: a revolution in observational technique that has made possible continuous scrutiny of the entire depth of the atmosphere and the surface features of the land and ocean; and the computational and data-storage capabilities associated with modern electronics. Only in recent decades, however, has the observation of climate approached the level where such sophistication is possible. During the previous 100 years or so, instrumental records were usually limited to temperature (air and sea surface), pressure, and precipitation (largely over land). Prior to that, instrumental records were fragmentary at best and are available only from about 1700. For the millennium or two before the beginning of the modern era, literary, historical, and archaeological sources provide some data; but discussion of climate and climatic change in the more remote past depends on "proxy" evidence. Most such evidence arises from the ability to interpret the world's sedimentary rocks—those deposited by still or running waters, by the winds, or by moving or stationary ice, both grounded and afloat. Annually accumulated tree growth (tree rings) has been a further source. Carefully extracted cores from trees, bogs, lakes, glaciers, and the ocean floors have provided an intelligible record for the climates of the past 750,000 years, as well as a growing body of evidence for climates that date back much further.

Proxy evidence of climatic change

The usefulness of these proxy sources would be limited, however, without the development of accurate dating and calibration methods. Radiocarbon dating of organic remains, for example, offers reasonable precision back 40,000 years or so. The decay of other radioactive (unstable) isotopes yields dates back to the early history of the planet. Stable isotope chemistry, chiefly using deuterium (heavy hydrogen, or hydrogen-2), oxygen-16 and oxygen-18, and carbon-12 and carbon-13, does not provide dating information, but it does provide evidence of temperature changes, ice volume, and, by inference, ocean circulation. One recent dating technique is based on the study of reversals of the polarity of the Earth's magnetic field. Advances in isotope chemistry have also provided vital tools.

The picture that emerges from the evidence yielded by these relatively new proxy and dating techniques is of a fairly stable global climate. Nevertheless, there have been prolonged periods, of which the present is one, in which large changes have occurred. Dynamically, both ocean and atmosphere have many of the characteristics of chaotic systems. They obey firmly established dynamic and thermodynamic laws but do so in a nonlinear fashion that permits seemingly unlimited variability. Thus, there are meaningful patterns to be described in this account of climate and climatic change, but the chaotic element tends to obscure these regularities on any short-term view. In spite of, for example, virtually constant input from the Sun, only slowly changing atmospheric composition, and an even more slowly changing Earth surface, the daily weather pattern never quite repeats itself. Each season in each locality places bounds within which this variety can unfold; but within these constraints the climate is free to choose a very diverse set of realizations.

The aim of the study of climatic change is thus to detect the signal of substantive change behind a deafening clamour of short-term noise.

### PRE-PLEISTOCENE CLIMATIC CHANGE

The present is an abnormal time for the climates of the world. Ice sheets cover Greenland and Antarctica, and permanent pack ice blocks the Arctic Ocean. The deep waters of the oceans are frigidly cold in all latitudes (mostly below 3° C). The vegetation map is strongly zonal, with treeless tundra on high-latitude land surfaces and luxuriant rain forest near the equator. It is now known that few of these characteristics were typical of most of the Earth's history, and some generalizations can be made about ancient climatic conditions, but detailed information is still fragmentary.

To reconstruct the Earth's climate in times prior to the Pleistocene epoch (that is, more than 1,600,000 years ago), it is necessary to examine the geologic record. The fossil content of the Earth's sedimentary rocks speaks for the most part of warmer climates than that of today. There appears always to have been a warm tropical belt, with twin cooler poles. Even near the poles, conditions were at most times temperate and apparently ice-free. Geologists once referred to these benign conditions as the "normal climate" of geologic times, but as early as the late 19th century it had become apparent that this rule had to allow exceptions. Nevertheless, the view persisted, and still persists, that the Earth at most times lacked snow and ice and had no extreme cold in the polar regions.

**Some governing conditions.** In one respect there is clear evidence that climate has not changed radically for more than 3,000,000,000 years. Metamorphic rocks of this age

in the Precambrian shields appear to contain metasediments that were originally deposited by torrential streams. Running or standing water appears to have been present on the planetary surface in all subsequent epochs. If so, and assuming that atmospheric pressure and the physics of water have not changed, temperatures in the range of 0° to 100° C must have pertained. Since, moreover, simple forms of life evolved in the oceans of early Precambrian times, a more realistic estimate of temperatures may be 0° to 50° C. Present-day temperatures at the surface average 15° C.

Nevertheless, early climates cannot be compared to modern climate with any confidence because other external controls must have been very different. For example,

Variables affecting climate

(1) The early atmosphere was different in composition from that of today. In particular it lacked oxygen and hence ozone. Powerful ultraviolet radiation reached the Earth's surface, inhibiting terrestrial life and guaranteeing an extremely active chemistry in the lower atmosphere. Free oxygen and ozone entered the system later in Precambrian times, as life in the oceans evolved; but the land surfaces were largely naked. This condition held until about 440,000,000 years ago (the Late Ordovician epoch).

(2) The distribution of land and sea was radically different. Vast crustal changes have occurred, such that the physical geography of the planet came to resemble that of the modern Earth only in the most recent 66,000,000 years (the Cenozoic era).

(3) The positions of the geographic and magnetic poles have both shifted relative to the distribution of land masses (or vice versa) throughout geologic history. Thus the south geographic pole migrated into the Antarctic continent more than 100,000,000 years ago, approaching its present position about 70,000,000 years ago. In addition, the magnetic polarity of the Earth has repeatedly reversed itself (a phenomenon that has vital importance for the climate of the upper atmosphere and is also useful in the dating process).

(4) The rate of spin of the Earth about its polar axis, which determines the length of day, has decreased from 425 days per year about 500,000,000 years ago to 365¼ days per year today. Hence days have become longer. This rate of spin determines the Coriolis parameter (see above *Wind*), one of the key controls of the circulation of atmosphere and ocean.

(5) The Sun itself, the source of virtually all the energy in the atmosphere, has gradually become hotter, so that the solar constant (the power per square metre of the incoming solar radiation) has increased throughout geologic history. Heat released from the Earth's interior (very small by comparison) has decreased during this period.

Because these are all factors that control or influence modern climate, ancient climates cannot be expected to have resembled modern climate. Moreover, because these variables have unknown values, reconstruction of ancient climates must rely not on theoretical deduction but on empirical evidence from the geologic record.

**The sedimentary record.** Deductions can be made about the climatic significance of certain classes of sedimentary rock. These include calcareous rocks (limestones, marbles, marls), arenaceous rocks (sands, sandstones, quartzites, conglomerates), carbonaceous rocks (coals, oil- and gas-bearing layers), red beds (assemblages of red and yellow sandstones, marls, clays, pebble and boulder beds), evaporites (salt, potash, gypsum, certain classes of limestone), and glacial deposits (morainic materials, including diamictites, tills, and outwash fans). Inferences about the climatic meaning of such rocks depend on the lithology of the materials, the known association of such materials with present-day climates, and the fossil biota they contain, if any.

*Calcareous rocks.* Deposits of calcareous rocks are widespread in the sediments of the entire geologic record. They occur as much-altered limestones in the Precambrian record, where they sometimes appear as crystalline marbles. They are also common in most later periods but are especially widespread in the Carboniferous (360,000,000 to 286,000,000 years ago), Jurassic (208,000,000 to 144,000,000 years ago), and Cretaceous (144,000,000 to 66,400,000 years ago) periods, often in richly fossiliferous forms. In the modern world calcareous sediments appear to form primarily in warm temperate or tropical seas, though foraminiferal oozes (formed from empty calcareous shells) occur even in seas in high latitudes. This modern preference for lower latitude sites, together with the richness of the marine macrofossil record, which is dominated by species typical of warm seas, leads to the view that calcareous rocks speak of warm conditions.

This view is reinforced if, as is often the case, the fossils present include corals. Corals are today largely confined to subtropical shorelines, typically in the 20°–25° C sea temperature range (with temperatures below 18° C fatal to some of the reef builders), and to waters with a salinity of 2.7 to 4.0 percent. Coralline limestones from the late Paleozoic era (*c.* 350,000,000 to 245,000,000 years ago) and the Mesozoic era (245,000,000 to 66,400,000 years ago) are found today in high-latitude outcrops. Thus, even when corrections for polar shift and continental drift are made, it is apparent that warm temperate conditions must have been present in high middle latitudes of both hemispheres during those times.

Climatic implications of coral growth

*Arenaceous rocks.* Arenaceous rocks are rich in silica and are often made up of fragments of preexisting rocks, implying the breakdown of continental surfaces by processes of rapid denudation. They are widespread in the sedimentary record. These sands, sandstones, and quartzites (the indurated form of sandstone) are usually deposited near shorelines, where their presence says little about climate. However, they may also be associated with specific bedding features that argue for torrential runoffs (deltaic structures, such as the Catskills Delta in the northeastern United States, are one example). Gravel sheets and boulder beds may point up the vigour of the erosion. In a few cases, it has been possible to estimate rates of erosion and from these to make inferences as to temperatures and precipitation. On the whole, however, the sandy rocks of the continents have not revealed much about the climatic history of the Earth.

*Carbonaceous rocks.* In contrast to the arenaceous rocks, carbonaceous rocks speak volumes, for they are derived from abundant living remains. The most dramatic evidence comes from the widespread occurrence of coal seams in rocks dating from the Carboniferous period onward. These coal deposits clearly represent the remains of forests, chiefly in wetland habitats (where anaerobic conditions prevented oxidation of the wastes). In many cases, the genera and even species of the dominants can be determined from macroscopic fossil contents. Land vegetation first became significant in the Devonian period (408,000,000 to 360,000,000 years ago), but it was in the Late Carboniferous epoch (320,000,000 to 286,000,000 years ago) that the first great development of wetland forests occurred. Forests of gymnosperms (*e.g.,* cycads, conifers, and seed ferns) flourished in the swamplands of coastal plains and in the floodplains of great rivers. Again the evidence is in favour of warm, humid conditions, even in relatively high latitudes. The drier land surfaces appear to have remained naked, however. The coal measures of many regions are intercalations of coal seams within sandy or silty materials washed into the wetlands off barren land surfaces during periods of marine transgression.

Climatic implications of coal deposits

Coal deposits occur in Mesozoic and even Tertiary (66,400,000 to 1,600,000 years ago) rocks, but they contain species that suggest cooler conditions. In the modern world, the most extensive swamp areas in which carbonaceous materials accumulate are in middle and high latitudes. It is therefore unsafe to assume that Mesozoic and Tertiary coals and lignites imply warm climates.

The widespread occurrence of petroleum and natural gas in sediments cannot readily be interpreted in climatic terms, because they appear to have originated primarily from unspecified marine biota. Their abundance is consistent with the view that ancient seas and oceans were biologically productive, and hence probably warm and nutrient-rich; but this consistency does little more than reinforce conclusions drawn from other sources.

*Red beds.* On all continents, and from many geologic epochs, widespread accumulations of red and yellow rocks

can be found. These red beds are equivocal in the evidence they yield. Massive red quartzites of Precambrian age and similar deposits from the Devonian and Carboniferous periods, from the Triassic period (245,000,000 to 208,000,000 years ago), and even from the Tertiary, all appear rich in hydrated iron oxides (accounting for the colour); they sometimes, especially the ancient rocks, also contain partially weathered feldspars. Because of their false bedding, it is clear that some of the red beds resulted from wind deposits; others appear to be associated with alluvial fans (fan-shaped accumulations of sediment deposited by streams draining from mountains to adjacent plains) or with the desiccation of former lakes. Whether deposited by wind or water, these beds appear to have accumulated in inland basins, with no discharge to the sea, as in modern desert areas. Thus, it is often assumed that red-bed assemblages imply warm desert conditions. The sparse fossil record (totally absent in the older sediments) usually gives no firm clue as to the climate. On dynamic grounds, it is reasonably certain that deserts have always existed in continental subtropical and tropical areas, so that the abundance of wind-deposited materials and the lack of drainage to the sea are unsurprising.

Climatic implications of duricrusts

In a different category are the duricrusts: hardened, dehydrated crusts formed at or near the Earth's surface. Duricrusts with high proportions of iron oxide can be found in Australia and sub-Saharan Africa. Many parts of these continents have remained above sea level for enormous periods, and the accumulated weathering layer, known as the regolith, may be many metres thick. Some of these ancient regoliths are hard enough to have resisted deflation and sheet erosion. Some appear to be laterites, layers of soil with a porous, claylike texture. Products of rock decay, laterites are formed under warm, humid conditions. Ancient laterites are mostly of Tertiary date, but some may have originated in the Triassic. In some areas they occur where laterization of the weathered material is no longer possible, providing evidence of climatic change. In other regoliths—even in rocks of Precambrian date—it is possible to detect buried layers of soil, or paleosoils, whose composition and structure may reveal climatic inferences.

*Evaporites.* Thick accumulations of common salt, potash, gypsum, and certain kinds of limestone indicate that rapid evaporation off marine lagoon or outlet-free lake basins has occurred. Such materials accumulate today on desert shorelines and in inland basins (*e.g.,* Great Salt Lake, Aral Sea, Caspian Sea, Lake Eyre). Their presence in ancient sediments points clearly to dry, warm conditions. They are especially common in Triassic sediments but have been deposited in many epochs in differing areas.

*Glacial deposits.* Because glaciation is still in progress and because there are very recently deglaciated surfaces, chiefly in North America and Eurasia, geologists have a detailed knowledge of the lithology, morphology, and genesis of typical glacial deposits. These include till, or ground moraine (unsorted assemblages of pebbles and boulders that are typically unrounded by water transport and are embedded in a clay or silt matrix); various fluvioglacial sands and diamictites; morainic materials; occasionally windblown sands, often with dune formation; and in some areas (though not of direct glacial origin) windblown loess deposits. Sea and lake ice may raft large boulders and other coarse material (such as faceted sand grains), eventually depositing them far from the margins of continental or alpine glaciers. Finally, glacial ice erodes the surfaces across which it moves, leaving striations and molded rock forms parallel to the direction of motion.

Ancient sediments and surfaces have been found to display some of the above characteristics, providing evidence of past cold conditions. Both tillites (indurated tills) and ice-rafted boulders in marine sediments have been described from a number of geologic epochs. In most cases, however, the identification of such evidence with glacial climates is controversial, because other modes of origin can be adduced.

Nevertheless, it is widely accepted that prolonged cold glacial intervals occurred long before the well-known glaciations of the Pleistocene epoch (1,600,000 to 10,000 years ago). The Precambrian record is too much shattered by processes of metamorphism for easy interpretation, though episodes of Proterozoic (2,500,000,000 to 570,000,000 years ago) glaciation have been demonstrated. (The celebrated Gowganda tillite is of this age.) An Infracambrian ice age, about 570,000,000 years ago, is widely accepted, as is another in the Ordovician period (505,000,000 to 438,000,000 years ago). There is also no doubt that a major and probably prolonged (perhaps 10,000,000 years) ice age occurred about 280,000,000 years ago between the Late Carboniferous and Early Permian epochs. Widespread glacially striated surfaces, covered in part by Permo-Carboniferous tillites and erratic boulders, have been exposed in South Africa and Australia. Permo-Carboniferous tillites have also been identified in several areas that must have been located in the high southern latitudes of Gondwanaland, the ancient proto-continent composed of the landmasses of the Southern Hemisphere.

Glacial episodes

Uncertainty remains, however. Because glacial deposits are thin and are readily removed by subsequent erosion, the surviving record is desperately inadequate to provide details. Recent attempts at synthesis have even questioned whether it is safe to conclude that glacial ice was ever absent from former polar regions. Moreover, there may be no surviving evidence of some glacial episodes.

**Summary of pre-Cenozoic climates.** The scattering of equivocal evidence provided by sedimentary records reveals little about climates more than 570,000,000 years ago (Precambrian time). It is clear that in order for extensive sediments to accumulate in lakes and seas strong erosion must have occurred, so temperatures must have been in the range in which water is liquid; but more precise figures cannot be determined. There are evidences of early glaciations, but, like more recent events, these appear to have been relatively short-lived.

For the roughly 500,000,000 years between Precambrian time and the Cenozoic era, during which time the basic geography of the planet changed as the result of continental drift and mountain-building phases, the following generalizations appear to be supported by reasonable evidence:

(1) In general, the Earth possessed a usually warm and moist climate, with latitudinal zonations. The equatorial belt was at least as warm and humid as it is today, but the polar caps lacked ice, the seas were open, and high-latitude land vegetation (as it evolved) appears to have responded to cool temperate conditions. The equator-pole temperature gradient (which drives the atmospheric general circulation) was usually near half today's values. Deserts existed in the continental areas, probably in subtropical latitudes, but at certain times they became more extensive.

(2) Strong increases in poleward temperature gradients occurred at least twice, once at the outset of the Cambrian period (about 570,000,000 years ago) and again in Permo-Carboniferous times. Allowing for continental drift, these episodes appear to have been effective in creating cool polar conditions, with westerly winds around the polar caps. Evidence from the fossil record also suggests that the entire globe was relatively cool and dry from 320,000,000 to 208,000,000 years ago, during the Late Carboniferous, Permian, and Triassic periods. The fossil record then indicates that the Earth warmed substantially during the following 140,000,000 years (the Jurassic and Cretaceous periods). The Cretaceous period especially appears to have been the heyday of world climate, with calm, equable conditions. It was a time of expanded ocean surface, with extensive shelf seas. Mean annual surface temperatures exceeded present-day values by about 6° C, with no polar ice caps.

The Jurassic and Cretaceous warming trend

**Summary of pre-Pleistocene Cenozoic climates.** As is widely known, the Cretaceous period came to an abrupt end about 66,000,000 years ago with what appears to have been a catastrophic event, possibly due to asteroid impact. There were massive extinctions of many forms of life, perhaps due to rapid cooling induced by stratospheric dust veils that reduced solar radiation heating of the Earth's surface, allowing freezing temperatures to occur at sea level over much of the planet. (This hypothesis has not yet been substantiated, but it has led many scientists to believe that the smoke and dust produced by explosions in a nuclear war could cause similar cold, semidark conditions

and that this so-called nuclear winter could also result in widespread destruction of plant and animal life.)

Whatever the nature of the catastrophe that preceded it, the early Cenozoic era seems to have continued the warm, fairly moist conditions of the mid-Cretaceous. The land surface became covered with dense forests of higher plants, including most of the genera that are typical of today. These forests extended to the land areas near the poles of both hemispheres, which were close to their present positions. (The fossil remains of stumps and embedded roots of trees in the so-called Arcto-Tertiary forest are exposed in situ on two of the northernmost islands of Canada, Axel Heiberg and Ellesmere.) Since the continents had by this time assumed approximately modern latitudes, the zonal divisions typical of the modern world—high- and mid-latitude forests, arid zone, and tropical rain forests—begin to appear in the fossil record.

Oceanic and atmospheric cooling in the Cenozoic

Evidence from deep-sea sediments and Antarctic sites, however, shows that quite early in the Cenozoic there began a slow, intermittent cooling of the world's oceans, both at the surface and in deep waters (see Figure 48). There was an especially sharp decrease in temperature at the outset of the Early Oligocene epoch (36,600,000 to 30,000,000 years ago). Renewed and more profound cooling began in the Middle Miocene epoch (16,600,000 to 11,200,000 years ago). The cooling appears to have been most pronounced in the bottom waters. Thus was introduced the familiar modern fact of cold deep ocean water and floor, which is still of immense importance in the mechanics of climate.

There is also evidence of profound atmospheric cooling in Antarctica, whose thick ice sheet still covers almost the entire continent. Analysis of two cores extracted as part of the Deep Sea Drilling Project (a research program undertaken by several major U.S. oceanographic institutions from 1968 to 1983) has revealed the existence of diamictites and other sediments on the Antarctic shelf about longitude 76° E and latitude 66°–68° S. Magnetic polarity dating assigns these glacial sediments to the Early Oligocene epoch and possibly to the Late Eocene epoch (43,600,000 to 36,600,000 years ago). The evidence suggests that a large glacier complex was present at those times, the ice extending northward 140 kilometres beyond its present limit. The deposits also contain a leaf from a southern beech tree (*Nothofagus*), indicating that in warmer phases an Antarctic forest had been established and that a forest could be reestablished on recently exposed glacial deposits. There is much similar evidence to show that the Antarctic continent was partly or extensively glacierized from the Oligocene to modern times, with the ice sheet reaching its maximum extent 6,000,000 to 5,000,000 years ago during the Messinian age of the Late Miocene epoch (11,200,000 to 5,300,000 years ago). According to present opinion the repeated growth and collapse of the Antarctic ice sheet caused sea-level fluctuations of 80 to 100 metres.

Climatic asymmetry

This summary of Cenozoic climates suggests that the climatic asymmetry of the modern world may date to antiquity. Today, the Northern Hemisphere has warmer temperatures in all latitudes and less glacial ice than the Southern Hemisphere (the area of the Greenland ice cap is only 14 percent that of the Antarctic). There are corresponding differences in circulation of both atmosphere and ocean, the latter being much influenced by the relative absence of land barriers in the Southern Hemisphere. These contrasts may also have existed during the late Cenozoic. The earliest record of Northern Hemisphere alpine glaciation (in Alaska) is on the order of 10,000,000 years ago, considerably later than the Oligocene dates claimed for the Antarctic. The West German climatologist Hermann Flohn has argued persuasively that there was a protracted period (5,000,000 to 3,500,000 years ago) in which the Arctic Ocean remained ice-free while the Antarctic was glacierized. This is disputed, however, by scientists who maintain that the chief mode of sedimentation in the Arctic Ocean has been glacial ice-rafting since at least the Late Miocene (with icebergs adrift in the central ocean). Presumably, therefore, the Greenland ice cap dates to Miocene times. The recent dating of Antarctic glacierization to early Oligocene times, and perhaps beyond, does suggest, however, that Flohn's unipolar glaciation may have existed at an earlier time.

The Cenozoic thus appears as the period in which the world began to assume its modern appearance in at least four ways: (1) the emergence on land of climatically adapted vegetation zones, or biomes, with the flowering plants providing a continuous cover in all but the arid zone (which was always approximately subtropical in latitude); (2) a cooling toward the present climatic regime in which there is extensive glacierization on land in both polar zones, with tundra and cold temperate forest along the subpolar margins; (3) a cooling of the oceans, especially in the Early Oligocene and Late Miocene epochs, with

From N.J. Shackleton and J.P. Kennett, "Late Cenozoic Oxygen and Carbon Isotopic Changes at DSDP Site 284: Implications for Glacial History of the Northern Hemisphere and Antarctica," *Initial Reports of the Deep Sea Drilling Project* (1975); and J.P. Kennett, "Cenozoic Evolution of Antarctic Glaciation, the Circum-Antarctic Ocean, and Their Impact on Global Paleoceanography," *Journal of Geophysical Research*, 82:3843-3860 (1977); published by the American Geophysical Union

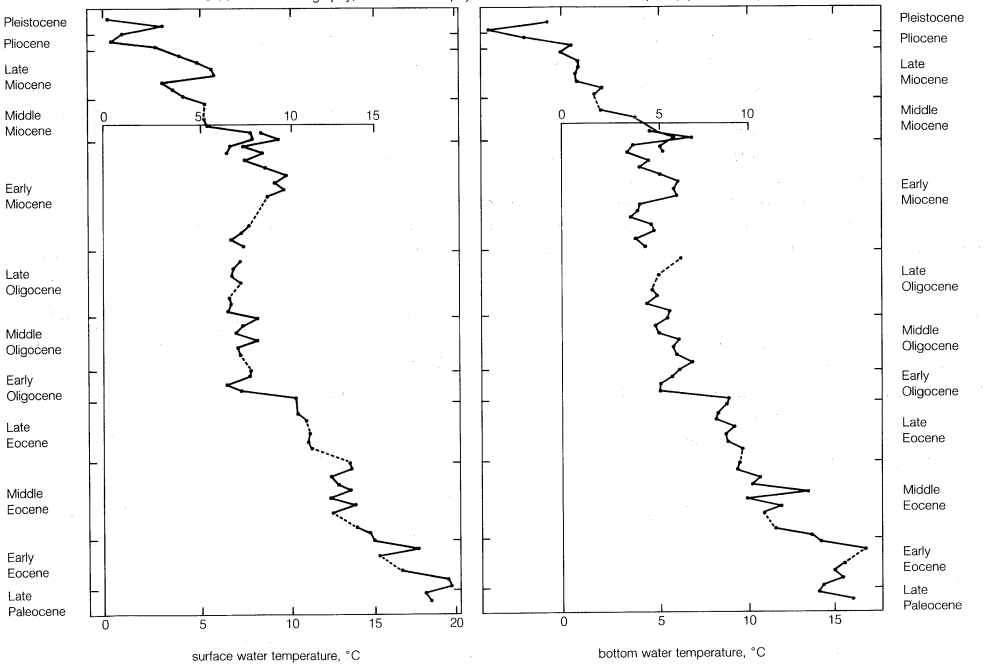



Figure 48: Temperature fluctuations of (left) ocean surface waters and (right) bottom waters. Data based on oxygen isotope analysis of cores from three sub-Antarctic sites.

the establishment of permanently very cold deep waters and ocean floors worldwide; and (4) the establishment of climatic asymmetry between the hemispheres, especially as regards the extent and magnitude of land glacierization.

## PLEISTOCENE CLIMATIC CHANGE

The cooling manifested in the middle and late Cenozoic led about 1,600,000 years ago (a precise date cannot be specified) into the remarkable epoch in which human beings slowly became the ecological dominants of the Earth, the Pleistocene. For almost a century it was believed that the Pleistocene epoch—often called the Ice Age in older literature—saw repeated expansions of ice sheets on the continents, with warm intervals during which the ice receded between the glacial episodes. Stratigraphic evidence, such as the mapping of moraines and till sheets with fossil soils and warm plant and animal assemblages sandwiched between them, suggested that there had been four glaciations of parts of Europe and North America, with four interglacial intervals. The Holocene epoch, which follows the Pleistocene and which constitutes the last 10,000 years to the present, was seen as the fourth of these interglacials, the ice being expected to return. The Holocene was designated separate from the Pleistocene because it encompasses the entire history of human civilization and is the epoch in which human beings overrode all competing lifeforms, but it differs in no other significant way from the interglacial phases of the Pleistocene.

The Ice Age

This comfortable account of four glacial and four interglacial episodes was shattered in the second half of the 20th century by the flood of new evidence introduced by a variety of new observational techniques. Among the new sources relevant to the reconstruction of past climates are: (1) the establishment of reasonably firm dating techniques that make possible more useful inferences from sedimentary and other records; (2) the growth, following a suggestion of the American chemist Harold C. Urey in 1947, of the study of stable isotopes of oxygen, hydrogen, and carbon in natural systems, and the realization that these isotopes could yield invaluable evidence of climatic variation; (3) the organization, initially by a small group of individuals from several disciplines, of systematic observations of the isotope record in glacial ice, ocean sediments, tree tissues, and ancient groundwater (including cave deposits); (4) the remarkable Deep Sea Drilling Project and its offshoots, which have yielded, with international collaboration, an immense array of sediment cores drawn from areas of the world ocean floor likely to yield direct evidence of climatic variation; and (5) the growth of paleolimnological and paleoecological studies of cores containing lake sediments and peat accumulations, chiefly by applying palynological analysis (the study of pollen content) to the record, mainly but not exclusively in Holocene sediments.

**Dating techniques.** There are three main methods used for dating geologic materials. One is through carbon-14 or other radioisotope analysis. Because radioactive elements decay into daughter products at known constant rates, unaffected by outside conditions, scientists can calculate the age of a material by measuring the ratio of original element to daughter product in the sample. A second dating technique involves geomagnetic polarity. Over the Earth's lifetime the polarity of its magnetic field has reversed many times. These reversals occur everywhere at the same instant. Thus, by studying the magnetic alignment of iron oxide particles, which depends on the direction of the Earth's polarity at the time the particles solidify, it is possible to establish the age of a material in relation to the known record of reversals. The third technique analyzes visual annual bands, such as tree rings, varve deposits (annual layers of sediment formed in lakes by glacial meltwater), and in some cases ice layers resulting from the seasonal thawing and freezing of glaciers. Occasionally, materials can be dated by determining their relative position to well-known, identifiable bands of sediment left by volcanic eruptions or periods of eolian dust depositions.

Radiocarbon dating

**Oxygen isotope analysis.** Once a sample has been dated, the climatic signal is then extracted in a variety of ways. For the Pleistocene epoch, it has been the application of oxygen isotope analysis to ocean sediments that has yielded the best results. Oxygen has two main stable isotopes, $^{16}O$ and $^{18}O$, and the waters of the ocean contain both. The $^{18}O$ is heavier than the more abundant $^{16}O$ because it has two extra neutrons in each nucleus. Evaporation favours water containing the lighter $^{16}O$. The result is that atmospheric water vapour has a measurable deficiency in $^{18}O$. Condensation, however, tends to reverse this process, so that the vapour is still more depleted in $^{18}O$, especially as temperatures fall (as they do when vapour is carried upward by convection or poleward in atmospheric disturbances). As a result, precipitation of the vapour falling in cold polar or cold temperate latitudes has a marked deficiency in $^{18}O$, and the water remaining in the ocean is correspondingly enriched in $^{18}O$.

The $^{18}O/^{16}O$ ratio, and hence $\delta^{18}O$ (the deviation, in parts per thousand, between the ratio of the sample and a standard ratio, usually that of standard mean ocean water or of a fossil belemnite known as PDB), is readily measured in research laboratories and can be used to analyze cores from ocean sediments, ice sheets, tree trunks, speleothems (stalactites and stalagmites from caves), and any other relevant material. Similar stable isotope techniques are applied to hydrogen and carbon. The carbon isotope ratio has been of special interest in tree-ring analysis.

In the case of ice cores, the $\delta^{18}O$ record is assumed to yield information about the temperature of the evaporation–condensation sequence that preceded the deposition of the material being analyzed. Suitable calibration curves are well established, though their use is actually quite complicated. Excellent records have been analyzed from Antarctica, Greenland, and some small Canadian ice sheets.

The climatic interpretation of marine sediments relies chiefly on the fact that much of the accumulated material consists of the solid remains (tests) of microorganisms. These may originally have been either shallow-water plankton or benthic species living near or on the ocean floor. In either case, the tests provide information about the $^{18}O/^{16}O$ ratio of the water that supported them, although not in unmodified form; the two isotopes are incorporated into the tests in a temperature-dependent fashion. Thus, ocean floor oozes, which under the microscope can readily be separated into their constituent organisms, contain a record of the local temperature of the original habitat of the species concerned (surface- or bottom-dwelling). Certain indicator species have been standardized for such analyses. In addition, the fact that the ice contained in continental ice sheets is poor in $^{18}O$ has meant that throughout the Pleistocene epoch ocean water has been enriched in $^{18}O$ at times of high ice volume. In other words, the oxygen isotope ratio record from the ocean sediments can also be used as a measure of the amount of water withdrawn from the oceans to lie on continental surfaces as ice. Clearly this withdrawal should also appear as changes in sea level (another frequently cited indicator of glacial/interglacial climatic contrasts).

Relation between oxygen isotope ratios and ice volume

Figure 49 shows a record of oxygen isotope deviations in sediments from an equatorial Pacific core spanning the past 850,000 years. The geomagnetic polarity reversal, at 730,000 BP, anchors the time scale. The numbers directly above the curve indicate the different stages of

From N.J. Shackleton and N.D. Opdyke, "Oxygen Isotope and Paleomagnetic Stratigraphy of Equatorial Core 28-238: Oxygen Isotope Temperatures on a $10^5$ and $10^6$ Year Time Scale," *Quaternary Research*, 3:39 (1973); and N.J. Shackleton in A.B. Pittock *et al.* (eds.), *Climatic Change and Variability* (1978), Cambridge University Press

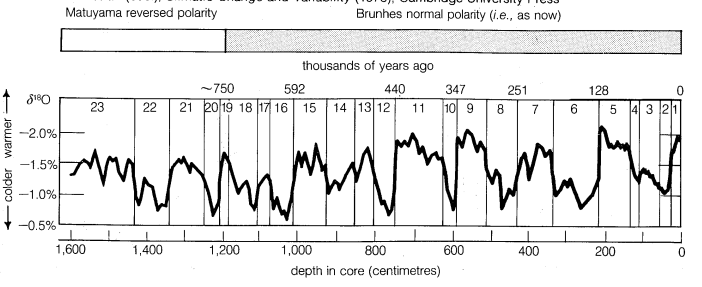



Figure 49: Oxygen isotope deviations in sediments from an equatorial Pacific core (V28-238, at roughly 1° N, 160° E), showing sharp and rapid alternations of high and low volumes of continental glacial ice (see text).

the climatic sequence. Odd numbers, corresponding to stages with more negative values of $\delta^{18}O$, mark times of low glacial volume. Even-numbered stages, or those with higher values of $\delta^{18}O$, mark cold glacial phases. As this and other records show, there appears to have been a high volume of glacial ice on land—*i.e.,* withdrawn from oceanic waters—many times in the past 3,000,000 years. Between these glacial episodes came shorter phases of low glacial ice volume, presumably because of warmer climates.

**Glacial and interglacial phases.** Using oxygen isotope analysis and other sources, climatologists have been able to piece together an extraordinarily detailed record of the Pleistocene climates, though much remains to be done. Overall, the record shows that the entire epoch was characterized by fluctuating glacial ice volumes, which were associated with swings of global air temperatures and equivalent oceanic changes (including small but extensive shifts in the temperature of the deep waters in all the main ocean basins). The number of glacial episodes and of interglacials clearly far exceeds the numbers that were once suggested by the study of Pleistocene sediments on land. Unfortunately, the complete history of the earlier glacial episodes on land may never be known because so much of the record was obscured or removed by subsequent glaciations.

The overall swings in planetary temperature between glacial and interglacial episodes have still not been fully determined. Interglacial temperatures as much as 5° C higher than those of today have been inferred from some records. Glacial temperatures on land in the range of 10° to 15° C below current values also have been cited. It is certain that the actual departures were lowest in equatorial oceanic areas and that the largest swings were in high latitudes.

Glacial temperatures

Considerable detail is available for the past 250,000 years. This long interval saw two major glacial episodes, the last of which, called the Wisconsin in North America and the Würm in the European Alps, left most of the available field evidence of glaciation on land surfaces. An interglacial separated the two glaciations, beginning about 130,000 years ago (though there is dispute over this date); it yielded to renewed glaciation at least 80,000 years and perhaps as long as 110,000 years ago. It is usually argued that the start of this warm interglacial interval coincided with a maximum of summer solar radiation in the Northern Hemisphere, which is indeed confirmed by the Milankovitch theory of orbital variations (see below). The evidence also points to an enduring characteristic of the global climate: *i.e.,* that major swings in temperature, such as glacials and interglacials, tend to affect both hemispheres simultaneously and thus are global events.

Figure 50 summarizes an important achievement of Antarctic field research, the analysis of an ice core from the Soviet research station Vostok, yielding a record that spans 160,000 years. Curve A shows the carbon dioxide concentration in trapped air bubbles in the ice as a function of date. Curve B shows the temperature at the time, as derived from $\delta^{18}O$ analysis. The resemblance between the two curves shows that temperature and carbon dioxide concentration went hand in hand through most of the late Pleistocene epoch (though it does not say which is cause and which effect). Also plotted in Figure 50 are curves showing estimated changes in sea level. Curve C, based on $\delta^{18}O$ analysis of an east equatorial core from the Pacific Ocean floor, shows how global sea level may have varied. Sea level was highest at the warmest times, corresponding to small volumes of ice in Antarctica (and in the Northern Hemisphere) and high levels of atmospheric carbon dioxide. Curve D shows the calculated dates of coral terraces, uplifted rapidly over the past 160,000 years or so on the coast of Papua New Guinea, in the Huon Peninsula. The sea-level curves bear strong resemblances to the Vostok data. Thus, there is parallelism between all four pieces of evidence, which are drawn from disparate sources, demonstrating the global nature of major climatic variations.

The global nature of climatic variation

Some uncertainty remains, however, concerning the correctness of the dates used on Figure 50. Figure 51 shows a third long record, extending from 50,000 to 310,000 years ago. It is drawn from borings through a calcite vein, made

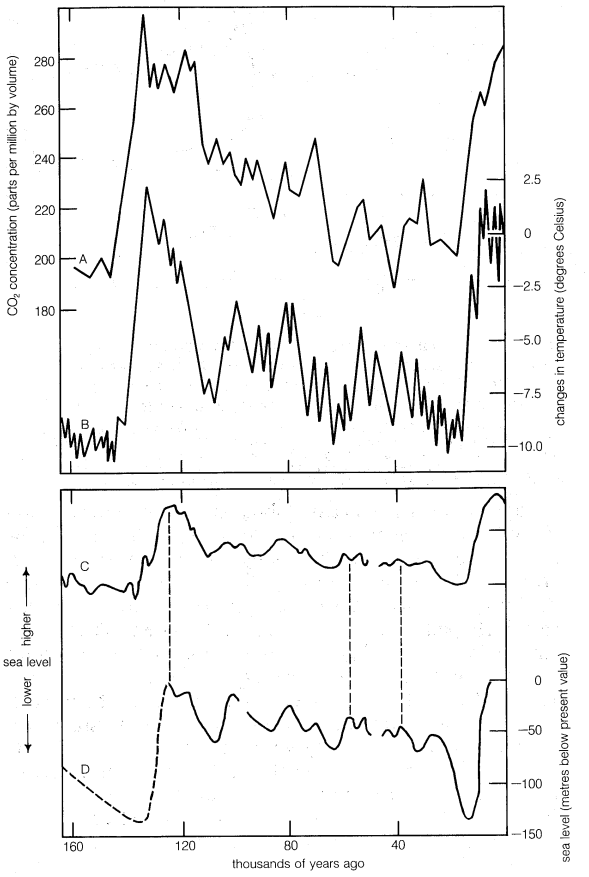



Figure 50: *Agreement among variations in four diverse climatic indicators (see text).*
(A) Carbon dioxide concentration in bubbles from the deep ice core at the Soviet Antarctic station Vostok (present-day value 350 p.p.m.v.). (B) Inferred temperatures at Vostok. (C) Estimated changes in global sea level. (D) Estimated changes in sea level at Huon Peninsula, Papua New Guinea.

Compiled by P.J. Barrett from *New Scientist,* vol. 118, no. 1608 (1988), which first appeared in *New Scientist,* London, the weekly review of science and technology; N.J. Shackleton and N.D. Opdyke, "Oxygen Isotope Paleomagnetic Stratigraphy of Equatorial Core 28-238: Oxygen Isotope Temperatures on a $10^5$ and $10^6$ Year Time Scale," *Quaternary Research,* 3:39 (1973); and J. Chappell and N.J. Shackleton, "Oxygen Isotopes and Sea Level," reprinted by permission of *Nature,* vol. 324, no. 6093, pp. 137–140 (1986), copyright © Macmillan Magazines Limited

of material deposited from percolating surface waters at Devils Hole, an open fault zone in the Great Basin of the United States near the Nevada–California border. In this case, the material was dated using two radioisotope techniques, which gave very similar dates. The chronology is believed to be correct within approximately 3,000 years at the 140,000-year point on the curves.

There is again a resemblance between the Great Basin curve and those in Figure 50, two of which are reproduced in Figure 51. The main peaks and troughs, however, fit closely only if the lines joining them are slanted (whereas they should be vertical if these are simultaneous global events). Termination II, the assumed beginning of the last interglacial, appears to have occurred in Nevada 147,000 ± 3,100 years ago, or 16,000 years before it appears in the oceanic record. Thus, the Great Basin curve calls into question the validity of the dating techniques used for the Antarctic ice and the marine sedimentary records. If the Great Basin chronology is correct, the view that there is a close fit between solar orbital data and temperature needs to be reexamined.

One fact that emerges clearly from these records, and from the earlier work of Willi Dansgaard and his colleagues in Greenland, is the apparent short length of the last interglacial. High temperatures, sea levels, and carbon dioxide content seem to have lasted barely 20,000 years. The period between 110,000 years ago and the culmination of the Wisconsin/Würm glaciation 18,000 years ago was one of predominantly cold conditions and low sea levels, though both sea level and temperature fluctuated rapidly as they sank to their lowest recent levels in 18,000 BP.

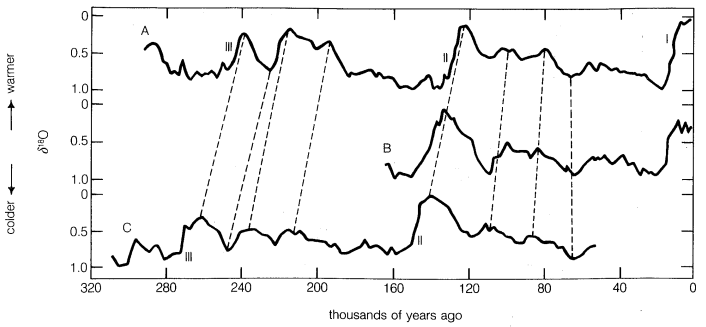



Figure 51: Comparison between timing of oxygen isotope deviations in sediments from (A) an equatorial deep sea core, (B) the Vostok glacial ice core, and (C) the U.S. Great Basin. Slanted dashed lines imply a discrepancy in dates (see text).

From I.J. Winograd *et al.*, "A 250,000 Year Climatic Record from Great Basin Vein Calcite: Implications for Milankovitch Theory," *Science*, vol. 242, no. 4883, pp. 1275–1280 (1988), Copyright by the American Association for the Advancement of Science; and C. Lorius *et al.*, "A 150,000 Year Climatic Record from Antarctic Ice," reprinted by permission of *Nature*, vol. 316, no. 6029, pp. 591–596 (1985), copyright © Macmillan Magazines Limited

**Summary of Pleistocene climates.** Uncertainties still abound, but a clear view of Pleistocene climate is now within reach. The epoch emerges as an interval of almost 2,000,000 years, during which temperatures were usually below those normal to geologic history and during which polar ice caps waxed and waned. There were many episodes of full glaciation when global temperatures may have been 5° C or more below present values, and perhaps 10° C below more ancient norms. There were, however, other intervals—interglacials or interstadials—of relative warmth, usually with enriched carbon dioxide concentration in the atmosphere. Most warm intervals appear to have been quite short, and they seem to have ended abruptly with a return to colder glacial conditions. These repeated changes (vastly more numerous and more complex than was once supposed) and the general coolness continued a trend clearly visible during the preceding Tertiary period.

The events of the Pleistocene epoch have left a remarkable record in ocean sediments, in the Greenland, Antarctic, and Canadian ice caps, and in certain continental sediments. Modern geochemical and geophysical techniques have made possible quantitative estimates of temperature, sea level, and atmospheric compositional changes. These have been coupled with paleoecological reconstructions of past living communities on land and in the sea. The Pleistocene is the earliest period for which the record is amenable to quantitative methods.

## LATE PLEISTOCENE AND HOLOCENE CLIMATIC CHANGE

Evidence bearing on the most recent 18,000 years of Earth history is more diverse and persuasive than that available for earlier epochs. The geophysical and geochemical methods described for the Pleistocene continue to be valuable, but emphasis now shifts toward paleolimnological and paleoecological data (lake sediments and peat deposits, interpreted chiefly for their pollen contents) and the art of climatic modeling. In the 1980s, research was further expanded by the coming together of two quite different traditions, those of field geologists and biologists on the one hand and those of atmospheric scientists on the other. The result has been a remarkable advance in climatic knowledge.

CLIMAP and COHMAP

Two exercises involving leading research figures from many disciplines and several countries have particularly contributed to the fund of knowledge. The acronyms by which they are known are CLIMAP (Climate: Long-range Interpretation, Mapping, and Prediction) and COHMAP (Cooperative Holocene Mapping Project). CLIMAP was an attempt to specify in considerable detail the condition of the Earth's surface, most notably the oceans, at the climax of the Wisconsin glaciation 18,000 years ago; it also included a series of mathematical modeling exercises aimed at defining the atmospheric circulation present at that time. COHMAP is a later exercise designed to unravel the history of deglaciation of North America and Eurasia, the recolonization of the northern land surfaces by plants and animals, and the equivalent changes in the tropics and the Southern Hemisphere. Much of the following discussion is based on information derived from these two studies.

At the time of the late Wisconsin maximum, and for several millennia on either side of 18,000 BP, two major ice sheets, the North American complex and the Fennoscandian ice sheet, covered a substantial part of the northern continents. These must have accumulated during the long, intermittent cooling in the interval following 110,000 BP. Furthermore, because the solar radiation regime at 18,000 BP did not differ much from that of today, the ice sheets must have been inherited from the past. They, nevertheless, imposed their own effects on world climate.

The North American complex included a deep cover over the Western Cordillera, the Laurentide sheet centred on Hudson Bay, and the Greenland sheet (in part linked with the Laurentide by floating shelf ice, also present along the Arctic Ocean flank). Its southern limit on land extended from Puget Sound across the Missouri and Ohio valleys to New England and the Canadian Atlantic provinces. The Fennoscandian ice sheet covered Scandinavia, Finland, much of Britain and Ireland (which had local glaciation as well), parts of the North European Plain, and the northwestern republics of the Soviet Union. Smaller ice sheets covered the Alps and certain mountain masses in central and eastern Asia. Alpine glaciers were extended in many of the world's mountain systems, including those of the tropics. There was a major expansion of the Arctic Ocean pack ice into the Norwegian Sea, with seasonal ice extending into the Atlantic as far south as a line from Portugal to the Grand Banks off Newfoundland. The southern limit of the cold Arctic water—the oceanic polar front—was pushed far south of its present locus.

The Antarctic ice sheet, then as now the largest glacial mass, was surrounded by a large expansion of the seasonal pack ice belt in the Antarctic Ocean. Alpine glaciation affected the Andes, the southeastern Australian ranges, and New Zealand, but the absence of extensive land removed any possibility of large-scale southern hemispheric glaciation outside Antarctica.

These conditions speak of a world with annual temperatures 5° to 6° C below present values, with averages over and near the ice sheets perhaps 10° to 15° C below. Because of the equatorial shift of the oceanic polar fronts, there were extensive areas of both southern and northern oceans that were 10° C colder than at present. A worldwide drop in humidity and precipitation (with local exceptions) must also have occurred.

**Causes of deglaciation and Holocene warming.** It is now widely held that the rapid recovery of temperatures after the Wisconsin maximum was caused by variations in the Earth's orbital behaviour. This astronomical or orbital variation theory of climatic change, which was developed in 1920 by Milutin Milankovitch, a Serbian scientist, explains not only the Holocene recovery but also many of the remarkable fluctuations of Pleistocene times. The theory is based on the fact that the orbit of the Earth around the Sun varies in three ways. All of the variations are small, but, especially when they reinforce one another, they are sufficient to cause significant redistributions of solar heating between latitude belts or hemispheres. Because the hemispheres differ greatly in surface properties and heat storage capacity, these orbitally caused variations in the pattern of solar heating are capable of changing global climate. The three modes of variation are the obliquity, or tilt, of the Earth's axis, which varies with an average period of about 41,000 years; the precession of the equinoxes, which refers to the movement of perihelion (the point when the Earth is nearest to the Sun) around the path of orbit and which varies with a period of roughly 22,000 years; and the eccentricity of the orbit, varying with a period near 100,000 years. Elaborate mathematical analysis of these parameters has made it possible to calculate the corresponding changes in solar heating of the Earth over the past 750,000 years.

The Milankovitch theory of orbital variations

These calculations show, for example, that over the past 18,000 years the amount of solar energy reaching the Earth in summer was concentrated in the Northern Hemisphere, reaching a peak of 40 watts per square metre

above 1950 values between 12,000 and 9,000 years ago. This corresponded to an 8 percent increase in northern high-latitude summer heating, enough, it is believed, to account for the wastage of the Fennoscandian and North American glaciers (only Greenland's ice survives today). The winter cooling that compensated for the extra summer heat (orbital changes produce such seesaw effects) was most effective in the subtropical latitudes of the Southern Hemisphere. After 9,000 BP these trends reversed, and the modern world has a solar heating pattern much like that of the late Wisconsin maximum in 18,000 BP. The climatic impact of these changes, as calculated for the COHMAP exercise, is shown in Figure 53.

The right-hand panels of the figure illustrate the main features of the atmospheric circulation over the North American-European region (at 3,000-year intervals) and over the African-Indian Ocean-Australasian area (at 9,000-year intervals). These sketches are based on more detailed originals derived from the application of an atmospheric general circulation model (AGCM) to the conditions applicable at the time, including the solar radiation, the extent of ice cover, sea-surface temperatures, and other boundary conditions. The left-hand panels are COHMAP reconstructions of the measured environmental changes in the two sectors during the past 18,000 years. For land areas these are based primarily on pollen analysis of dated lake sediments, with some data from long peat sequences (a recent synthesis extends back into the last interglacial). The striking northward march of the forest biomes into the glaciated northern landscapes dominates events on land. Also visible is the rapid shrinkage of glacial ice on land: the Fennoscandian ice sheet was almost extinct by 9,000 BP and the Laurentide sheet by 6,000 BP. At sea there was a major reopening of the northern North Atlantic and Norwegian seas as the pack ice front retreated. In addition, as glacial ice melted, sea levels rose, reducing land area in, for example, northwestern Europe and Southeast Asia.

The hypsithermal interval

By 9,000 BP summer temperatures and precipitation had risen above present-day values in much of the land area of the Northern Hemisphere. These warm conditions—the so-called hypsithermal interval (or, in older usage, climatic optimum)—remained in place until about 6,000 BP, after which temperatures fell slowly to their present values (with minor irregularities discussed later) some centuries before the Christian era. This period of warmer climate occurred earlier in western North America than in the east, where the shrinking Laurentide ice kept temperatures cool. It also made an early appearance in parts of the Southern Hemisphere. In New Zealand, for example (where the glacial maximum is believed to have preceded that of the Northern Hemisphere by 6,000 to 7,000 years), warmest conditions appear to have occurred between 10,000 and 8,000 BP.

The Holocene warming was accompanied by far-reaching climatic changes in the arid zone and in the monsoon lands of Africa and Asia. Like the warming, these changes were in response to the orbitally caused changes in the amounts of incoming solar radiation. Geologic data show that rainfall in these areas was especially affected.

Relative moisture during the Holocene warming

An excellent measure of rainfall in the arid zone is the water level in closed lake basins (many of which are now dry). Figure 52 shows the changes in lake levels (in generalized form) on the chief continents. In North America, where the obstacle of the Laurentide ice sheet and associated anticyclonic circulation drove the main westerly jet stream southward throughout the Wisconsin, lake levels in the western United States (especially the Great Basin) were high until about 10,000 BP, when the ice sheet receded and circulation assumed its modern pattern (with the jet stream entering the continent further north at about 50° N). Because the jet stream steers the cyclonic storms bringing rain and snow, the now-arid southwest was abundantly watered until this northward shift.

African lakes, by contrast, were low in glacial times. After about 12,000 BP, however, they rose to high levels, especially between 10,000 and 7,000 BP, until general desiccation occurred after 5,000 BP. While general cooling affected high northern latitudes at this time, there seems to have been a marked slackening in the monsoon rainfall

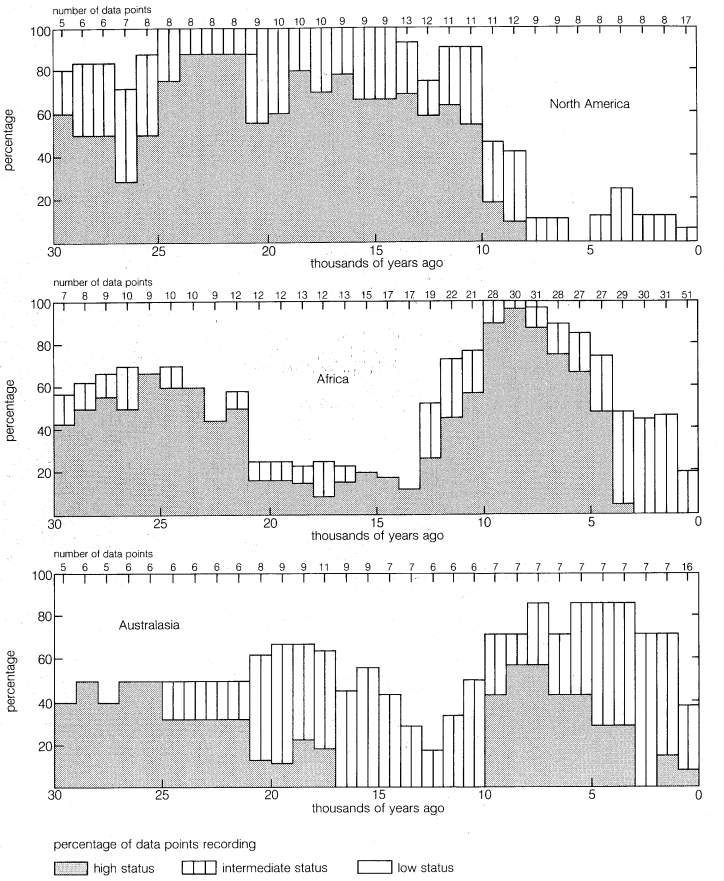


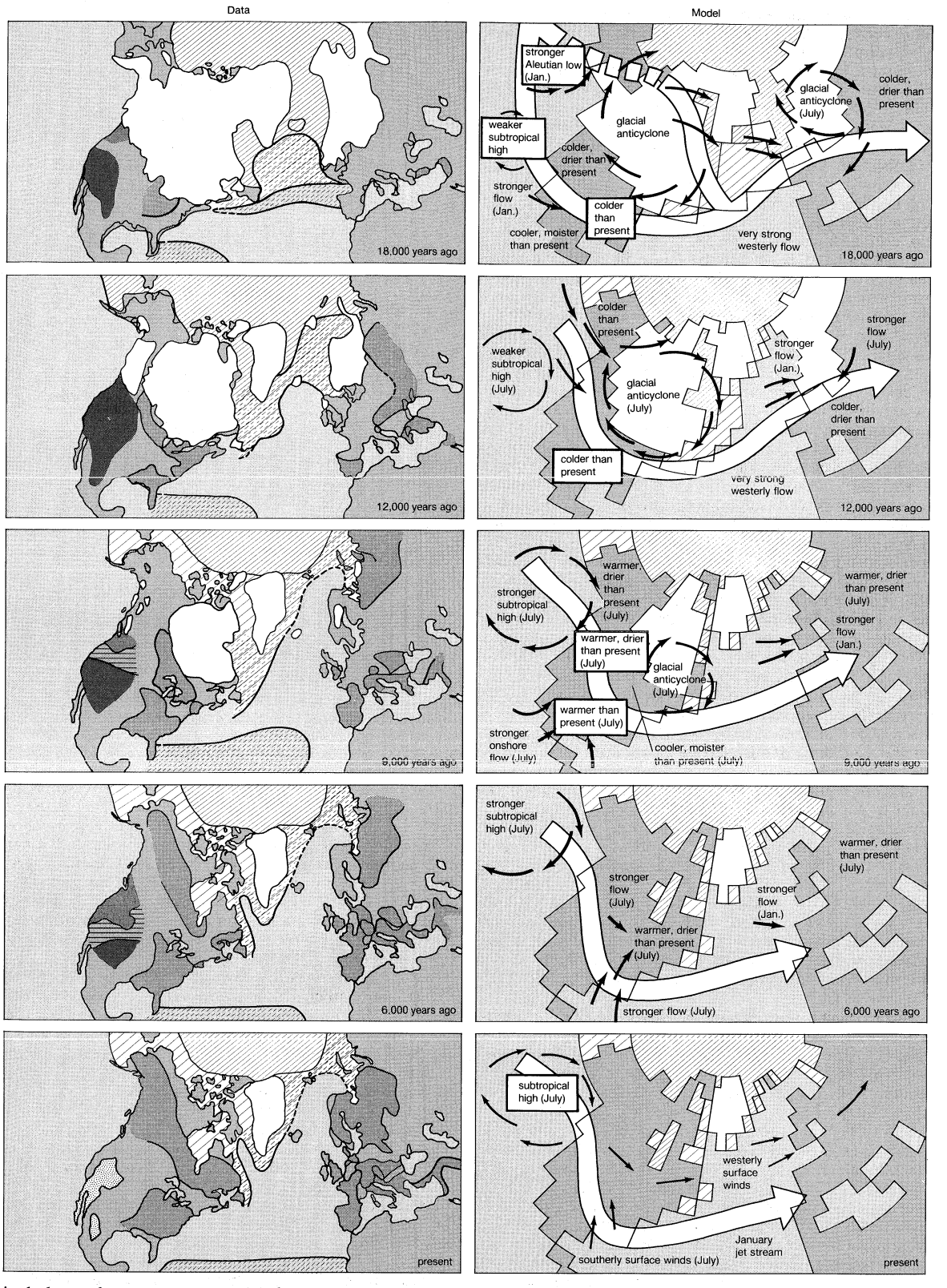

Figure 52: Estimates of percentage of high, intermediate, and low water levels in inland lakes.

From. F.A. Street and A.T. Grove, "Global Maps of Lake Level Fluctuations Since 30,000 B.P.," *Quaternary Research*, 12:83-118 (1979), Academic Press

that fed many of the African lakes. In Northern Africa, in particular, the moist phase of the early Holocene attested to by the high lake levels corresponded to the appearance of lush savannas supporting hippopotamuses and other large animals throughout the present-day Sahara desert. This early Holocene moist phase was widespread elsewhere in the Middle East and other parts of Asia, enabling (as it closed) early human innovators the opportunity to move from hunting to pastoralism and from gathering to arable agriculture.

Research has confirmed that the Milankovitch orbital changes strengthened the monsoonal currents of the Northern Hemisphere during the early Holocene. At the time of the Wisconsin maximum the monsoon was weak, but between 12,000 and 5,000 BP summer temperatures in inner Asia were as much as 4° C warmer than today, making possible the strong monsoon of this period.

**The Recent (pre-instrumental) phase.** Although climatically inseparable from the Holocene epoch, the time period in which human civilization developed is sometimes referred to as the Recent phase. This phase comprises the past two millennia globally and the past nine locally. A much more secure chronology of events can be compiled for this period, and there is a substantial increase in the body of climatic evidence available. In fact, the volume of evidence is so great that only a short list of the major sources can be presented here. Prior to the 17th century, when instrumental records began to be kept, these sources included: (1) the methods already defined for the Holocene epoch, except for those with low time resolution capabilities (such as the solar orbital data); (2) direct glaciologic and glacial geomorphologic observation, especially in intensively studied areas such as the European Alps and the Scandinavian countries; (3) tree-ring analysis, or dendrochronology, which provides evidence for the past nine millennia in favourable areas; (4) full archaeological evidence; and (5) the vast documentary record, which tends to provide anecdotal or episodic evidence and




includes such sources as records of grape, hay, and grain harvests, logs of harbour conditions, and land-use records. Content analysis has recently sharpened use of some of these sources.

The picture that emerges from the evidence is of a climate quite similar to that of today, yet not discontinuous with that of middle and late Holocene times. The differences that have been demonstrated are often subtle and poorly defined. Many long-term records, such as those for stream flow, can even be interpreted as a form of noise, lacking meaningful pattern. Others have obvious and prolonged fluctuations that deserve mention, but it is usually difficult to establish the geographic extent of these fluctuations, because the record is badly broken in time and space.

One of the most detailed local records is that provided by the pollen analysis of the great peat bogs of northern Europe. During the late prehistoric and early historic phases of the past 6,000 years, these bogs appear to have undergone a number of desiccations (warm, dry summers), revealed in the bog cores as dry, often wooded, surfaces. The dry phases were generally followed by wet conditions in which peat accumulation was rapid. These overlying

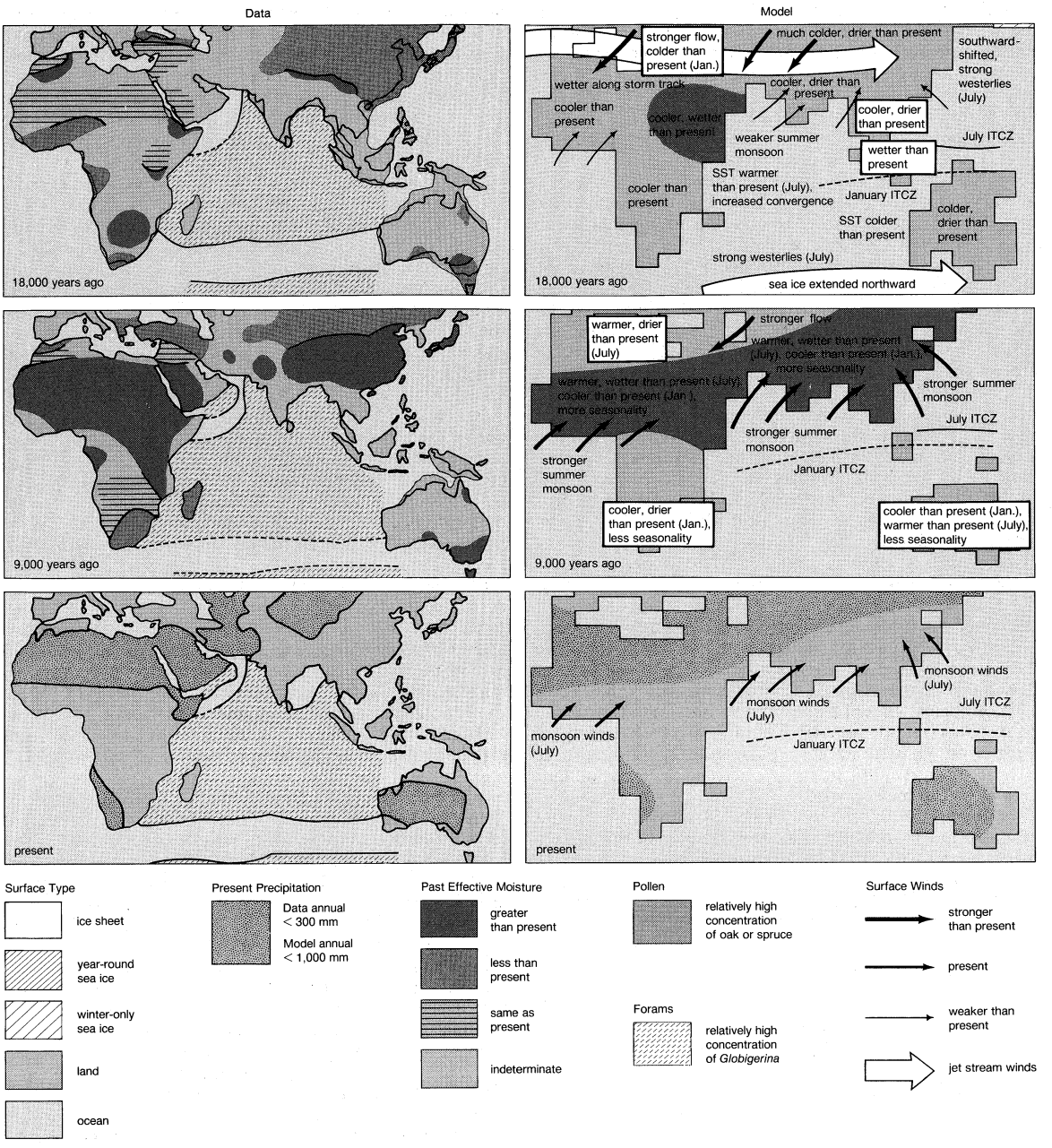



Figure 53: COHMAP reconstructions of (left panels) changes in environmental distributions, based on observed data, and (right panels) estimated changes in climate, based on general circulation model (see text).

From COHMAP Project Members (1988), "Climatic Changes of the Last 18,000 Years: Observations and Model Simulations," *Science*, vol. 241, no. 4869, pp. 1043–1052 (1988), copyright 1988 by the American Association for the Advancement of Science

layers of renewed peat growth are known as "recurrence" horizons. For a long time the analysis of these and other such singularities in the peat and lake-floor record seemed indispensable to the interpretation of the post-glacial climatic record. Today, they are viewed as evidence of minor fluctuations in a record that does not greatly depart from present conditions. Thus, the slight cooling that followed the hypsithermal interval is recognized as spasmodic. Similarly, it is clear that Europe in the period 500 BC to AD 1000, which was crucial to the continent's human settlement, was subject to climatic fluctuations whose sociological importance far outweighed their small magnitude.

The medieval warm period

There is general agreement that from perhaps AD 600 to 1400 European and North Atlantic climates became warmer and probably drier. Evidence of this warming comes from many regions and includes such sources as agricultural records as well as limited pollen, water-level, and macrofossil data. Although the evidence is often rather flimsy, it implies that during those eight centuries temperatures were perhaps as much as 1.5° C higher than those of today. As a result of this medieval warm period, Viking settlers were able to establish thriving colonies in Iceland, Greenland, and Newfoundland. The northern limit of viticulture, which requires sunny, warm conditions, extended farther north than it does today, into northern France and Germany and southern England. In Arctic Canada and Greenland the highly developed Thule culture, which depended on whaling for sustenance, flourished late in the period, until colder conditions made whaling impossible.

Although it is clear that the Middle Ages were characterized by warm conditions, any objective compilation of local records must reveal that no real coherence exists. The climax of the warming occurred at different times in different places, and no region experienced stable, continuous warmth for more than two or three centuries. Alpine glaciers were far advanced in their valleys at the height of the period, and there is much other evidence that confirms the variability of the regime. Outside Europe and the North Atlantic area, in fact, there is little evidence that this warm phase ever took place.

The Little Ice Age

Studies of glacial sediments, tree rings, and written records show that, from the beginning of the 16th century until the mid-19th century, cooler and harsher conditions prevailed in most parts of the world. As with the medieval warm period, the climax of the so-called Little Ice Age arrived at different times in different areas. In general, the Northern Hemisphere, especially North America, the Arctic, and East Asia, seems to have been affected earlier than the Southern Hemisphere. Some conditions appear to have been globally synchronous, but their magnitude varied from place to place. In addition to regional variations, there was increased variability from season to season, from year to year, and from one group of years to the next, so that, for example, central England experienced several of its hottest summers in the 17th century, one of the coldest centuries on record. Some of the effects of the Little Ice Age on human life are reflected in records of crop failures and famines, the abandonment of northern farms and villages, and changes made in sailing routes to avoid advancing ice. Although the main cold phase ended about 1700, the bleak winters and cool, moist summers that characterized the Little Ice Age continued in many parts of the world until as late as the mid-19th century. In 1850, for example, meteorological instrument records show that temperatures were still about 0.7° C below present-day global values.

**Contemporary climatic change.** Meteorological instrument records survive from as early as the 17th century, but it was not until about 1850 that the number of well-calibrated instrumental records of temperature and precipitation was adequate for statistical analysis. Since then, the data files have accumulated rapidly, and in recent decades reasonable quality control has been achieved by the World Meteorological Organization. The record extends over the entire globe (now observed continuously by a variety of sensors on geostationary or polar-orbiting satellites), as well as upward into the stratosphere, with less-abundant data from the mesosphere.

*Temperature fluctuations.* Figure 54 shows an estimate of global mean annual temperature changes at the surface from 1860 to the 1980s. The estimates are adjusted to allow for the unequal spacing of data points between land and sea, between mid-latitudes and the polar regions, and between the hemispheres. Data from cities with a population of more than 100,000 have been omitted or adjusted because of the so-called urban heat island effect. (Urban areas are warmer than the surrounding countryside because asphalt, concrete, and other urban surfaces absorb and store heat more efficiently than rural vegetation.) The curves show: (1) an almost linear increase of temperature throughout the record, amounting to 0.7° C since 1860; (2) an irregular sequence in the Northern Hemisphere, where in particular the period 1940–70 showed no increase, or even a slight decline; (3) a much more uniform upward trend in the ocean-dominated Southern Hemisphere; and (4) an especially vigorous warming in the 1980s, the warmest decade in the entire record.

From (top) P.D. Jones, T.M.L. Wigley, and P.B. Wright, "Global Temperature Variations Between 1861 and 1984," *Nature*, vol. 322, no. 6078, pp. 430–434 (1986); and (bottom) P.D. Jones *et al.*, "Evidence for Global Warming in the Past Decade," *Nature*, vol. 332, no. 6167, p. 790 (1988); both reprinted by permission of *Nature*, copyright © Macmillan Magazines Limited

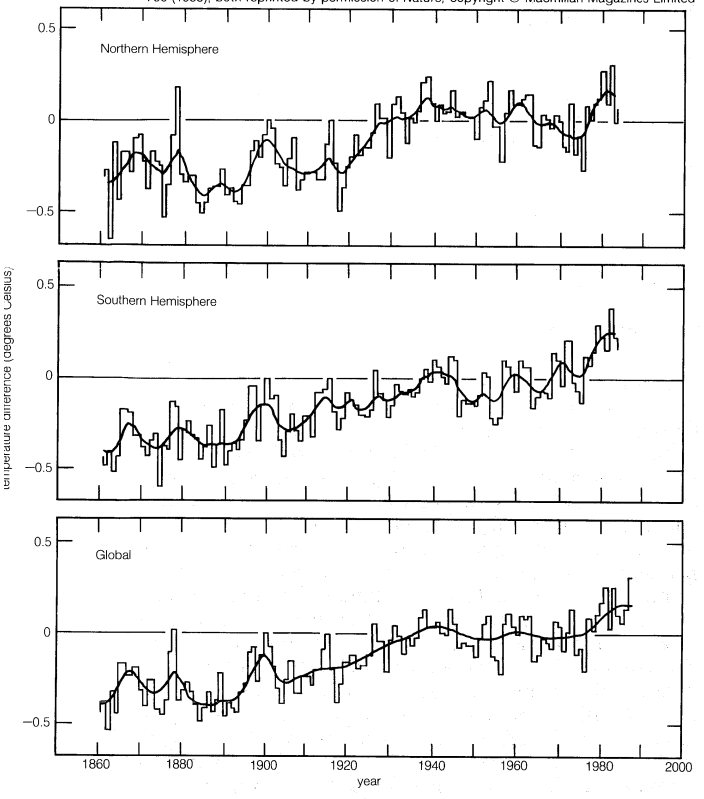



Figure 54: Variation of mean surface air temperature from 1860 to the 1980s.

There have been questions as to whether the effect of urban heating on local temperatures has been truly eliminated from such data. An investigation suggested that as much as 0.1° C of the apparent warming might have been due to this contamination of the record. One similar temperature analysis is believed to have been seriously compromised by urban heating. The self-heating of cities is, of course, a real effect that has had significant economic consequences. On the global scale, however, it is in reality negligible. Unfortunately, this may not be true of its impact on the record.

The contemporary global warming trend

It appears that a real global warming of 0.4° C has occurred in the 20th century and that it accelerated in the 1980s after a three-decade hesitation. If the responsible agent is the greenhouse effect (see below *Climate and life*), the warming is at the lower end of the model-predicted range; higher temperatures might have been expected from the increase in greenhouse gases observed since the 19th century. Greenhouse heating might also have been expected to be concentrated in polar and subpolar latitudes, while, in fact, temperature increases in the Northern Hemisphere have been greatest in a high mid-latitude belt. Many large regions, including much of North America, so far show little or no upward trend of temperature.

*Precipitation fluctuations.* Precipitation measurements are less amenable to study and in any case are only available on land. Figure 55 shows an analysis of precipitation data from Northern Hemisphere land stations. In each graph the percentile position is relative to frequency distribution, as specified by a mathematical representation. (For example, a percentile value of 0.6 in any given year means that precipitation in that year was greater than or equal to 60 percent of the other annual values.) This analysis shows that little has changed in the equatorial belt since about 1900, but a significant decrease in precipitation has affected subtropical areas (5°–35° N) since 1960. Mid- and high-latitude stations (35°–70° N) show a substantial increase since 1950. Again the effect is nonuniform, large areas—notably Europe—having escaped any obvious change. Studies have also shown that there are marked seasonal differences in precipitation.

In the 1980s there have been many severe droughts, and there is a natural tendency for lay observers to view these as manifestations of the greenhouse effect, several formulations of which predict drought in mid-latitude areas. Although this is a possibility, most authorities agree that the droughts and other fluctuations are expressions of climate's natural internal variability and not of externally forced change. The reasons for such large-scale variability remain unclear, however. Some authorities have argued that variability tends to increase as mean temperatures decrease, so that the possibility of damaging cold extremes is enhanced. Others see the variability as largely random.

*Other fluctuations.* Two remaining aspects of recent climatic fluctuations (*i.e.*, nonlasting change) that deserve attention are the great African desiccation of the 1960s, '70s, and '80s and the recognition of the El Niño/Southern Oscillation (ENSO) phenomenon as a source of worldwide variability.

Rainfall in Africa has been highly disturbed since the 1950s, especially in sub-Saharan regions. Numerous studies have demonstrated that the generally high values applicable to the 1950s and early '60s gave way to a prolonged

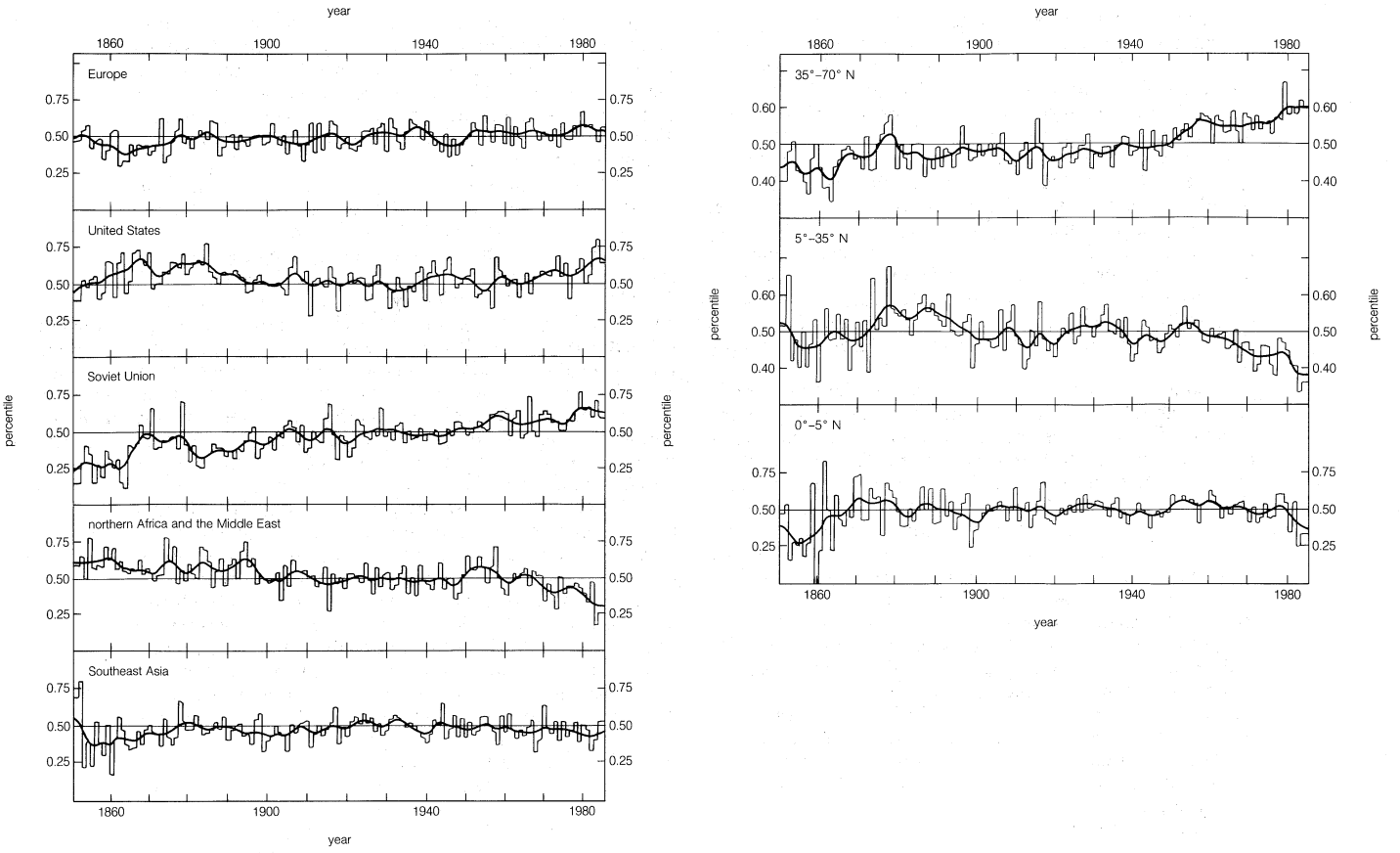



Figure 55: Variation of precipitation over land areas in the Northern Hemisphere from 1850 to 1985 (see text).

From R.S. Bradley *et al.*, "Precipitation Fluctuations Since the Mid-19th Century," *Science*, vol. 237, no. 4811, pp. 171–175 (1987), copyright 1987 by the American Association for the Advancement of Science

The African drought

desiccation affecting much of Africa, but especially the margins of the deserts. Many countries in the Sahelian and Sudanian zones of North Africa experienced a decline in the already inadequate annual rainfall extending from the early 1960s (when it was abundant) until the mid-1980s. Indeed in some areas the decline had not been reversed by the late 1980s. The impact on food supplies was catastrophic, the effects being exacerbated in Ethiopia, Chad, and Sudan by wars. Over the whole continent the two or three decades of drought were tragically unhelpful to the emerging countries of the formerly colonial area. Although no credible cause can be assigned to this fluctuation, which is not yet fully understood, it has been alleged that human misuse of the land has exacerbated a natural fluctuation.

ENSO events recur every three to seven years. The immediate symptom is the flooding of the normally cold Pacific coastal waters off Peru by warm equatorial water from the west. Such El Niño events are accompanied by changes in the trade winds over the entire Pacific and by large rainfall anomalies over Australasia, India, and parts of southeast Asia (the Southern Oscillation component). Recent research, especially at the Scripps Institution of Oceanography at La Jolla, Calif., has shown that major ENSO events are accompanied by worldwide changes in pressure distribution and by anomalies in temperature and rainfall over North America, Australia, and even Eurasia. The dynamics of these huge teleconnections of climate are not understood, but it is recognized that the year-to-year variability of the modern climate is not a local phenomenon. Moreover, such global interconnections have almost certainly existed throughout the Earth's history.

## CAUSES OF CLIMATIC VARIATION

**General considerations.** Explanation—the identification of causes—is always difficult when one discusses complex phenomena, and climate is very complex. In many ways the atmosphere and oceans function as if they were each closed systems, in which change in behaviour is achieved by purely internal processes. In other ways, atmosphere and ocean are linked, and changes in either may arise from interaction with the other. The climatic system, like most other natural systems, behaves in a nonlinear fashion; that is, it does not necessarily change proportionally in response to a change in external forcing. In fact, the system is choked with feedbacks, both positive and negative, that ensure that this nonlinearity will prevail. It might seem, for example, that local temperatures ought to be controlled by local input of solar energy. In no simple way is this true, however. In the first place, the solar heating triggers a return flow of energy to space (whose dependence on temperature is extremely nonlinear). Furthermore, local temperature is much affected by the transport of heat by ocean currents and the winds—*i.e.*, advection of heat. This is why, for example, Sweden and Hudson Bay have such different climates, in spite of being in the same latitudes. Advection depends in turn on the distribution of solar heating over the whole Earth—and the transport terms in the equations governing atmospheric motion are the reason why these equations are nonlinear. In short, cause can never really be separated from effect in the real world.

Nevertheless, it is reasonable to try to answer the question "Why does climate vary?" To do so, the nature of climate must be considered:

The nature of climate

(1) Climate is best described as the prevailing, and hence expected, sequence of states of the atmosphere (loosely, weather) at a given locality. This definition can be extended to cover regions, hemispheres, or even the entire globe. Climatic change is the shift of this sequence to some new apparently lasting state.

(2) The perceived scope of climate has recently been broadened to include a wider range of measurable parameters (*e.g.*, humidity, particle load, ionization, cloud regime, salinity) and to extend over the entire depth and height of the lower atmosphere as well as the surface layers of the ocean.

(3) Climate may be viewed as part of a larger climatic system, an idea closely related to that of ecosystem. Thus, climate pervades not only the atmosphere and ocean but also the soil, ice masses, lakes and rivers, and the living cover of the Earth—all of which interact with the Sun-driven climate of the atmosphere.

(4) The atmosphere and ocean behave in part as chaotic systems. Although they obey well-known laws, they respond to the laws in a bewilderingly complex way. In spite, for example, of reliably periodic forcing by the Sun and tides, the sea and air appear never to repeat themselves precisely, showing instead a preference for nonperiodic behaviour. Tiny events—in principle even the batting of a butterfly's wings—can amplify into consequences that simply cannot be predicted from the known governing laws.

In some ways the wonder is not that weather and climate are so variable but that regularities do exist behind the complex facade. Nevertheless, it is clear that climate is not just a randomly behaving system. Distinct rhythms and spatial patterns recur, and their signals can be isolated from the noise so evident in climatic statistics. In describing climates, therefore, it is necessary to define not only the mean, modal, or median values of the appropriate parameters, such as temperature, pressure, precipitation, and wind velocity, but also the frequencies and characteristics of the variability, whether it is periodic, nonperiodic, or quasi-periodic. It is vital to recognize that variability within the chosen averaging period is as much part of climate as the mean values of each parameter. Thus the year-to-year variability of rainfall in the Great Plains of the United States or in Sahelian Africa is quite as important (and deserving of explanation) as the mean value of rainfall. In economic terms the variability may be the most important aspect.

Climatic interdependence

Another component of climate requiring definition is interdependence. Correlations often exist between variables as measured locally, for example, between precipitation and humidity or between wind direction and visibility. Teleconnections, or correlations between variables measured in different places, also exist. Finally, a long time series of a quantity such as temperature often displays autocorrelation; *i.e.,* its present behaviour is related to its behaviour at other times. Some of these correlations are of fundamental importance. For example, the frictional drag of the winds on the sea surface, which partly governs the oceanic circulation, depends on the covariance (a measure of correlation) of vertical and horizontal components of the wind and on similar covariances within the sea.

Explanation of climatic change, if it is to be attempted fundamentally, must take these points into consideration. If a popular theory of climatic change, such as the Milankovitch orbital variation theory, can only explain part of the observed change (as is commonly the case), it is necessary to look beyond it, until most of that change is understood. Theory, moreover, should be expressed in quantitative terms. Qualitative guesses no longer appeal (physicists call such guesses "hand-waving"), although they still have a place in the generation of hypotheses.

**Identified causes.** Theory suggests (and evidence confirms) that the following may be causes of certain types of climatic variation.

*Variations in the solar constant.* Calculations suggest that a 1 percent change in the solar constant can produce a change in surface global temperature in the range of 1° to 1.5° C. Recent observations show that there have been minor aperiodic changes (±0.2 percent) in the solar constant but none sufficient to account for observed intervening fluctuations (on the order of ±1° C since the Little Ice Age).

*Variations in the Earth's orbit.* As described above, variations in the Earth's orbit around the Sun can cause variations in the distribution and seasonality of solar radiation, which in turn may result in climatic changes. The orbital variation effect has been identified clearly in the geologic, glaciologic, and biotic record of the past 300,000 years, but it is less successful in explaining climatic conditions dating farther back in time. Orbital variations, moreover, can only be used to explain fluctuations on the order of 1,000 years or more, because all three components—precession of the equinoxes, obliquity, and eccentricity—work on long, predictable time scales. It is generally agreed that the waxing and waning of ice sheets and variations of sea level can be related to these parameters in the past 300,000 years, as can many aspects of Holocene variation, notably the monsoons.

*Variations in atmospheric composition.* The concentration of carbon dioxide in the Earth's atmosphere is known to have varied greatly in geologic time and also in Pleistocene, Holocene, and Recent times, including the past century. It is also known that times of increased carbon dioxide levels have coincided with times of higher temperatures. Although it is clear that a close correlation exists between carbon dioxide and air temperature, it is still not known whether increases in carbon dioxide cause temperature increases or vice versa. Theory suggests, however, that carbon dioxide (as well as other trace gases in the atmosphere) allows short-wave energy (visible light) from the Sun to reach Earth relatively unobstructed but hinders long-wave energy (infrared radiation) reradiating from the Earth's surface from escaping back into space and that this trapping of energy results in a rise in temperature. Variations in climate during the next century are expected to be dominated by this so-called greenhouse effect.

Relation between carbon dioxide and air temperature

*Volcanic dust loading of the atmosphere.* Explosive eruptions (*e.g.,* El Chichón, Mex., 1982) are known to introduce material into the stratosphere that can remain in suspension for one to three years. This dust veil can reduce solar input, possibly leading to a fall in temperature. The biggest explosive eruptions (*e.g.,* Mount Tambora, 1815; Krakatoa, 1883; both in Indonesia) are known to have produced marked cooling for a few years after the event. It has been argued that a high frequency of explosive eruptions may account for past cool periods. Such eruptions may also have significantly modified the solar-orbital control of late Pleistocene and Holocene climates.

*Variations in the distribution of land and sea due to crustal changes.* Mountain building, continental uplift, and the drifting of continents have profoundly influenced world climate. They operate on the scale of geologic periods, requiring many millions of years to complete their effects. An example of a climatic change caused by topographic variations is the creation of the cold deep waters of the ocean, which was favoured by such changes as the opening of the Antarctic Ocean (as continental drift separated Australia and Antarctica), the rise of the Isthmus of Panama, and the closing off of the Tethys, the ancient ancestor of the Mediterranean Sea.

Together, these five theories provide possible explanations for much of the known climatic variation on the global scale and on many of the observed time scales. Yet, for the most part they do so only in a crude way, usually in relation to mean annual global or polar air temperatures, sea-ice and land-ice distribution, the position of oceanic fronts, and other integrators of long-term climatic features. They have little to say, in particular, about the internal variability of climate.

**Contemporary areas of research.** Meteorologists and geologists recognize, in fact, that major gaps still exist in the explanation of climatic variation. On the geologic time scale, for example, it has proved difficult to explain the approximately 100,000-year recurrence period for the cold, glacial episodes of the Pleistocene. This period resembles that of the dominant mode of the orbital eccentricity of the Earth, but this aspect of orbital variation is thought inadequate to explain the glacial episodes so vividly shown in the deep-sea core record.

Climatic modeling

Very much closer to hand, the numerical modeling of atmospheric behaviour (see below *Numerical weather prediction [NWP] models*) has proved effective in the prediction of weather events as much as five or 10 days ahead, starting from a specified initial state. Such models try to predict actual events but lose that ability quickly as the chaotic nature of the dynamics blurs the outcome. Nevertheless, the models can be modified into atmospheric general circulation models (AGCMs). These AGCMs, which do not start from a specified initial state (the weather of the moment) and which do not predict actual events, can

be used, with specified boundary conditions (such as carbon dioxide content, sea surface conditions, relief features, and continental coastal outlines), to create the statistics of hypothetical climates—including that of the present. Reference was made above to the use of one such model in analysis of the COHMAP results for the Holocene epoch.

AGCMs and their cousins, oceanic general circulation models (OGCMs), can be refined to the point where they predict not only mean temperature, pressure, and rainfall but also the annually repeated seasons and nonperiodic variability. Unfortunately, the several models now operational do not compare well with one another in their predictions of regional detail. They also usually differ from observed fact on this scale to an unacceptable degree, especially as regards precipitation.

The most difficult part of climate to predict is unfortunately the one of maximum economic import—variations from month to month, from year to year, and from decade to decade. Marked and persistent anomalies occur on all these time scales (and have probably existed throughout geologic history), but at present they can be neither explained nor predicted. It is thought that they may be related to such periodic or quasi-periodic fluctuations as the so-called "Chandler wobbles" of the Earth's axis, sunspot cycles, the ENSO phenomenon, and the so-called Julian–Madden waves in the tropical circulation, but until the mechanics of such phenomena are more completely understood, no real theories can be advanced.

Finally, it remains uncertain whether or not the atmosphere–ocean system responds exclusively to physical controls. It is possible that biologic regulation is also at work. The idea that the living cover of the Earth interacts with the atmosphere is well established, but meteorologists and glaciologists have generally seen climate as the control and the rest of the ecosystem as response. In 1972 the British chemist James Lovelock challenged this view when he put forward the Gaia hypothesis. Briefly, the Gaia hypothesis suggests that the climate of the Earth has for eons been modulated—perhaps even controlled—by the biota, which regulates the concentration of atmospheric carbon dioxide and other organically derived substances so as to keep temperature and precipitation at advantageous levels. In effect the Earth is, in Lovelock's most recent restatement (1988), a living organism, with self-regulating processes (homeostasis) capable of ensuring the survival of a life-sustaining global climate. (This matter is discussed at greater length below in the section *Climate and life.*) Lovelock's view has been widely disputed, but it is recognized that the role of life in climate and climatic change cannot be lightly dismissed. (F.K.H.)

# CLIMATIC CLASSIFICATION

## General considerations

The climate of an area, as previously noted, is the synthesis of the weather conditions that have prevailed there over a long period of time (usually 30 years). This synthesis involves both averages of the climatic elements and measurements of variability (such as extreme values and probabilities). Climate is a complex, abstract concept involving data on temperature, humidity, precipitation type and amount, wind speed and direction, atmospheric pressure, sunshine, cloud types and coverage, and such weather phenomena as fog, thunderstorms, and frost and the relationships among them. As such, no two localities on Earth may be said to have exactly the same climate. Nevertheless, it is readily apparent that, over restricted areas of the planet, climates vary within a limited range and that climatic regions are discernible within which some uniformity is apparent in the patterns of climatic elements. Moreover, widely separated areas of the world possess similar climates, which tend to recur in similar geographic relationships to each other. This symmetry and organization of the climatic environment suggests an underlying worldwide regularity and order in the phenomena causing climate (*e.g.,* patterns of radiation, atmospheric pressure, winds, fronts, and air masses), which were discussed in earlier sections.

Climate classification is an attempt to formalize this process of recognizing climatic similarity, of organizing, simplifying, and clarifying the vast amount of weather data collected by the meteorological services of the world, and of systematizing the long-term effects of interacting climatic processes to enhance scientific understanding of climates. Users of climate classifications should be aware of the limitations of the procedure, however.

Limitations of climate classification

First, climate is a multidimensional concept, and it is not an obvious decision as to which of the many observed weather variables should be selected as the basis of the classification. This choice must be made on a number of grounds, both practical and theoretical. For example, using too many different elements opens up the possibilities that the classification will have too many categories to be readily interpreted and that many of the categories will not correspond to real climates. Moreover, measurements of many of the elements of climate are not available for large areas of the world or have been collected for only a short time. The major exceptions are temperature and precipitation data, which are available almost universally and have been recorded for extended periods of time.

The choice of variables also is determined by the purpose of the classification (*e.g.,* to account for distribution of natural vegetation, to explain soil formation processes, or to classify climates in terms of human comfort). The variables relevant in the classification will be determined by this purpose, as will the threshold values of the variables chosen to differentiate climatic zones.

A second difficulty results from the generally gradual nature of changes in the climatic elements over the Earth's surface. Except in unusual situations due to mountain ranges or coastlines, temperature, precipitation, and other climatic variables tend to change only slowly over distance. As a result, climate types tend to change imperceptibly as one moves from one locale to an adjacent one. Choosing a set of criteria to distinguish one climatic type from another is thus equivalent to drawing a line on a map to distinguish the climatic region possessing one type from that having the other. While this is in no way different from many other classification decisions that one makes routinely in daily life, it must always be remembered that boundaries between adjacent climatic regions are placed somewhat arbitrarily through regions of continuous, gradual change and that the areas defined within these boundaries are far from homogeneous in terms of their climatic characteristics.

Most classification schemes are intended for global- or continental-scale application and define regions that are major subdivisions of continents hundreds to thousands of kilometres across. These may be termed macroclimates. Not only will there be slow changes (from wet to dry, hot to cold, etc.) across such a region as a result of the geographic gradients of climatic elements over the continent of which the region is a part, but there will exist mesoclimates within these regions associated with climatic processes occurring at a scale of tens to hundreds of kilometres that are created by elevation differences, slope aspect, bodies of water, differences in vegetation cover, urban areas, and the like. Mesoclimates, in turn, may be resolved into numerous microclimates, which occur at scales of less than 0.1 kilometre, as in the climatic differences between forests, crops, and bare soil, at various depths in a plant canopy, at different depths in the soil, on different sides of a building, and so on.

Mesoclimates and microclimates

These limitations notwithstanding, climate classification plays a key role as a means of generalizing the geographic distribution and interactions among climatic elements, of identifying mixes of climatic influences important to various climatically dependent phenomena, of stimulating the search to identify the controlling processes of climate, and, as an educational tool, to show some of the ways in

which distant areas of the Earth are both different from and similar to one's own home region.

## Approaches to climatic classification

The earliest known climatic classifications were those of classical Greek times. Such schemes generally divided the Earth into latitudinal zones based on the significant parallels of 0°, 23.5°, and 66.5° of latitude and on the length of day. Modern climate classification has its origins in the mid-19th century, with the first published maps of temperature and precipitation over the Earth, which permitted the development of methods of climate grouping that used both variables simultaneously.

Principal types of classification schemes

Many different schemes of classifying climate have been devised (more than 100), but all of them may be broadly differentiated as either empiric or genetic methods. This distinction is based on the nature of the data used for classification. Empiric methods make use of observed climatic data, such as temperature, humidity, and precipitation, or simple quantities derived from them (*e.g.,* evaporation). In contrast, a genetic method classifies climate on the basis of its causal elements, the activity and characteristics of all factors (air masses, circulation systems, fronts, jet streams, solar radiation, topographic effects, and so forth) that give rise to the spatial and temporal patterns of climatic data. Hence, while empiric classifications are largely descriptive of climate, genetic methods are (or should be) explanatory. Unfortunately, genetic schemes, while scientifically more desirable, are inherently more difficult to implement because they do not use simple observations of the atmosphere. As a result, such schemes are both less common and less successful overall. Moreover, the regions defined by the two types of classification schemes do not necessarily correspond; in particular, it is not uncommon for similar climatic forms resulting from different climatic processes to be grouped together by many common empiric schemes.

Empiric methods may be subdivided further based on the manner in which the climatic data are used. Some approaches attempt to employ climatic data directly and to group patterns of their statistics on some "rational" basis, drawing guidance from the numbers themselves. More commonly, however, climates have been classified on the basis of the response of some phenomenon to climate. For example, climatic types have been differentiated on the basis of natural vegetation distributions, human comfort, soil formation, rock weathering, and myriad other factors whose intensity, nature, or distribution are felt to be controlled or influenced by climate. As such, criteria for differentiating different climatic groups are determined by the assumed behaviour of the phenomenon chosen to relate to climate rather than the statistical properties of the data themselves. The distinction between these two empiric approaches is not as clear as might at first be thought, however. While the thresholds chosen to differentiate groups might be determined by the characteristics of the climatic statistics themselves, the choice of variables on which to base the classification is made by its author. That choice is made on the basis of some perceived significance in the variables selected. This perception implies some subjective view of the nature of climate or its significance to some dependent phenomenon.

### GENETIC CLASSIFICATIONS

Major types of genetic classification

Genetic classifications group climates by their causes. Among such methods, three types may be distinguished: (1) those based on the geographic determinants of climate, (2) those based on the surface energy budget, and (3) those derived from air-mass analysis.

In the first class are a number of schemes (largely the work of German climatologists) that categorize climates according to such factors as latitudinal control of temperature, continentality versus oceanicity, location with respect to pressure and wind belts, and effects of mountains. These classifications all share a common shortcoming: they are qualitative, so that climatic regions are designated in a subjective manner rather than as a result of the application of some rigorous differentiating formula.

An interesting example of a method based on the energy balance of the Earth's surface is the 1970 classification of Werner H. Terjung, an American geographer. His method utilizes data for more than 1,000 locations worldwide on the net radiation received at the surface, the available energy for evaporating water and for heating the air and subsurface. The annual patterns are classified according to the maximum energy input, the annual range in input, the shape of the annual curve, and the number of months with negative magnitudes. The combination of characteristics for a location is represented by a label consisting of several letters with defined meanings, and regions having similar net radiation climates are mapped.

Probably the most extensively used genetic systems, however, are those that employ air-mass concepts. Air masses are large bodies of air that in principle possess relatively homogeneous properties of temperature, humidity, etc., in the horizontal. Weather on individual days may be interpreted in terms of these features and their contrasts at fronts, which suggests that climate, as a synthesis of weather, can be treated likewise. While air masses are the proximate causes of climate rather than the ultimate controls (which are the factors that gave the air masses their characteristics), this approach can still be properly regarded as genetic.

Genetic systems employing air-mass concepts

Two American geographer-climatologists have been most influential in classifications based on air mass. In 1951 Arthur N. Strahler described a qualitative classification based on the combination of air masses present at a given location throughout the year. Some years later (1968 and 1970), John E. Oliver placed this type of classification on a firmer footing by providing a quantitative framework that designated particular air masses and air mass combinations as "dominant," "subdominant," or "seasonal" at particular locations. He also provided a means of identifying air masses from diagrams of mean monthly temperature and precipitation plotted on a "thermohyet diagram," a procedure that obviates the need for less common upper-air data to make the classification.

### EMPIRIC CLASSIFICATIONS

Most empiric classifications are those that seek to group climates according to their effects on some climate-dependent phenomenon. While many such phenomena have been used in this way, natural vegetation stands out as one of prime importance. The view held by many botanist-climatologists is that natural vegetation functions as an integrator of the characteristics of climate in a region; the vegetation, in effect, is an instrument for measuring climate in the same way that a thermometer measures temperature. That this view is an oversimplified one is undoubtedly true. Nevertheless, it has been a prime motivation of many climatologists, and its preeminence is apparent in the fact that many textbooks and other sources refer to climates using the names of vegetation, as, for example, rain-forest, taiga, or tundra.

Natural vegetation in empiric classification

Wladimir Köppen, a German botanist-climatologist, developed the most popular (but not the first) of these vegetation-based classifications. His aim was to devise formulas that would define climatic boundaries in such a way as to correspond to those of the vegetation zones that were being mapped for the first time during his lifetime. Köppen published his first scheme in 1900 and a revised version in 1918. He continued to revise his system of classification until his death in 1940. Other climatologists have modified portions of Köppen's procedure on the basis of their experience in various parts of the world.

Köppen's classification is based on a subdivision of terrestrial climates into five major types, which are represented by the capital letters A, B, C, D, and E. Each of these climate types except for B is defined by temperature criteria. Type B designates climates in which the controlling factor on vegetation is dryness (rather than coldness). Aridity is not a matter of precipitation alone but is defined by the relationship between the precipitation input to the soil in which the plants grow and the evaporative losses. Since evaporation is difficult to evaluate and is not a conventional measurement at meteorological stations, Köppen was forced to substitute a formula that identifies

The Köppen classification system

aridity in terms of a temperature-precipitation index (*i.e.*, evaporation is assumed to be controlled by temperature). Dry climates are divided into arid (BW) and semiarid (BS) subtypes, and each may be differentiated further by adding a third code, h for warm and k for cold.

As noted above, temperature defines the other four major climate types. These are subdivided, with additional letters again used to designate the various subtypes. Type A climates (the warmest) are differentiated on the basis of the seasonality of precipitation: Af (no dry season), Am (short dry season), or Aw (winter dry season). Type E climates (the coldest) are conventionally separated into tundra (ET) and snow/ice climates (EF). The mid-latitude C and D climates are given a second letter, f (no dry season), w (winter dry), or s (summer dry), and a third symbol (a, b, c, or d [the last subclass exists only for D climates]), indicating the warmth of the summer or the coldness of the winter. Table 6 gives the specific criteria for the Köppen–Geiger–Pohl system of 1953.

The Köppen classification has been criticized on many grounds. It has been argued that extreme events, such as a periodic drought or an unusual cold spell, are just as significant in controlling vegetation distributions as the mean conditions upon which Köppen's scheme is based. It also has been pointed out that meteorological factors other than those used in the classification, such as sunshine and wind, are important to vegetation. Moreover, it has been contended that vegetation can respond only slowly to climate, so that the vegetation zones observable today are in part adjusted to past climates. Many critics have drawn attention to the rather poor correspondence between the Köppen zones and the observed vegetation distribution in many areas of the world. In spite of these and other limitations, the Köppen system remains the most popular climatic classification in use today.

Thornthwaite's classification

A major contribution to climate grouping was made by the American geographer-climatologist C. Warren Thornthwaite in 1931 and 1948. He first used a vegetation-based approach that made use of the derived concepts of temperature efficiency and precipitation effectiveness as a means of specifying climatic effects on vegetation. His second classification retained these concepts in the form of a moisture index and a thermal efficiency index but radically changed the classification criteria and rejected the idea of using vegetation as the climatic integrator, attempting instead to classify "rationally" on the basis of the numerical values of these indices. His 1948 scheme is encountered in many climatology texts, but it has not gained as large a following among a wider audience as the Köppen classification system, perhaps because of its complexity and the large number of climatic regions it defines.

While vegetation-based climate classifications could be regarded as having relevance to human activity through what they may indicate about agricultural potential and natural environment, they cannot give any sense of how human beings would feel within the various climate types. Terjung's 1966 scheme was an attempt to group climates on the basis of their effects on human comfort. The classification makes use of four physiologically relevant parameters: temperature, relative humidity, wind speed, and solar radiation. The first two are combined in a comfort index to express atmospheric conditions in terms perceived as extremely hot, hot, oppressive, warm, comfortable, cool, keen, cold, very cold, extremely cold, and ultra cold. Temperature, wind speed, and solar radiation are combined in a wind effect index expressing the net effect of wind chill and addition of heat to the human body by solar radiation. These indices are combined for different seasons in different ways to express how humans feel in various geographic areas on a yearly basis. Terjung visualized that his classification would find applicability in medical geography, climatological education, tourism, housing, clothing, and as a general analytical tool.

Terjung's comfort index

Many other specialized empirical classifications have been devised. For example, there are those that differentiate between types of desert and coastal climates, those that account for different rates of rock weathering or soil formation, and those based on the identification of similar agricultural climates.

**Table 6: Criteria for Classifying Major Climatic Types According to the Köppen–Geiger–Pohl Scheme (1953)**

| letter symbol | | | criterion |
|---|---|---|---|
| 1st | 2nd | 3rd | |
| A | | | temperature of coolest month 18° C or higher |
| | f | | precipitation in driest month at least 60 mm |
| | m | | precipitation in driest month less than 60 mm but equal to or greater than $100-(r/25)$* |
| | w | | precipitation in driest month less than 60 mm and less than $100-(r/25)$ |
| B† | | | 70% or more of annual precipitation falls in the summer half of the year and $r$ less than $20t+280$, or 70% or more of annual precipitation falls in the winter half of the year and $r$ less than $20t$, or neither half of the year has 70% or more of annual precipitation and $r$ less than $20t+140$‡ |
| | W | | $r$ is less than one-half of the upper limit for classification as a B type (see above) |
| | S | | $r$ is less than the upper limit for classification as a B type but is more than one-half of that amount |
| | | h | $t$ equal to or greater than 18° C |
| | | k | $t$ less than 18° C |
| C | | | temperature of warmest month greater than or equal to 10° C, and temperature of coldest month less than 18° C but greater than −3° C |
| | s | | precipitation in driest month of summer half of the year is less than 30 mm and less than one-third of the wettest month of the winter half |
| | w | | precipitation in driest month of the winter half of the year less than one-tenth of the amount in the wettest month of the summer half |
| | f | | precipitation more evenly distributed throughout year; criteria for neither s nor w satisfied |
| | | a | temperature of warmest month 22° C or above |
| | | b | temperature of each of four warmest months 10° C or above but warmest month less than 22° C |
| | | c | temperature of one to three months 10° C or above but warmest month less than 22° C |
| D | | | temperature of warmest month greater than or equal to 10° C, and temperature of coldest month −3° C or lower |
| | s | | same as for type C |
| | w | | same as for type C |
| | f | | same as for type C |
| | | a | same as for type C |
| | | b | same as for type C |
| | | c | same as for type C |
| | | d | temperature of coldest month less than −38° C (d designation then used instead of a, b, or c) |
| E | | | temperature of warmest month less than 10° C |
| | T | | temperature of warmest month greater than 0° C but less than 10° C |
| | F | | temperature of warmest month 0° C or below |

*In the formulas above, $r$ is average annual precipitation total (mm) and $t$ is average annual temperature (° C). All other temperatures are monthly means (° C), and all other precipation amounts are mean monthly totals (mm). †Any climate that satisfies the criteria for designation as a B type is classified as such, irrespective of its other characteristics. ‡The summer half of the year is defined as the months April–September for the Northern Hemisphere and October–March for the Southern Hemisphere.

## World distribution of major climatic types

The following discussion of the climates of the world is based on groupings of Köppen's climatic types. It should be read in conjunction with Table 6 (for the specific criteria defining each type), Figure 56 (for the geographic extent of each type), and Figure 57 (which provides sample annual temperature and precipitation data for many of the climates discussed).

TYPE A CLIMATES

Köppen's A climates are found in a nearly unbroken belt around the Earth at low latitudes, mostly within 15° N and S. Their location within a region in which available net solar radiation is large and relatively constant from month to month ensures both high temperatures (generally in excess of 18° C) and a virtual absence of thermal seasons. Typically, the temperature difference between day and night is greater than that between the warmest and the coolest month, the opposite of the situation in mid-latitudes. The terms winter and summer have little meaning, but in many locations annual rhythm is provided by the occurrence of wet and dry seasons. Type A climates are controlled mainly by the seasonal fluctuations of the trade winds, the intertropical convergence zone (ITCZ), and the Asian monsoon (see above *Wind*).

**Wet equatorial climate (Af).** Within about 12° latitude of the Equator lies a region of consistently high temperatures (around 30° C), with plentiful precipitation (150–1,000 centimetres), heavy cloud cover, and high humidity, with very little annual temperature variation. Such regions lie within the influence of the ITCZ in all months; the converging, ascending air spawns convectional thunderstorm activity with much of the rainfall occurring in late afternoon or early evening when the atmosphere is most susceptible to thunderstorms. While precipitation is profuse in all months, variations do occur in response to the precise location of the ITCZ, drier months resulting when the circulation brings the region to its periphery. Other zones of Af climate are found beyond the usual margins of ITCZ activity, in coastal Madagascar, southeast coastal Brazil, and much of Central America and western Colombia, where trade winds blow onshore all year to encounter coastlines backed by mountain barriers that stimulate the formation of precipitation as warm, moist tropical air is forced to ascend and cool. Such regions also may receive precipitation from tropical disturbances, including hurricanes.

Effect of the ITCZ

**Tropical monsoon and trade-wind littoral climates (Am).** These climates resemble the Af in most characteristics, with small annual temperature ranges, high temperatures, and plentiful precipitation (often more than Af climates in annual total). They differ from the latter, however, in that they exhibit a short dry season, usually in the low-sun ("winter") season. The highest temperatures generally occur at the end of this clear spell. Two distinct processes can give rise to Am climate types. The largest areas, mostly in southern and southeastern Asia, result from the Asian monsoon circulation that brings convective and orographic precipitation in the summer when warm, moist maritime tropical air moves over land to converge into the low-pressure zone north of the Himalayas. In winter, by contrast, cool, dry air diverges out of the Siberian anticyclone to the north, bringing a cooler, drier, and clearer period of variable length. In the Americas and in Africa, continents that are not large enough to generate a true monsoon circulation, Am climates are of the trade-wind littoral variety. These areas receive precipitation on narrow coastal strips through orographic effects as the moist air of the trade winds ascends mountain chains. Seasonal migrations and changes in the intensity of these winds give rise to short, moderately dry seasons. Summer precipitation may be enhanced by tropical disturbances traveling in the trade winds.

**Tropical wet-dry climate (Aw).** This climate has distinct wet and dry seasons, with most of the precipitation occurring in "summer." Total amounts of rainfall are less than in the previous two climate types (50–175 centimetres), most of which occur in convectional thunderstorms. The dry season is longer than in the Am climates and becomes progressively longer as one moves poleward through the region. Temperatures are high throughout the year but show a greater range than Af and Am climates (19°–20° C in winter and 24°–27° C in summer). Throughout most of the region, the cause of the seasonal cycle is the shift in the tropical circulation throughout the year. During the high-sun season, the ITCZ moves poleward and brings convergent and ascending air to these locations, which stimulates convective rainfall. During the low-sun season, the convergence zone moves off to the winter hemisphere and is replaced by the periphery or core of the subtropical anticyclone, with its subsiding, stable air resulting in a period of dry, clear weather, the intensity and length of which depend on latitude.

TYPE B CLIMATES

Arid and semiarid climates cover about a quarter of the Earth's land surface, mostly between 50° N and 50° S, but they are mainly found in the 15°–30° latitude belt in both hemispheres. They exhibit low precipitation, great variability in precipitation from year to year, low relative humidity, high evaporation rates (when water is available), clear skies, intense solar radiation, and high wind speeds.

Arid and semiarid climates

**Tropical and subtropical desert climate (BWh, part of BWk).** Most of the tropical, true desert (BW) climates of the Earth occur between 15° and 30° latitude, at the poleward end of the Hadley cell circulation (see above). These regions are dominated in all months by the subtropical anticyclone, with its descending air, elevated inversions, and clear skies. This is an atmospheric environment that

Adapted from Arthur N. Strahler, *Physical Geography*, third edition; John Wiley & Sons, Inc.

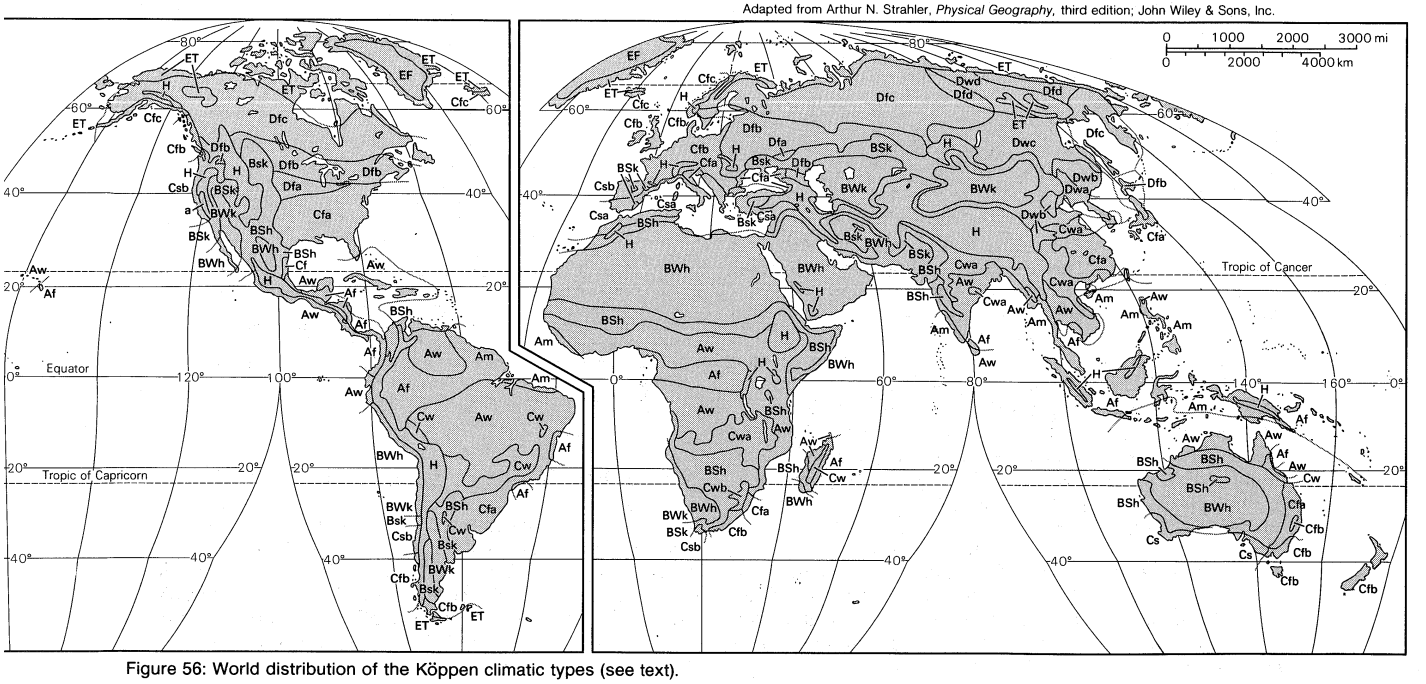



Figure 56: World distribution of the Köppen climatic types (see text).

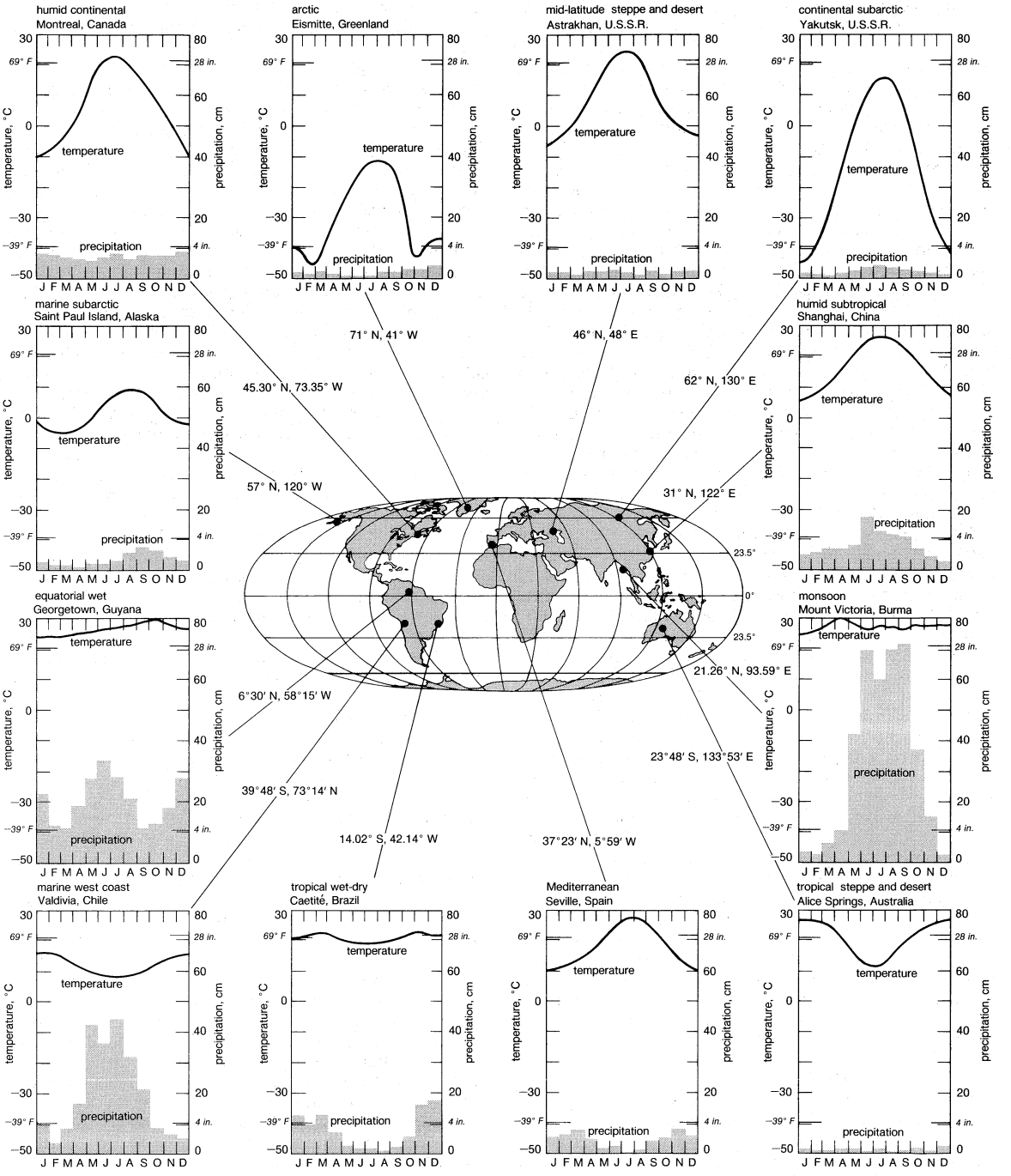



Figure 57: Annual temperature and precipitation patterns for several locations with different climatic conditions.

From W.M. Marsh and J. Dozier, *Landscape: An Introduction to Physical Geography*, copyright © 1981 John Wiley & Sons, Inc.; reprinted by permission of John Wiley & Sons, Inc.

inhibits precipitation. The most extreme arid areas also are far removed from sources of moisture-bearing winds in the interiors of continents and are best developed on the western sides of continents, where the subtropical anticyclone shows its most intense development. An exception to the general tendency for aridity to be associated with subsidence is in the so-called Horn of Africa region, where the dryness of Somalia is caused more by the orientation of the landmass in relation to the atmospheric circulation. Both the high- and low-sun monsoonal winds blow parallel to the coast, so that moisture-laden maritime air can penetrate over land only infrequently. In most low-latitude deserts, cloud cover is uncommon (fewer than 30 days per year have clouds in some areas). Precipitation amounts are mostly in the range 0–25 centimetres, although the unreliability of precipitation is more significant than the small totals. Average figures have little meaning; a location with a 10-year mean of five centimetres, for example, might have received 50 centimetres in one year as a result of an unusual intrusion of moist air, followed by nine years with no measurable precipitation.

Temperatures are high, with monthly means in the range 21°–32° C. Daily temperature variations are extreme. Ranges of 35° C are not unknown when daytime maxima in excess of 40° C are followed by a rapid nocturnal temperature drop brought about by the limited capacity of the dry, cloudless desert air to emit infrared radiation to the ground to offset radiation loss from the surface at night. The highest air temperatures recorded on Earth have been in the BWh regions; for example, Death Valley in the western United States has reached 57° C, while al-ʿAzīzīyah in Libya has had a recorded high of 58° C. Actual surface temperatures may reach 82° C on dry sand under intense sunshine.

An interesting variant of tropical and subtropical deserts are the so-called West Coast Desert areas found on the western coastal margins of the regions discussed above (*e.g.,* in the Sonoran Desert of North America, the Peru and Atacama deserts of South America, and the Sahara [Moroccan part] and Namib deserts of Africa). These areas are much cooler than their latitude would suggest (monthly mean temperatures of only 15°–21° C), and parts are classified as BWk in Köppen's scheme. The cooling results from air flow off adjacent coastal waters where upwelling of the ocean gives rise to cold currents. Deserts of this sort are subject to frequent fog and low-level clouds; yet they are extremely arid. Some parts of the Atacama Desert, for example, have recorded no precipitation for 20 years.

**Tropical and subtropical steppe climate (BSh).** The low-latitude semiarid (or steppe) climate occurs primarily on the periphery of the true deserts treated above. It is transitional to the Aw climate on the equatorward side (showing a summer rainfall maximum associated with the ITCZ and a small annual temperature range) and to the Mediterranean climate on its poleward margin (with a cooler, wetter winter resulting from the higher latitude and mid-latitude frontal cyclone activity). Annual precipitation totals are greater than in BW climates (38–63 centimetres). Yearly variations in amount are not as extreme as in the true deserts but are nevertheless large.

**Mid-latitude steppe and desert climate (BSk, part of BWk).** Although these climates are contiguous with the tropical dry climates of North and South America and of central Asia, they have different origins. Cool true deserts extend to 50° latitude and cool steppes reach nearly 60° N in the Canadian Prairies, well beyond the limits of the subtropical anticyclone. These climates owe their origins to locations deep within continental interiors, far from the windward coasts and sources of moist, maritime air. Remoteness from sources of water vapour is enhanced in some regions (*e.g.,* the Great Plains of the United States) by mountain barriers upwind. Temperature conditions are extremely variable, with annual means decreasing and annual ranges increasing poleward. In the higher latitudes, winters are severely cold, with meager precipitation (much of it in the form of snow) associated with polar and arctic air in frontal cyclones. Summer precipitation is more often convective, arriving in the form of scattered thunderstorm activity brought about by irregular incursions of moist air. Both BWk and BSk climates in the mid-latitudes owe their origins to these mechanisms, but the steppe type tends to be located peripheral to the true desert, either adjacent to the moister C and D climates or at the poleward extent of the range, where reduced evaporation under cooler conditions makes more of the scarce precipitation available as soil moisture for plant growth.

### TYPE C AND D CLIMATES

Through a major portion of the middle and high latitudes (mostly from 25° to 70° N and S) lies a group of climates classified within the Köppen scheme as C and D types. Most of these regions lie beneath the upper-level, mid-latitude westerlies throughout the year, and it is in the seasonal variations in location and intensity of these winds and their associated features that the explanation of their climatic character must be sought. During summer, the polar-front and its jet stream move poleward, and air masses of tropical origin are able to extend to high latitudes. During winter, as the circulation moves equatorward, tropical air retreats and cold polar outbreaks influence weather, even within the subtropical zone. The relative frequency of these air masses of different origins varies gradually from low to high latitude and is largely responsible for the observed temperature change across the belt (which is most marked in winter). The air masses interact in the frontal systems commonly found embedded within the traveling cyclones that lie beneath the polar-front jet stream. Ascent induced by convergence into these low-pressure cells and by uplift at fronts induces precipitation, the main location of which shifts with the seasonal circulation cycle. Other important sources of precipitation are convection, mainly in tropical air, and forced uplift at mountain barriers. Monsoon effects modify this general pattern in eastern Asia, while the subtropical anticyclone plays a role in the explanation of climate on the western sides of the continents in the subtropics.

Principal sources of precipitation

**Humid subtropical climate (Cfa, Cwa).** These climates are found on the eastern sides of the continents between 20° and 35° N and S latitude. Most show a relatively uniform distribution of precipitation throughout the year (the Cfa types), with totals in the range 75–150 centimetres. In summer these regions are largely under the influence of moist, maritime airflow from the western side of the subtropical anticyclonic cells over low-latitude ocean waters. Temperatures are high; the warmest months generally average about 27° C, with mean daily maxima from 30 ° to 38° C and warm, oppressive nights. Summers are usually somewhat wetter than winters, with much of the rainfall coming from convectional thunderstorm activity; tropical cyclones also enhance warm-season rainfall in some regions. The coldest month is usually quite mild (5°–12° C), although frosts are not uncommon, and winter precipitation is derived primarily from frontal cyclones along the polar front. In North America the spring and early summer seasons, when the front begins its northward return, are notorious for the outbreak of tornadoes associated with frontal thunderstorms along the zone of interaction between tropical and polar air. In eastern and southern Asia the monsoon influence results in a modified humid subtropical climate (Cwa) that has a clearly defined dry winter when air diverges from the Siberian anticyclone and the polar front and cyclone paths are deflected around the region. These areas generally lie on the poleward side of Am and Aw climates and exhibit a somewhat larger annual temperature range than Cfa types. Winters are sunny and rather cool. Annual precipitation totals average about 100 centimetres but vary from 75 centimetres to 200 centimetres.

High incidence of tornadoes

**Mediterranean climate (Csa, Csb).** Between about 30° and 45° latitude on the western sides of the continents is found a series of climates that show the unusual combination of hot, dry summers and cool, wet winters. Poleward extension and expansion of the subtropical anticyclonic cells over the oceans bring subsiding air to the region in summer, with clear skies and high temperatures. When the anticyclone moves equatorward in winter, it is replaced by traveling, frontal cyclones with their attendant precipitation. Annual temperature ranges are generally smaller than those found in the Cfa climates, since locations on the western sides of continents are not well positioned to receive the coldest polar air, which develops over land rather than over the ocean. Mediterranean climates also tend to be drier than humid subtropical ones, with precipitation totals ranging from 35 to 90 centimetres; the lowest amounts occur in interior regions adjacent to the semiarid steppe climates. Some coastal locations (*e.g.,* southern California in the western United States) exhibit relatively cool summer conditions and frequent fogs where cold offshore currents prevail. Only in Europe where the latitude for this climate type fortuitously corresponds to an ocean basin (that of the Mediterranean, from which this climate derives its name) does the Cs type extend eastward away from the coast for any significant distance.

**Marine west coast climate (Cfb, Cfc).** Poleward of the Mediterranean climate region on the western sides of the continents, between 35° and 60° N and S latitude are regions that exhibit ample precipitation in all months. Unlike their equatorward neighbours, these areas are located beyond the farthest poleward extent of the subtropical anticyclone, and they experience the mid-latitude westerlies and traveling frontal cyclones all year. Precipitation totals vary somewhat throughout the year in response to the changing location and intensity of these storm systems, but annual accumulations generally range from 50 to 250 centimetres, with local totals exceeding 500 centimetres where onshore winds encounter mountain ranges. Not only is precipitation plentiful but it is also reliable and frequent. Many areas have rainfall more than 150 days per year, although the precipitation is often of low intensity. Fog is common in autumn and winter, but thunderstorms, tornadoes, and hurricanes are infrequent. Strong gales with high winds, however, may be encountered in

Equable climates

winter. These are equable climates with few extremes of temperature. Annual ranges are rather small (10°–15° C), about half those encountered farther to the east in the continental interior at the same latitude. Mean annual temperatures are usually 7°–13° C in lowland areas, the winters are mild, and the summers are relatively moderate, rarely having monthly temperatures above 20° C.

In North and South America, Australia, and New Zealand, north–south mountain ranges backing the west coasts of the landmasses at these latitudes confine the marine west coast climate to relatively narrow coastal strips (but enhance precipitation). By contrast, in Europe the major mountain chains (the Alps and Pyrenees) run east–west, permitting Cfb and Cfc climates to extend inland some 2,000 kilometres into East Germany and Poland.

**Humid continental climate (Dfa, Dfb, Dwa, Dwb).** The D climates are primarily northern hemispheric phenomena, since landmasses are absent at the significant latitudes in the Southern Hemisphere. The humid continental subgroup occupies a region between 30° and 60° N in central and eastern North America and Asia in the major zone of conflict between polar and tropical air masses. These regions exhibit large seasonal temperature contrasts with hot summers and cold winters. Precipitation tends to be ample throughout the year in the Df section, being derived both from frontal cyclones and, in summer months, from convectional showers when maritime tropical air pushes northward behind the retreating polar front. Many areas show a distinct summer precipitation maximum because of this convective activity, although more uniform patterns are not uncommon. Severe thunderstorms and tornadoes are an early summer occurrence when the polar front is in the southern margin of the Dfa region. Winter precipitation often occurs in the form of snow, and a continuous snow cover is established for from one to four months in many parts of the region, especially in the north. This snow often arrives in conjunction with high winds from an intense frontal cyclone, giving rise to a blizzard.

Large contrasts in seasonal temperatures

Winters tend to be cold but are subject to occasional frigid or mild spells brought about by periodic incursions of arctic or tropical air. Indeed the changeable nature of weather in all seasons is a characteristic feature of the climate, especially in such areas as the eastern United States and Canada where there are few topographic barriers to limit the exchange of air masses between high and low latitudes. Mean temperatures are typically below freezing from one to several months, and the frost-free season varies from fewer than 150 to 200 days per year. Annual precipitation totals range from 50 to 125 centimetres, with higher amounts in the south of the region and in the uplands.

In eastern Asia (Manchuria and Korea), a monsoonal variant of the humid continental climate (Dwa, Dwb) occurs. This climate type has a pronounced summer precipitation maximum and a cold, dry winter dominated by continental polar air diverging out of the nearby Siberian anticyclone.

**Continental subarctic climate (Dfc, Dfd, Dwc, Dwd).** North of the humid continental climate, from about 50° to 70° N, in a broad swath extending from Alaska to Newfoundland in North America and from northern Scandinavia to Siberia in Eurasia, lie the continental subarctic climates. These are regions dominated by the winter season, a long, bitterly cold period with short, clear days, relatively little precipitation (mostly in the form of snow), and low humidity. In Asia the Siberian anticyclone, the source of continental polar air, dominates the interior, and mean temperatures 40°–50° C below freezing are not unusual. The North American representative of this climate is not as severe but is still profoundly cold. Mean monthly temperatures are below freezing for six to eight months, with an average frost-free period of only 50–90 days per year, and snow remains on the ground for many months. Summers are short and mild, with long days and a prevalence of frontal precipitation associated with maritime tropical air within traveling cyclones. Mean temperatures in summer only rarely exceed 16° C, except in interior regions where values near 25° C are possible. As a result of these temperature extremes, annual temperature ranges are larger in continental subarctic climates than in any other climate type on Earth, up to 30° C through much of the area and more than 60° C in central Siberia, although coastal areas are more moderate. Annual precipitation totals are mostly less than 50 centimetres, with a concentration in the summer. Higher totals, however, occur in marine areas near warm ocean currents. Such areas also are generally somewhat more equable and may be designated marine subarctic climates (see Figure 57). Areas with a distinct dry season in winter, which results in the Köppen climate types Dwc and Dwd, occur in eastern Siberia, both in the region where the wintertime anticyclone is established and in the peripheral areas subject to dry, divergent airflow from it.

Dominance of the Siberian anticyclone

### TYPE E CLIMATES

Köppen's type E climates are controlled by the polar and arctic air masses of high latitudes (60° N and S and higher). These climates are characterized by low temperatures and precipitation and by a surprisingly great diversity of subtypes.

Low precipitation and low temperatures

**Tundra climate (ET).** Tundra climates occur between 60° and 75° of latitude, mostly along the Arctic coast of North America and Eurasia and on the coastal margins of Greenland. Mean annual temperatures are below freezing and annual ranges are large (but not as large as in the adjacent continental subarctic climate). Summers are generally mild, with daily maxima from 15° to 18° C, although the mean temperature of the warmest month must be less than 10° C. Days are long (a result of the high latitude), but they are often cloudy. The snow cover of winter melts in the warmer season (though in places with mean annual temperatures of −9° C or less the ground at depth remains permanently frozen as permafrost); however, frosts and snow are possible in any month. Winters are long and cold (temperatures may be below 0° C for six to 10 months), especially in the region north of the Arctic Circle where, for at least one day in the year, the Sun does not rise. Winter precipitation generally consists of dry snow, with seasonal totals less than in the summer when cyclonic storms that develop along the boundary between the open sea and sea ice yield more rainfall. Typical annual totals are less than 35 centimetres, but a range from 25 to 100 centimetres is possible, with higher totals in upland areas and downwind of coasts. The latter regions also exhibit warmer winters, even in areas where the sea freezes; some climatologists regard them as a more equable polar marine subtype.

**Snow and ice climate (EF).** This climate (designated as Arctic in Figure 57) occurs poleward of 65° N and S latitude over the ice caps of Greenland and Antarctica and over the permanently frozen portion of the Arctic Ocean, the source regions for arctic air masses. Temperatures are below freezing throughout the year, and annual temperature ranges are large, but again not as large as in the continental subarctic climates. Winters are frigid, with mean monthly temperatures from −20° C to −65° C; the lowest temperatures occur at the end of the long polar night. The EF climate holds the distinction for the lowest recorded temperatures at the Earth's surface: the Vostok II research station in Antarctica holds the record for the lowest extreme temperature (−89° C), while the Plateau Station has the lowest annual mean (−56° C). Daily temperature variations are very small, because at such high latitudes the Sun's elevation varies little over the daylight period. Precipitation is meager in the cold, stable air (in most cases, five to 50 centimetres), with the largest amounts occurring on the coastal margins. Most of this precipitation results from the periodic penetration of a cyclone into the region, which brings snow and ice pellets and, with strong winds, blizzards. High winds also occur in the outer portions of the Greenland and Antarctic EF climates, where cold, dense air drains off the higher, central sections of the ice caps as katabatic winds (see above *Local wind systems*).

### HIGHLAND CLIMATES

The major highland regions of the world (the Cascades, Sierra Nevada, and Rockies of North America, the Andes

of South America, the Himalayas and adjacent ranges and the Tibetan Highlands [or Plateau] of Asia, the eastern highlands of Africa, and the central portions of Borneo and New Guinea) cannot be classified realistically at this scale of consideration, since the effects of altitude and relief give rise to myriad mesoclimates and microclimates. This diversity over short horizontal distances is unmappable at the continental scale. Very little of a universal nature can be written about such mountain areas except to note that, as a rough approximation, they tend to resemble cooler, wetter versions of the climates of nearby lowlands in terms of their annual temperature ranges and seasonality of precipitation. Otherwise, only the most general characteristics may be noted.

Multitude of mesoclimates and microclimates

With increasing height, temperature, pressure, atmospheric humidity, and dust content decrease. The latter effect results in high atmospheric transparency and enhanced receipt of solar radiation (especially of ultraviolet wavelength) at elevation. Altitude also tends to increase precipitation, at least for the first 4,000 metres. The orientation of mountain slopes has a major impact on solar radiation receipt and temperature and also governs exposure to wind. Mountains can have other effects on the wind climate; valleys can increase wind speeds by "funneling" regional flows and may generate mesoscale mountain- and valley-wind circulations as well. Cold air also may drain from higher elevations to create "frost pockets" in low-lying valleys. Furthermore, mountains can act as barriers to the movement of air masses, can cause differences in precipitation amounts between windward and leeward slopes, and, if high enough, can collect permanent snow and ice on their peaks and ridges; the snow line varies in elevation from sea level in the subarctic to about 5,500 metres at 15°–25° latitude. (A.J.A.)

# CLIMATE AND LIFE

## Coevolution of climate and life

It has been well established that life and climate interact and that they have mutually altered each other over geologic history. In a sense, they have undergone coevolution, a term coined by the American biologists Paul R. Ehrlich and Peter H. Raven to describe the process whereby two or more species depend on the interactions between them. It might be said that the coevolution of life and climate during the past 4,000,000,000 years of Earth history is an expression for the incredibly complex mixture of forces causing climatic change.

### THE DEVELOPING ATMOSPHERE

According to a widely held theory, the release of gases from the interior of the Earth during volcanic eruptions, along with other primitive processes, produced the planet's early atmosphere. Such processes determined the Earth's albedo (reflectivity to incoming solar radiation) and its heat-trapping "greenhouse" properties (see below *Causes of climatic change: The greenhouse effect*). Eventually carbon, hydrogen, oxygen, and nitrogen—the basic chemical constituents of life—were synthesized into primitive organic molecules from which early life forms subsequently evolved. This probably occurred during the first 500,000,000 years of Earth history. Anaerobic bacteria (those capable of living in the absence of oxygen) thrived in this early atmosphere, which consisted primarily of water vapour and carbon dioxide, with trace amounts of ammonia, methane, and various other gases. These organisms survived during this period because they found their ecological niche—*i.e.*, an environment having suitable chemical and physical conditions.

**The evolution of atmospheric oxygen.** Over an extended period of time, life contributes to the evolution of the natural environment, just as changing environmental conditions influence the development of organisms and their ability to survive. This interaction is exemplified by the evolution of atmospheric oxygen. The large volume of oxygen found today in the atmosphere accumulated over billions of years largely as a result of two processes. One of these consists of mechanisms for removing oxygen from the air and ocean and converting it to an oxide sediment, such as iron oxide. The other process is photosynthesis, a primary producer of oxygen. It seems likely that, once the continental crust was well oxidized (a comparatively efficient process), the absorption and conversion of free oxygen was controlled by oceanic mechanisms, which became less efficient. Thus, the production of oxygen by photosynthesizing organisms on an ever-expanding scale eventually outstripped the oxygen-removing capabilities of the environment. Ironically, the increasing supply of atmospheric oxygen was a disaster for many of the primitive anaerobic bacteria that produced it. Yet, the oxygen-rich atmosphere opened the door to the development of higher life forms.

Importance of photosynthesis

**The Gaia hypothesis.** The close interrelation between life and its environment, and its philosophical significance, was noted by the British chemist James E. Lovelock and the American biologist Lynn Margulis. They called this idea of complementary evolution of life and environment the Gaia hypothesis after Gaia, the ancient Greek goddess of the Earth. As Lovelock put it, this is "a new insight into the interactions between the living and the inorganic parts of the planet. From this has arisen the hypothesis, the model, in which the Earth's living matter, air, oceans, and land surface form a complex system which can be seen as a single organism and which has the capacity to keep our planet a fit place for life."

The Gaia hypothesis is highly controversial because it intimates that individual species (*e.g.*, ancient anaerobic bacteria) might sacrifice themselves for the benefit of all living things. Furthermore, the hypothesis has yet to be formulated quantitatively and in a scientifically testable manner. However, regardless of the eventual validity of the idea that life controls its environment for its own benefit, the recognition that the Earth's physical, chemical, and biological components interact and mutually alter their collective destiny, by accident or design, is a profound insight.

### CAUSES OF CLIMATIC CHANGE

The fact that climate influences terrestrial life is readily apparent. At the same time, it is all too clear that humans and other constituents of the biosphere affect climate significantly. Although many details about the interdependence of climate and the biosphere remain unknown, it has been firmly determined that there are relatively fixed supplies of certain elements essential to life (*i.e.*, nutrients) that circulate in the environment. These materials are transported (by wind or water, for example) and transformed chemically into forms that can be used by living things.

Interdependence of climate and the biosphere

Nutrients move in biogeochemical cycles. Climate, through atmospheric circulation, influences life by its effect on the flow of nutrients through these cycles. Many of these materials help to determine the composition of the atmosphere, which in turn affects climate. Water vapour is one such material. When it forms clouds, the Earth's albedo changes and more of the Sun's rays are reflected back into space. There is less heating, and so the climate is altered. (Water vapour and clouds also are important elements in the greenhouse warming effect; see below.)

Water is one of the most important nutrients for sustaining life, and its movement is termed the hydrologic cycle. What kinds of vegetation will grow in a given location is largely determined by the availability of water. Water is transferred to the air by means of evaporation and transpiration. Evaporation is the process whereby liquid water both from bodies of water (*e.g.*, rivers and lakes) and from the soil is transformed to water vapour. Transpiration involves the discharge of water vapour to the atmosphere from the surfaces of plants, particularly from their leaves. The total transfer of moisture to the

air via both these processes is termed evapotranspiration. On a global average, evapotranspiration on land is about six times less than evaporation of water from the ocean. Evapotranspiration, however, is often the principal local source of water vapour at the centres of continents.

Precipitation in the hydrologic cycle interacts with the so-called sedimentary cycle. Water helps to transport materials from the land to the sea, where they are incorporated into sediments. In the short term, this cycle consists of the processes of erosion, nutrient transport, and sediment formation. In the geologic, long-term sense, the sedimentary cycle utilizes mechanisms other than water flow. The processes involved are sedimentation, uplift, seafloor spreading, and continental drift.

The hydrologic and sedimentary cycles work together in the distribution of hydrogen, carbon, oxygen, nitrogen, phosphorus, and sulfur. These elements are the macronutrients, elements that account for 95 percent of the composition of all living organisms; they are necessary for sustaining life. The natural supply of these nutrients is fairly constant, with ample amounts existing within the Earth's crust. They are not always accessible, however, and must be recycled for life to continue.

**The carbon cycle.** The carbon cycle is the biogeochemical cycle that is probably of greatest interest to climatologists. Carbon exists in trace amounts in the atmosphere as carbon dioxide $CO_2$. It also exists in this and other forms in greater amounts in soil and bodies of water. Plants build their tissues with carbohydrates that are synthesized from $CO_2$ and water through photosynthesis, which utilizes solar energy. In spring and summer this process becomes more active because increasing sunlight and warmer temperatures help plants take $CO_2$ out of the air at a more rapid rate. In the Northern Hemisphere the concentration of $CO_2$ in the air decreases by a few percent from spring to fall (tens of billions of tons of $CO_2$ are involved). In the Southern Hemisphere the uptake of $CO_2$ is only about a third as much because there is less landmass and so fewer plants. In fall and winter the other part of the carbon cycle predominates as photosynthesis decreases and plants die and decay, introducing more $CO_2$ into the atmosphere than is removed by photosynthesis. Furthermore, the carbon cycle is influenced by exchanges of $CO_2$ between the atmosphere and the ocean. These exchanges are controlled by complex chemical processes in the ocean. In addition, other nutrients (*e.g.*, water and nitrogen) interact with carbon and life forms in an interlocking set of biogeochemical cycles. The very large carbon reservoirs in sedimentary rocks also are part of this biogeochemical cycling on a geologic time scale.

Absorption of infrared radiation by carbon dioxide

Although $CO_2$ constitutes only about 0.03 percent of the air, it affects the atmospheric heat balance because it absorbs infrared radiation. As a result, some of Earth's heat that ordinarily escapes as infrared rays into space is trapped by the atmosphere.

**The greenhouse effect.** This trapping of heat in the lower levels of the atmosphere is the so-called greenhouse effect. Next to water vapour, carbon dioxide is the most important gas in this process. There are other trace gases that, in concentration in the atmosphere, can create a strong greenhouse effect. Notable among these is methane ($CH_4$), which is produced by insects, humans, and other animals, and nitrous oxide ($N_2O$), which may be increasing due to the rapid growth in the use of nitrogen fertilizers. In satisfying its energy and agricultural needs, humankind has probably increased the amount of $CO_2$ in the atmosphere by 20 to 30 percent since the Industrial Revolution. Most projections point to roughly another 10 percent increase by the year 2000 and possibly as much as a 100 percent increase over preindustrial levels by the mid-21st century.

Using these estimates, global mean surface temperature could rise by approximately 1° C by the turn of the century and up to 5° C by the end of the 21st century. The latter figure approaches the magnitude of the average global temperature changes from glacial to interglacial periods. The implications of this warming are discussed below.

At this juncture, it is important to recall that climate largely determines what can grow where and the rate at which nutrients cycle through life and the environment. Completing the coevolutionary link between climate and life is the important fact that life affects the Earth's albedo and the composition of its atmosphere. Both of these regulate the energy balance of the planet, which in turn determines climate.

**Reduction in rain-forest cover.** Tropical rain forests play an important role in the exchange of gases between the biosphere and atmosphere. Significant amounts of nitrous oxide, carbon monoxide, and methane, for example, are released into the atmosphere from these forests. This metabolism is being changed by human activity. More than half of the carbon monoxide derived from tropical forests comes from their clearing and burning, which is reducing the size of such forests around the world.

Effects of deforestation

Another consequence of this deforestation must be examined. In the upper Amazon basin of South America, the rain forest recycles rains brought primarily by easterly trade winds. Surface transpiration and evaporation supply about half the rainfall for the entire region. In basins of dense forest cover far from the ocean, these local processes can account for most of the local rainfall. Should the Amazon Rain Forest, which accounts for 30 percent of the land area in the equatorial belt, disappear, drought would likely follow, and it might well affect the global energy balance. All around this equatorial belt, including Africa and the South Pacific islands, "forest farmers" are clearing portions of the rain forests each year. In some areas, it is being done to provide grazing land for more cattle with which to feed, at least in part, the nations of the Northern Hemisphere. In others, the deforestation is a survival response; that is to say, increasing population and uneven distribution of food supplies force the populace to expand their fields. Unfortunately, the soils in many parts of these cleared lands are not suitable for sustained agricultural use.

## Impact of climate on human life

### GENERAL OBSERVATIONS

Significance of short-term climatic variability

The interest in long-term atmospheric changes induced by human activities (anthropogenic changes) sometimes overshadows the fact that the natural short-term variability of climate also must be considered in assessing the influence of climate on society. Average, or mean, conditions are not a sufficient representation of the role climate plays in environmental and human affairs. Biologic systems are strongly affected by extremes in weather and climate. Accordingly, the character of climatic variability has an important effect on both natural and human communities and on the health of those communities.

A few obvious examples of the negative impact that natural climatic variability has had in the United States include the following. During the 1930s, crop and soil damage was the worst ever sustained in the Dust Bowl region of the central United States, and a forced emigration from the Great Plains reinforced the economic problems brought on by the Great Depression. In the winter of 1977, drought in the west and extreme cold in the east created economic disruptions estimated at billions of dollars. In the hot summer of 1980, the elderly, poor, and infirm in the south central states suffered abnormally high rates of mortality and morbidity. In addition, yields of such summer crops as corn and soybeans were greatly reduced. No thorough analysis of the distribution of losses and benefits resulting from climatic fluctuations of this sort has been made, but estimates put the net loss in the range of billions of dollars.

Experience with natural climatic fluctuations should be drawn upon both to develop policy measures with which to minimize society's vulnerability to anthropogenic climatic changes and to take advantage of new climatic resources. Atmospheric impact assessments for dealing with long-term problems usually involve an extrapolation of experience with the shorter-term effects of natural variability. Such experience, however, suggests that population size and the distribution and availability of resources mediate how natural climatic variability already affects society regardless of the existence of long-term climatic trends.

Thus, it is necessary to project how both climatic and societal factors will evolve and interact.

Day-to-day variability is an important atmospheric factor because society and the living environment respond nonlinearly to such variations. For example, the freezing point is a threshold below which small changes may have major effects on vegetation. Similarly, above certain temperatures, plant, animal, and human vitality can be seriously threatened. Since present-day climatic variability could well be superimposed on long-term climatic trends, one of the most significant effects of a seemingly small shift in the "mean" climate could be a large change in the frequency of harmful extreme situations.

## WEATHER AND TECHNOLOGY

Climate-induced problems

In the latter part of the 1800s during the homesteading of the Great Plains, farmers were for the most part successful because the weather was generally favourable. Severe drought conditions in the 1890s, however, shattered the illusion that life on the plains was necessarily good and drove out many settlers. This pattern of boom-and-bust farming recurred several times: good weather and high-production years were followed by periods of drought, economic ruin, and serious soil erosion. The worst drought and resultant soil degradation occurred during the 1930s in the area of the Dust Bowl. Average wheat and corn yields fell by as much as 75 percent. Worse, millions of tons of valuable topsoil were lost.

As a result of this climate-induced disaster, the federal government established the U.S. Soil Conservation Service to help farmers protect the soil. Decades later new crop strains better adapted to regional climates were developed, and irrigation and chemical fertilizers were made available to take advantage of the new genetic strains and to promote production. These advances, together with the development of such technological aids as tractors, mechanical harvesters, irrigation pumps, and agrochemicals (*e.g.*, pesticides and herbicides), have increased the productivity (average yield of grain per harvested area) of the Great Plains by 200 to 300 percent since the 1930s. Total production (productivity multiplied by total harvested area) also has risen. In addition, the amount of year-to-year variability in yield, relative to the average yield, for most grains has decreased over time. Yet, while the relative variability (year-to-year variability as a percentage of long-term average yield) has generally declined for crops in recent years, the absolute variability (year-to-year variability in yield by itself) has increased on the whole in spite of all the technological advances.

The impact of agrotechnology

This has led to an ongoing debate about the relative role of climate versus the role of technological advances in influencing both average-yield and yield-variability trends during the 1960s and '70s. While agrotechnology appears to have been the prime factor behind the general increase in annual grain yield, there is some doubt as to its actual contribution to the favourable decrease of variability in relative yield. Some investigators, most notably the American geographer Richard Warrick, have argued that this decrease cannot be entirely attributed to modern farming practices. Only if the weather anomalies of the 1960s and '70s had been as bad as those from roughly 1930 to the late 1950s could agrotechnology be credited with the reduced impact of climatic stress on crop yields. Warrick and his associates compiled an index of the severity of summer droughts in the Great Plains in the period 1931–77 (Figure 58), which indicates quite clearly that since the late 1950s weather conditions have not been bad enough to test the hypothesis that technological advances have truly reduced the annual variations of grain yield. Given this situation, it is risky to assume that modern technology in the "breadbasket" of North America can stably maintain yearly productivity.

## WORLD FOOD PRODUCTION AND CLIMATE

World food production outpaced population growth through the 1960s. Many began to believe that the technologies responsible for the so-called green revolution had ended the threat of Malthusian disaster. The worldwide events of 1972 and those of 1974 in the United States and Canada badly shook the conviction that technology had finally solved the problem of world hunger.

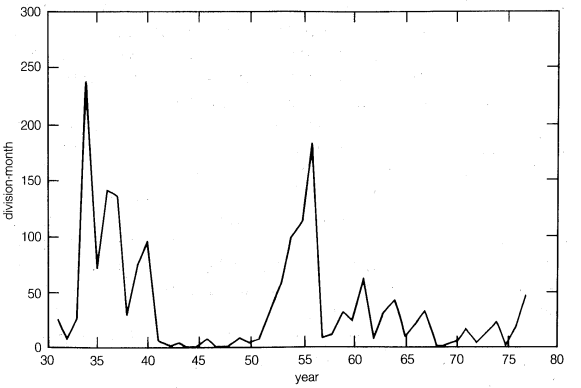



Figure 58: U.S. Great Plains drought area, 1931–77. Shown here is the number of divisions exhibiting severe or extreme drought (≦ −3.00) conditions, based on Palmer index values summed over five months, April through August. The more intense the drought, the higher the division-month number.

From R.A. Warrick, "Drought in the Great Plains: A Case Study of Research on Climate and Society in the U.S.A.," in J. Ausubel and A.K. Biswas (eds.), *Climatic Constraints and Human Activities* (1980), Pergamon Press Ltd; reprinted with permission

**Adverse effects of global weather conditions.** The year 1972 was not a good year for much of the world. There were serious climatic, economic, and human setbacks: severe droughts occurred in the Soviet Union, India, Southeast Asia, Australia, Central and South America, and the Sahel region of Africa; Peru's protein-rich anchovy fishery was devastated as a result of an El Niño event (see below); and grain supplies in many major food-producing areas were depleted. The resulting famines eventually killed or debilitated tens of millions of people. The total number of deaths in India and Bangladesh attributed to this bad-weather year was a million or more.

Shortfalls in Soviet, Indian, African, and Peruvian food production led to a 3 percent drop in global grain production in 1972. Such a seemingly small loss, when combined with a growing need for food and a 2 percent annual population growth rate, proved to be a significant problem. Climatologists publicly debated the role of climate in these events and the likelihood that climate-induced troubles would increase.

**Grain crises in the mid-1970s.** Except for 1974, world grain consumption rose steadily after 1966. Total grain yields began dropping in 1973; they were down by roughly 4.5 percent in 1974 from the 1973 yields. Harvested area was up only about 0.4 percent in 1974 over 1973. Consequently, worldwide grain production in 1974 was set back by about 3 percent—tens of millions of metric tons. In addition, the upward trend in world grain trade broke in 1974. All of this occurred at a time when grain stocks were about the lowest in decades.

**Grain security in the 1980s.** A number of years of good weather returned to most of the world near the end of the 1970s. However, 1980 was hot and dry in the central United States, which resulted in low yields in summer crops. There were poor harvests in the Soviet Union as well. Consequently, the world had about the same level of grain security—approximately 14 percent—at the beginning of the 1980s as it did entering the 1970s. This level dropped another percentage point the next year, but it recovered somewhat by 1982.

A weather event occurred in 1983 in the United States that illustrates some of the consequences of governmental crop planning and controls. There was a large surplus of grains in 1982 that had farmers complaining about depressed prices for their crops. The U.S. Department of Agriculture introduced a new program that paid grain farmers to withhold acreage from production. It estimated that harvested area in the United States for coarse grains would drop from more than 43,000,000 hectares (106,000,000 acres) in 1982 to an estimated 33,000,000 hectares in 1983. This decrease of 23 percent in crop area by the summer of 1983 occurred at the same time that a heat wave struck the corn belt. Yields of coarse grains

dropped, and production fell to a little under 140,000,000 metric tons, a reduction of nearly half compared to that of 1982. This depleted the existing stocks of coarse grains to about two-thirds by 1984. Fortunately, poor yields did not continue, and there was some recovery the following year. This situation, however, initially had some uncomfortable parallels with that which prevailed just before the food crisis of the early to mid-1970s.

The climatic events of the 1970s and '80s show all too clearly that it is becoming more and more risky to ignore the impact of abnormal weather when making food policy. Climate is not the constant it was assumed to be in past food-policy analyses.

**Sahelian drought.** The Sahel, which borders the southern fringe of the Sahara in Africa, is extremely prone to drought. Persistent drought conditions, coupled with substantial population growth in the region (an increase of more than 30 percent since the early 1950s) and a doubling of the livestock herd, have resulted in a gradual desertification of the Sahel. Desertification is extremely difficult to reverse. At first, there is a slow loss of soil fertility; this is followed by a process of rapid degradation, with the soil eventually losing its structure and ability to retain enough moisture to sustain vegetation.

Desertification

While climatic variations have played a major role in this disaster, economic and social priorities have been involved as well. The introduction of Western technology made it possible for the inhabitants of the Sahel to drill deep water wells. This situation encouraged the herdsmen not only to give up their nomadic way of life and remain near the wells but also to raise more livestock. Overgrazing resulted and, as drought conditions persisted, competition for forage became more and more intense. Herds of goats tore up the remaining indigenous plants by their roots, thereby destroying the ability of the plants to reproduce by themselves. At the same time, the increasing human population led to the cultivating of more and more marginal, ecologically fragile lands for subsistence farming. (The most fertile lands were frequently used to grow cash crops—namely, cotton and peanuts for foreign markets.) Furthermore, Sahelian farmers began reworking the marginal land within one to five years, whereas they had traditionally allowed these lands to remain fallow for 15 to 20 years, giving them ample time to recover. In changing their farming practices, the Sahelians contributed to the destruction of their lands. As the desert moved southward, so did the people and their livestock. This resulted in further denudation and deforestation. Desertification had set in by the late 1960s, followed by widespread famine. A major international project was undertaken to keep millions of Sahelians alive. Normal rains returned briefly to the Sahel during the mid-1970s; however, since no effective, long-term remedies were applied in animal, agricultural, and human control, outside aid again became necessary in the mid-1980s as drought conditions prevailed once more for a prolonged period (especially in the eastern sections of the sub-Sahara in Ethiopia), causing widespread famine and death. Figure 59 provides an index of rainfall for the Sahel during this period. What caused these droughts is still debated, but one of the theories involves unusual or anomalous temperature patterns in certain parts of the ocean.

From P. Lamb, "Rainfall in Subsaharan West Africa During 1941-83," *Zeitschrift fur Gletscherkunde und Glazialgeologie*, No. 21 (1985)

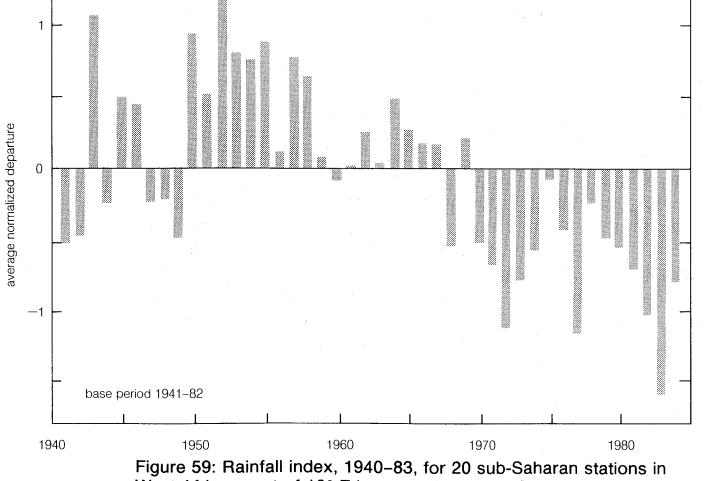



Figure 59: Rainfall index, 1940–83, for 20 sub-Saharan stations in West Africa, west of 10° E between 1° N and 19° N.

**El Niño/Southern Oscillation.** The interaction between the ocean and the atmosphere also can have a marked impact on life, health, and food. A manifestation of this is the El Niño phenomenon that occurs in the Pacific Ocean. Every few years the temperature of the normally cool surface waters of the eastern equatorial Pacific increases. In turn, the warmer waters affect the atmosphere, and rainfall and surface temperatures along western South America increase substantially. The reduction of cool upwelling water off the western coast disrupts commercial anchovy fishing in the region because the plankton on which the anchovies feed are nourished by nutrients brought up by the colder upwelling water. When the plankton decrease, so do the anchovies.

The El Niño phenomenon is not confined to the waters of South America. The normally warm Pacific water along Australia is replaced by an upwelling of cold water; precipitation in the western Pacific seems to decrease as a result. In addition, changes in atmospheric pressure occur off the shores of Australia, just as they do along the western coast of South America, and wind patterns deviate from their normal course in both cases. Extra-severe drought in Australia and flood-producing torrential rains and heat in South America occur concurrently with and are blamed on El Niño. This entire effect has been referred to as ENSO, for El Niño/Southern Oscillation.

Effect of the ENSO on mid-latitude climates

Studies suggest that the ENSO can affect mid-latitude climates, modulating the position and intensity of the polar-front jet stream (see above). An El Niño event that began in early 1982 lasted well into 1983. It was accompanied by unusual weather events outside of the equatorial Pacific region. Western Europe suffered from record summer heat. Late-fall temperatures in the United States were very cold; winter was mild; spring rains were far above normal and spring temperatures in the central part of the country were extremely cold; and summer conditions in the Midwest and Southeast were extremely hot and dry. These abnormal weather conditions were consistent with a shift in the jet-stream pattern away from its normal one and were with little doubt caused in part by the intense El Niño.

The ENSO appears to be a truly disruptive force that wreaks havoc on life. In 1982–83 it not only caused the drought in Australia and the flooding along the western coast of South America and the loss of the anchovy catch there, but it also damaged corn, soybean, and other summer crops in the United States, which resulted in losses amounting to billions of dollars. Clearly a better understanding of the El Niño and its associated atmospheric effects is needed, leading perhaps to predictive skill.

## Impact of human activities on climate

Three distinct atmospheric problems have been debated intensely since about the mid-1970s, though two of them are quite old issues: the possible reduction of stratospheric ozone from chemical emissions; the generation of acid rain; and climatic change stemming from the greenhouse effect. What these three problems have in common is quite simple: they all (1) are complex and punctuated by large uncertainties, (2) could be long-lasting, (3) transcend state and even national boundaries, (4) may be difficult to reverse, (5) are inadvertent by-products of widely supported economic activities, and (6) may require substantial investments of present resources to hedge against the prospect of large future environmental changes.

### OZONE DEPLETION

Of these problems, the only one to have received any substantial public policy action is that centring on the reduction of stratospheric ozone. Ironically, it is perhaps the easiest of the problems to reverse.

The importance of the stratospheric ozone layer in shielding the Earth's surface from the harmful effects of solar ultraviolet radiation has been recognized for several decades. It was not until the early 1970s, however, that scientists began actually to grapple with the fact that even relatively small decreases in the stratospheric ozone concentration can have a serious impact on human health—an increased incidence of skin cancer, particularly among fair-skinned peoples. Plans in the United States, Great Britain, and France to build a commercial fleet of supersonic aircraft triggered much heated discussion over the potential reduction of the ozone layer by the exhaust gases (*e.g.,* nitric oxide) emitted by such high-altitude planes. The debate in turn stimulated intensive scientific research on the stratosphere, which resulted in new findings and new concerns.

Efforts to reverse the problem

By the mid-1970s, various U.S. investigators had determined that chlorofluorocarbons (CFCs), widely employed as propellants in aerosol spray cans, could reduce the amount of stratospheric ozone significantly. A temporary ban was imposed on the use of certain CFCs in the United States, but only after much emotional debate among environmental and industrial scientists, reports by the National Academy of Sciences, and the development by industry of economically viable substitutes for spray-can propellants. (For a more detailed discussion of this issue, see ATMOSPHERE: *Depletion of stratospheric ozone.*)

## ACID RAIN

Acid precipitation has been known for centuries in locales such as London where sulfur discharged by the burning of coal produces toxic smogs; however, the problem did not assume scientific, economic, and political prominence until the early 1980s. As it transcends national boundaries, the acid rain problem has become a subject of heated controversy between otherwise friendly neighbours like the United States and Canada or West Germany and the Scandinavian countries.

Scientific studies have shown that the process that results in the formation of acid rain generally begins with the discharge of sulfur dioxide and nitrogen oxides into the atmosphere. These waste gases are released by the combustion of fossil fuels by automobiles, electric power plants, and smelting and refining facilities. They also are emitted by some biologic processes. The gases combine with atmospheric water vapour to form sulfuric and nitric acids. When rain or some other form of precipitation falls to the surface, it is highly acidic, frequently with a pH value of less than 4. (The term pH is defined as the negative logarithm of the hydrogen ion concentration in kilograms per cubic metre. The pH scale ranges from 0 to 14, with lower numbers representing greater acidity.) The consequent acidification of surface and subsurface waters is widely believed to have a detrimental effect on the ecology of the affected areas. Such regions as the Canadian Shield in Quebec and the Adirondack Mountains in New York are especially susceptible to contamination, because the snowpack buildup in winter allows a deadly pulse of acidic meltwater to occur during spring. As highly acidic water is toxic to many aquatic organisms, many lakes in these regions are biologically damaged. It also has been found that acid precipitation is harmful to trees and other forms of vegetation, causing foliar injury and reduction in growth (see also ATMOSPHERE: *Acid rain and allied problems*).

Detrimental effects of acid rain

The "cause-and-effect linkages" of the acid rain problem have been more clearly demonstrated in scientific terms than those related to ozone depletion; yet, the former has received much less direct policy action. The primary reason is the potential economic impact of efforts to remedy the problem—*i.e.,* the enormous expenditure that would be required to control the emission of sulfur compounds from power plants, refineries, and facilities of other smokestack industries. There continues to be loud and angry debate as to the environmental and economic benefits of corrective action, particularly since the relationship between the discharge of potentially acidic compounds and the ultimate delivery of acid precipitation to specific geographic areas is not straightforward.

## GREENHOUSE EFFECT INDUCED BY CARBON DIOXIDE AND OTHER TRACE GASES

Finally, the most long-lasting and potentially least reversible global problem is the greenhouse effect. As noted above, this effect is induced by carbon dioxide, chlorofluorocarbons, methane, and more than a dozen other gases in concentration in the atmosphere. The role played by carbon dioxide is the most significant. The amount of $CO_2$ in the atmosphere has risen steadily since the mid-1800s largely as a result of the combustion of coal, oil, and natural gas on an ever-widening scale. In 1850 the global $CO_2$ level of the atmosphere was roughly 280 parts per million, whereas by the late 1980s it had increased to approximately 350 parts per million.

The possibility of major climatic changes and their consequences

Should present trends in the emission of greenhouse gases, particularly of $CO_2$, continue beyond another 100 years, climatic changes larger than any ever experienced during recent geologic periods can be expected. This could substantially alter natural and agricultural ecosystems, human and animal health, and the distribution of climatic resources. In addition, any significant greenhouse warming could cause a rapid melting of some polar ice, resulting in a rise in sea level and the consequent flooding of coastal areas.

In spite of these long-term possibilities, the greenhouse problem has received the least policy-oriented attention of any of the three major issues at hand. There are various reasons for this: (1) The problem is fraught with technical uncertainties. (2) It has perceived "winners" and "losers"—economic and otherwise. (3) No one nation acting alone can do much to counteract the $CO_2$ buildup in the atmosphere. (4) Dealing with the problem substantively could be expensive and even alter life-styles. (5) There is no way of proving the validity of the greenhouse theory to everyone's satisfaction except by "performing the experiment" on the real climatic system, which would necessarily involve all living things on Earth. (6) The principal greenhouse gas, $CO_2$, is an inherent by-product of the utilization of a commodity that is most fundamental to the economic viability of the world—fossil-fuel energy. (This fact more than any other explains why the greenhouse problem is so difficult to solve.)

**Detailed analysis of the problem.** It seems appropriate to break down the issue of greenhouse warming into a series of stages and then consider how policy questions might be addressed against the background of these more technical stages. The present discussion will deal with the problem specifically as it relates to increasing atmospheric $CO_2$ for the sake of simplicity, though other related questions certainly can be dealt with in the same manner.

*Behavioral assumptions.* At the very basis of the problem is the need to make behavioral assumptions about the future use of fossil fuels (or alternatively of the projected extent of deforestation, because this, too, can affect the amount of carbon dioxide in the atmosphere; see above). In essence, this issue has to do with social science rather than with chemistry, physics, or biology. It depends on projections of human population, the per-capita consumption of fossil fuel, deforestation rates, reforestation activities, and perhaps even countermeasures for dealing with the additional $CO_2$. These projections, of course, are contingent on such questions as the likelihood of alternative energy systems or conservation measures becoming available, their costs, and their acceptability to society. Furthermore, trade in carbon-based fuels will be determined not only by energy requirements and available alternatives but also by the economic health of potential importing nations. This, in turn, will depend on whether those nations have adequate capital resources to spend on energy, rather than on other precious strategic commodities such as food or fertilizer or even weaponry. The future can be projected by drawing up scenarios such as that in Figure 60, which shows different projected $CO_2$ concentrations based on assumed rates of growth in the use of fossil fuels. Most typical projections are in the 1–2 percent annual growth range, implying doubling of atmospheric $CO_2$ by the middle of the 21st century. It has already increased by some 25 percent in the 20th century.

*Carbon cycle.* Once the plausible scenarios for $CO_2$

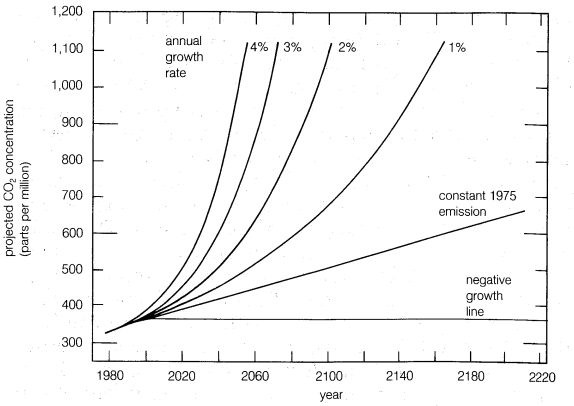



Figure 60: *Projected carbon dioxide concentrations from 1980 for different annual growth rates in fossil-fuel energy use.* The projections reflect the assumption that no increase in fossil-fuel energy use will occur (constant 1975 emission) and the possibility of negative growth, with annual energy growth after 1985 assumed to be reduced by a fixed amount (about 2 percent of present demand) each year.

From (data for negative growth line) A.B. Lovins *et al., Least-Cost Energy: Solving the $CO_2$ Problem* (1982), Brick House; (remaining data) S.H. Schneider and R. Londer, *The Coevolution of Climate and Life*, copyright ©1984 by S.H. Schneider and R. Londer, reprinted with permission of Sierra Club Books

buildup have been devised, it is necessary to determine exactly which interacting biogeochemical processes control the global distribution of carbon and its stocks. This involves uptake by green plants (since $CO_2$ is the basis of photosynthesis, more $CO_2$ in the air means faster rates of photosynthesis), changes in the amount of forested area, the types of vegetation planted, and the way in which climatic change affects natural ecosystems. The growth rate of photosynthesizers, such as grain plants and trees, may well increase. On the other hand, weeds and vegetation that harbour disease-bearing insects would also become more vigorous. Moreover, since there is a slow removal of $CO_2$ from the atmosphere, largely accomplished through chemical processes in the ocean that take from decades to centuries, the rates at which climatic change modifies mixing processes in the ocean also need to be taken into account. There is considerable uncertainty over just how much $CO_2$ will remain in the air, but most present estimates put the so-called airborne fraction at about 50 percent, which suggests that, over the time frame of a century or two at least, something like half the $CO_2$ injected into the atmosphere will remain and exacerbate the greenhouse effect.

*Global climatic response.* Once the amount of carbon dioxide that may exist in the atmosphere over the next century or so has been projected, its significance in terms of climate has to be estimated. The greenhouse effect, notwithstanding all of the controversy that surrounds the term, is not a scientifically controversial subject. In fact, it is one of the best, most well-established theories in the atmospheric sciences. For example, with its extremely dense atmosphere composed largely of $CO_2$, Venus has very high surface temperatures (up to about 500° C). By contrast, Mars, with its very thin $CO_2$ atmosphere, has temperatures comparable to those that prevail at the Earth's poles in winter. The explanation for the Venus hothouse and the Martian deep freeze is really quite clear—the greenhouse effect. This mechanism works because some gases and particles in a planet's atmosphere preferentially allow sunlight to filter through to the surface of the planet relative to the amount of radiant energy that the atmosphere allows to escape back to space. This latter kind of energy (infrared energy) is affected by the amount of greenhouse material in the atmosphere. Therefore, increasing the amount of greenhouse gases raises the surface temperature of the planet by increasing the amount of heat that is trapped in the lowest part of its atmosphere. While that part of the subject is not controversial, what is open to debate is exactly how much the Earth's surface temperature will rise given a certain increase in a trace greenhouse gas such as $CO_2$. Complications arise due to processes known as feedback mechanisms. For example, if the $CO_2$ added to the atmosphere were to cause a given temperature increase on Earth, warming would melt some of the snow and ice that now exist. Thus, the white surface, originally covered by the melted snow and ice, would be replaced with darker blue ocean or brown soil, surface conditions that would absorb more sunlight than the snow and ice. Consequently, the initial warming would create a darker planet that absorbs more solar energy and thereby produces greater warming in the end. This is only one of a number of possible feedback mechanisms, however. Because many of them are interacting simultaneously in the climatic system, it is extremely difficult to estimate quantitatively how many degrees of warming the climate will undergo for any given increase in greenhouse trace gases.

The effect of feedback mechanisms

Unfortunately, there is no period in Earth history that investigators can examine when carbon dioxide concentrations in the atmosphere were, say, twice what they are today and whose climatic conditions are known with a high degree of certainty. For this reason, investigators cannot directly verify their quantitative predictions of greenhouse warming on the basis of historical analogs. Instead, they must base their estimates on climatic models. These are not laboratory models, since no laboratory could approach the complexity of the real world. Rather, they are mathematical models in which basic physical laws are applied to the atmosphere, ocean, and glaciers; the equations representing these laws are solved with computers with the aim of simulating the present terrestrial climate.

Estimates of greenhouse warming based on climatic models

Many such models have been built during the past few decades. The calculations roughly agree that, if the atmospheric $CO_2$ concentrations were to double, the Earth's surface temperature would warm up somewhere between 1° and 5° C. As a point of comparison, the global surface temperature of the Earth during the Ice Age 18,000 years ago was on average about 5° C lower than it is today. Thus, a temperature change of more than one or two degrees worldwide represents a very substantial alteration.

*Regional climatic response.* To estimate the importance of climatic changes to society, researchers need not, however, study global average temperature so much as the possible regional distribution of evolving patterns of climatic change in the future. Will it, in the year 2010, be drier in Iowa, wetter in Africa, more humid in New York, or too hot in India? Unfortunately, to predict reliably the fine-scale regional response of variables, such as temperature and rainfall, requires climatic models of greater complexity (and cost) than are currently available. At present, there is simply no consensus among knowledgeable atmospheric scientists that the regional predictions of state-of-the-art models are reliable. Nevertheless, most experts agree that the following coherent regional features might well occur by about the year 2035: wetter subtropical monsoonal rain belts; longer growing seasons in high latitudes; wetter springs in high and middle latitudes; drier midsummer conditions in some mid-latitude areas (a potentially serious agricultural and water supply problem in major grain-producing nations); increased probability of extreme heat waves (with possible health consequences for people and animals in already warm climates); and an increase in sea level by a few tens of centimetres. Considerable uncertainty remains in these regional estimates, even though many plausible scenarios have been investigated.

Projected climatic variations resulting from greenhouse warming

*Environmental impact.* The possible impact on environment and society needs to be determined from a given set of scenarios for regional climatic change. Most important are the effects on crop yields and water supply. Also of concern is the potential for altering the range or number of pests that affect plants and diseases that threaten the health of humans or lower animal forms. Another point of interest is the effect on unmanaged ecosystems. For example, ecologists are much concerned that the rapid rate at which tropical forests are being destroyed due to human expansion is eroding the genetic diversity of the Earth. Since the tropical forests are in a sense repositories for the bulk of living genetic materials on Earth, the world is losing some of its irreplaceable biologic resources. In addition, substantial future changes to tropical rainfall have been predicted by climatic models. This means that present-day reserves (or refugia) may be unable to sustain

those species that they are designed to protect if rapidly evolving climatic change produces conditions in the refugia sufficiently different from those of today.

*Economic, social, and political effects.* Estimating the distribution of economic winners and losers, given a scenario of climatic change, involves more than simply looking at the total dollars lost and gained, were it possible to make credible calculations. It also requires looking at such important equity questions as who wins and who loses and how the losers might be compensated and the winners taxed. If the corn belt in the United States, for example, were to "move" north by several hundred kilometres as a result of greenhouse warming, then $1,000,000,000 a year lost on Iowa farms could well become Minnesota's $1,000,000,000 gain. Some macroeconomic views of this hypothetical problem, from the perspective of the United States as a whole, might see no net losses. Much social consternation, however, would be generated by such a shift in climate, particularly since the cause of the change would be economic, $CO_2$-producing activities. Even the perception that the economic activities of one nation could create climatic changes that would be detrimental to another has the potential for disrupting international relations, as is already occurring in the case of acid rain. While the details are still difficult to establish, there is considerable scientific consensus that regional climatic changes of environmental significance are likely to occur over the next few generations. In essence, the environmental changes induced by $CO_2$ and other greenhouse gases create what might be termed a problem of "redistributive justice."

*Policy responses.* The last stage in dealing with the greenhouse effect concerns the matter of appropriate policy responses. Three classes of action could be considered.

Possible ways of dealing with the greenhouse effect

The first is mitigation through purposeful intervention to minimize the potential effects on the environment. For example, dust might be deliberately spread in the stratosphere to reflect some sunlight and cool the climate as a countermeasure to inadvertent warming by a $CO_2$ buildup. This solution suffers from the obvious flaw that, since there is uncertainty associated with predicting the inadvertent consequences of human activities, substantial uncertainty must surround any deliberate attempts at climatic modification. It may be that existing computer models overestimate changes and underestimate proposed modification schemes. If so, human intervention would be the proverbial "cure worse than the disease." In that case, the prospect for international tensions is so staggering and society's legal instruments to deal with the problem so immature that it is hard to imagine any substantial mitigation strategies for the foreseeable future.

The second is simply adaptation. Adaptive strategies, favoured by many economists, would simply allow society to adjust to environmental changes without attempting to mitigate or prevent the changes in advance. It would be possible to accommodate climate change, for instance, by planting alternative crop strains that would be more widely adapted to a whole range of plausible climatic conditions anticipated for the future. Such adaptive strategies are often recommended because of the uncertain nature of the redistributive character of future climatic change.

The third type of policy response, namely prevention, is the most active. It might involve discontinuing the use of chlorofluorocarbons and other potential ozone-reducing gases, or a reduction in the amount of fossil fuel used around the world. These policies, often advocated by environmentalists, are controversial because in some cases they require substantial immediate investment as a hedge against large future environmental change, change that cannot be predicted precisely. What can be considered practical options are increasing the efficiency of energy end use (in a word, conservation), the development of alternative energy systems that are not based on fossil fuels, or, more radically, establishing a "law of the air." Such a measure was proposed in 1976 by two American scientists, the anthropologist Margaret Mead and the climatologist William W. Kellogg. To curtail $CO_2$ emissions, they suggested setting up a global standard and assigning various nations the right to generate certain levels of the gas.

The "appropriate" policy response will depend not only on scientific information about the probabilities and consequences of physical, biologic, and social impact scenarios but also on the value judgments of individuals, groups, corporations, and nations as to how to deal with the potential distribution of gains and losses implied by the buildup of carbon dioxide and other trace gases. There is no scientific answer as to how society should act and no scientific basis for any particular policy choice. All science can do is provide scenarios and assess the probabilities and consequences of various plausible alternatives. The public and government leaders need to understand that decisions have to be made in the face of scientific uncertainty by optimizing clearly stated sets of often conflicting values.

Atmospheric problems are fundamentally global in both cause and effect. Moreover, they are inextricably interwoven with the overall problem of global economic development and cannot be removed from the discussion of population, resources, environment, and economic justice. Rich nations cannot ask poor nations to abandon their development plans simply because of the potential $CO_2$ problem. Any global strategies for preventing $CO_2$ buildup will require cooperation between rich and poor nations on the transfer of knowledge, technology, and capital.

A further point of contention in the dialogue between developed and developing countries will be the question of population-growth rates. This is quite relevant to the $CO_2$ issue simply because total emission is the per-capita emission rate times the population size. If there is a movement in the future toward parity between rich and poor in per-capita use of fossil fuels, then population growth (which is occurring predominantly in Third World countries) will become as important a factor in the long-term $CO_2$/climate problem as is high per-capita use of fossil fuels today (largely a problem of developed nations).

Furthermore, there is an ethical question associated with atmospheric problems: Do we have the right to commit future generations to unprecedented atmospheric perturbations without actively attempting to prevent or at least anticipate them? To be sure, there is much uncertainty. It is safe to say, however, that humankind is abusing the atmospheric environment faster than it is understanding it. Clearly, some of the uncertain consequences could be serious and even irreversible.

From S. Manabe and R.J. Stouffer, *Journal of Geophysical Research*, 85:5529–54 (1980); published by the American Geophysical Union

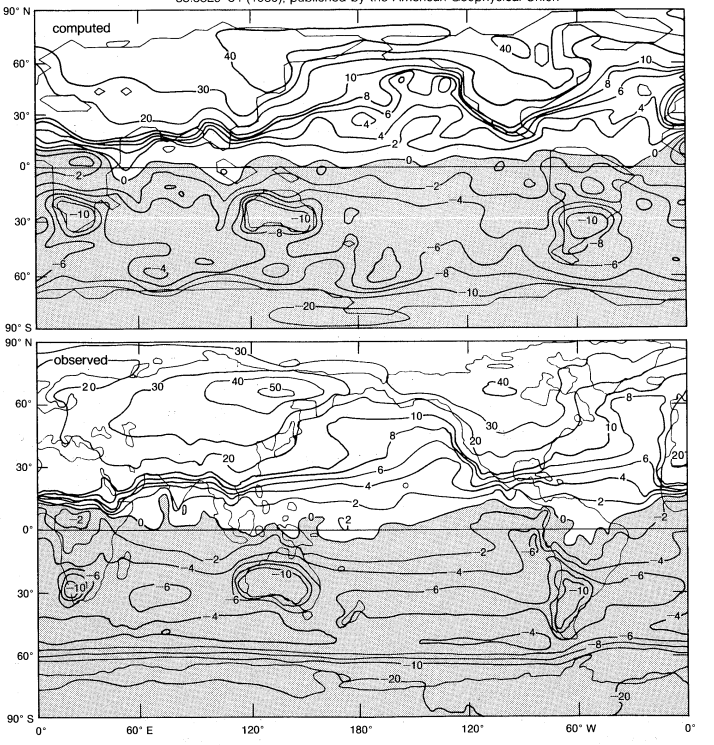



Figure 61: Global winter-to-summer temperature extremes, computed (top) and observed (bottom). A three-dimensional climate model was used for the top map.

**Verification of model predictions.** The largest climatic change to occur on Earth since the dinosaurs inhabited the planet some 100,000,000 years ago is the seasons. Temperature differences between winter and summer are many tens of degrees—in some cases 10 times—larger than those anticipated over the next century from an intensification of the greenhouse effect. How well, then, do existing climatic models fare in predicting this massively large seasonal change? They do quite well, in fact, as Figure 61 suggests. Thus, it is not likely that climatic models are making a fundamental error in predicting future warming and at the same time reproducing the very large seasonal differences that they do. On the other hand, seasonal variations occur within a year, whereas the rate of change for an increase of trace gases in the atmosphere is many decades. In effect, seasonal simulation skill is a piece of circumstantial evidence and not final proof of model validity. Nevertheless, it certainly lends a great deal of credence to the basic theoretical claims behind the greenhouse effect.

Comparisons with Venus and Mars

What other bits of supporting evidence can be brought to bear? The furnacelike conditions on Venus and the frigid air of Mars were noted above. When data on the atmospheric composition of these planets are fed to the computer models used to predict greenhouse warming on the Earth, corresponding changes are indeed produced—*i.e.,* temperatures on Venus are hundreds of degrees warmer and those of Mars are hundreds of degrees colder. Is this proof that the mathematical models are accurately predicting the terrestrial greenhouse effect? Again, the evidence is only circumstantial but strong. One other piece of circumstantial evidence can be cited—namely, the ability of present climatic models to reproduce the vastly different conditions on Earth during ancient times when ice stretched all the way from the Arctic to northern Europe and the mid-Atlantic region of the United States.

Finally, what has been happening to the Earth's climate during the past 100 years? If one takes all of the reliable records of temperature readings and averages them for the world, one finds that the Earth has warmed up about 0.5° C over the past century. At the same time, it has been determined that the $CO_2$ level is about 25 percent higher today than it was a century ago. Therefore, what has happened on Earth is broadly consistent with what the climatic models suggest should have been happening. The warming of the planet over the past century is yet another piece of circumstantial evidence, not conclusive proof in and of itself, but still important. (S.H.S./A.B.Ri.)

# METEOROLOGICAL MEASUREMENT AND WEATHER FORECASTING

## General considerations

### MEASUREMENTS AND IDEAS AS THE BASIS FOR WEATHER PREDICTION

The link between meteorological measurement and weather forecasting, the two topics of this section, is both necessary and convenient: necessary because all weather forecasts are grounded in data that have only very recently been measured, and convenient because attention is focused on the central importance of observations for the practice of science. With few other scientific enterprises are observations so vital as in the case of weather forecasting. From the days when early humans ventured from caves and other natural shelters, perceptive individuals in all likelihood became leaders by being able to detect nature's signs of impending snow, rain, or wind, indeed of any change in weather. With such information they must have enjoyed greater success in the search for food and safety, the major objectives of that time.

In a sense, weather forecasting is still carried out in basically the same way as it was by the earliest humans—namely, by making observations and predicting changes. The tools used to measure temperature, pressure, wind, and humidity in the late 20th century would certainly amaze them, and the results obviously are better. Yet, even the most sophisticated numerically calculated forecast made on a supercomputer requires a set of measurements of the condition of the atmosphere—an initial picture of temperature, wind, and other basic elements, somewhat comparable to that formed by our forebears when they looked out of their cave dwellings. The primeval approach entailed insights based on the accumulated experience of the perceptive observer, while the modern technique consists of solving equations. Although seemingly quite different, there are underlying similarities between both practices. In each case the forecaster asks "What is?" in the sense of "What kind of weather prevails today?" and then seeks to determine how it will change in order to extrapolate what it will be.

Because observations are so critical to weather prediction, an account of meteorological measurements and weather forecasting is a story in which ideas and technology are closely intertwined, with creative thinkers drawing new insights from available observations and pointing to the need for new or better measurements, and technology providing the means for making new observations and for processing the data derived from measurements. It tells of the theories of the ancient Greek philosophers, Renaissance scientists, and the scientific revolution of the 17th and 18th centuries and of the creativity of 20th-century atmospheric scientists and meteorologists. Likewise, it tells of the development of the "synoptic" idea—that of characterizing the weather over a large region at exactly the same time in order to organize information about prevailing conditions. In synoptic meteorology, simultaneous observations for a specific time are plotted on a map for a broad area whereby a general view of the weather in that region is gained. (The term synoptic is derived from the Greek word meaning "general or comprehensive view.") The so-called synoptic weather map came to be the principal tool of 19th-century meteorologists and continues to be used today in weather stations and on television weather reports around the world.

The synoptic approach

Figure 62 provides an example of an early synoptic map, which depicts the conditions on the morning of March 13, 1888, during the famous blizzard that paralyzed New York City for days and caused about 400 deaths. The storm delivered 60 to 120 centimetres of blowing, drifting snow, amid very cold temperatures, to a wide area from western New England to New Jersey on the heels of mild springlike weather. The map shows that the storm centre, where the sea-level pressure is lowest, was off the eastern tip of Long Island on the morning of the 13th, with a new centre forming far out at sea. Lines of equal pressure (isobars) circle the storm, as do the winds. The crowded isobars around the very low pressure at the centre imply strong winds. The dashed lines of equal temperature (isotherms) show that the storm occurred in connection with a great outbreak of subzero (in Fahrenheit) air from the continental interior, which was carried far into the Atlantic by the storm circulation. Tracing the 60° and 50° isotherms in the Atlantic shows that it was 55° F that morning in Bermuda, a rare event even at that distance from the storm.

Synoptic maps of this sort created a comprehensive picture. The synoptic meteorologist became an expert at drawing weather maps and interpreting prevailing conditions to be able to infer future conditions.

Since the mid-20th century, digital computers have made it possible to calculate changes in atmospheric conditions mathematically and objectively—*i.e.,* in such a way that anyone can obtain the same result from the same initial conditions. The widespread adoption of numerical weather prediction models brought a whole new group of players—computer specialists and experts in numerical processing and statistics—to the scene to work with

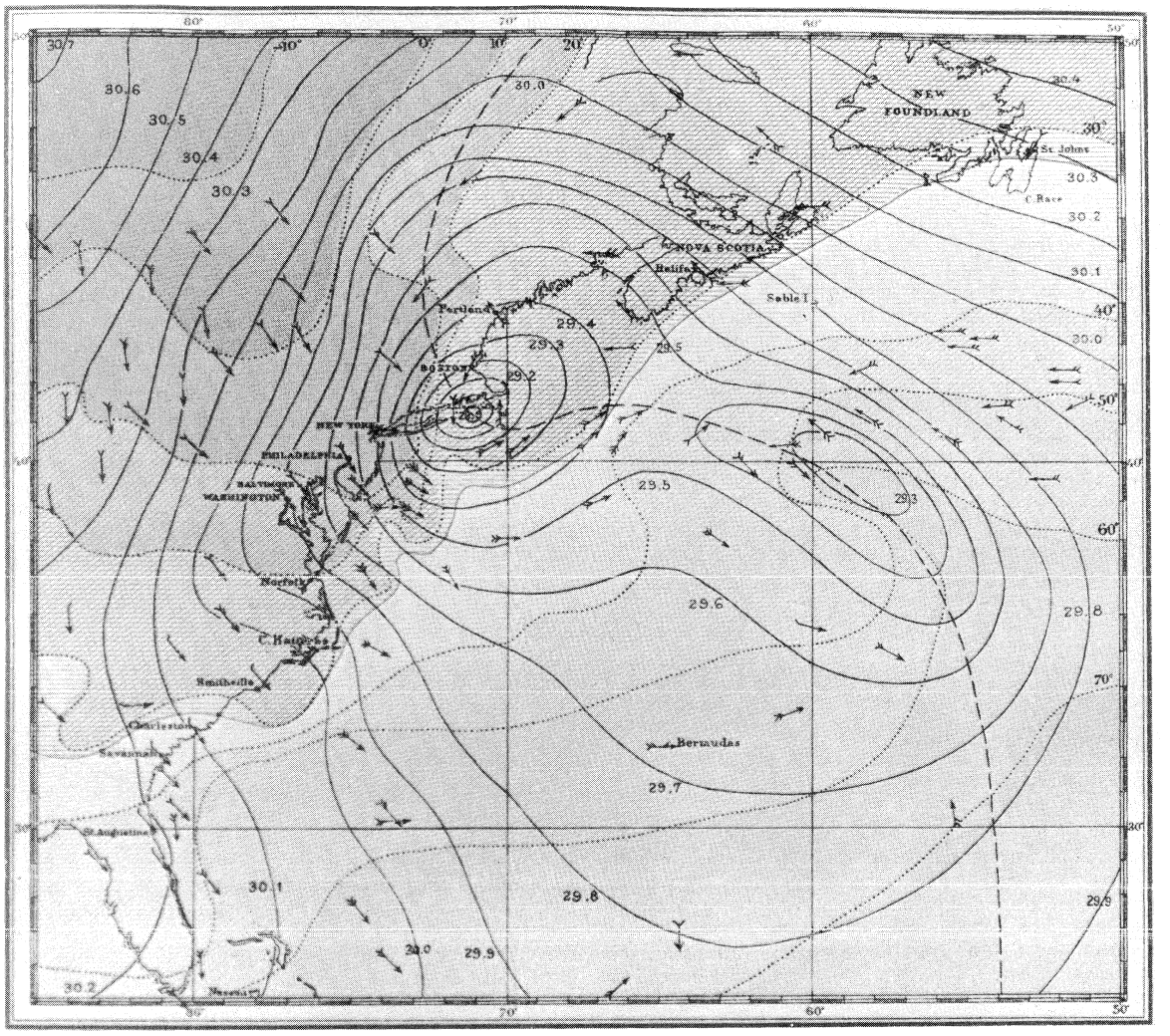



Figure 62: Surface weather map showing the meteorological conditions at 7:00 AM, EST, on March 13, 1888, during the great storm of March 11–14; from the first issue of *National Geographic Magazine*, October 1888 (see text).

By courtesy of the National Geographic Society

atmospheric scientists and meteorologists. Moreover, the enhanced capability to process and analyze weather data stimulated the long-standing interest of meteorologists in securing more observations of greater accuracy. Technological advances since the 1960s have led to a growing reliance on remote sensing, particularly the gathering of data with specially instrumented Earth-orbiting satellites. By the late 1980s, forecasts of weather were largely based on the determinations of numerical models integrated by high-speed supercomputers, except some shorter-range predictions, particularly of local weather conditions, which were made by specialists directly interpreting radar and satellite measurements.

### PRACTICAL APPLICATIONS OF WEATHER FORECASTING

Systematic weather records were kept after instruments for measuring atmospheric conditions became available during the 17th century. Undoubtedly these early records were employed mainly by those engaged in agriculture. Planting and harvesting obviously can be planned better and carried out more efficiently if long-term weather patterns can be estimated. In the United States, national weather services were first provided by the Army Signal Corps beginning in 1870. These operations were taken over by the Department of Agriculture in 1891. By the early 1900s free mail service and telephone were providing forecasts daily to millions of American farmers. The U.S. Weather Bureau established a Fruit-Frost (forecasting) Service during World War I, and by the 1920s radio broadcasts to agricultural interests were being made in most states.

Weather forecasting became an important tool for aviation during the 1920s and '30s. Its application in this area gained in importance after Francis W. Reichelderfer was appointed chief of the U.S. Weather Bureau in 1939. Reichelderfer had previously modernized the navy's meteorological service and made it a model of support for naval aviation. During World War II the discovery of very strong wind currents at high altitudes (the jet streams, which can affect aircraft speed) and the general susceptibility of military operations in Europe to weather led to a special interest in weather forecasting.

The most famous wartime forecasting problem was for Operation Overlord, the invasion of the European mainland at Normandy by Allied forces. An unusually intense June storm brought high seas and gales to the French coast, but a moderation of the weather that was successfully predicted by Colonel J.M. Stagg of the British forces (after consultation with both British and American forecasters) enabled General Dwight D. Eisenhower, supreme commander of the Allied Expeditionary Forces, to make his critical decision to invade on June 6, 1944.

Importance of commercial weather-forecasting organizations

The second half of the 20th century has seen unprecedented growth of commercial weather-forecasting firms in the United States and elsewhere. Marketing organizations and stores commonly hire weather-forecasting consultants to help with the timing of sales and promotions of products ranging from snow tires and roofing materials to summer clothes and resort vacations. Many oceangoing shipping vessels as well as military ships use optimum ship routing forecasts to plan their routes in order to minimize lost time, potential damage, and fuel consumption

in heavy seas. Similarly, airlines carefully consider atmospheric conditions when planning long-distance flights so as to avoid the strongest head winds and to ride with the strongest tail winds.

International trading of foodstuffs such as wheat, corn (maize), beans, sugar, cocoa, and coffee can be severely affected by weather news. For example, in 1975 a severe freeze in Brazil caused the price of coffee to increase substantially within just a few weeks, and in 1977 a freeze in Florida nearly doubled the price of frozen concentrated orange juice in a matter of days. Weather-forecasting organizations are thus frequently called upon by banks, commodity traders, and food companies to give them advance knowledge of the possibility of such sudden changes.

The cost of all sorts of commodities and services, whether they are tents for outdoor events or plastic covers for the daily newspapers, can be reduced or eliminated if reliable information about possible precipitation can be obtained in advance.

Forecasts must be quite precise for applications that are tailored to specific industries. Gas and electric utilities, for example, may require forecasts of temperature within one or two degrees of the mean a day ahead of time, or ski-resort operators may need prior estimates of nighttime relative humidity on the slopes within 5 to 10 percent in order to schedule snow making.

## History of weather forecasting

### EARLY MEASUREMENTS AND IDEAS

The Greek philosophers had much to say about meteorology, and many who subsequently engaged in weather forecasting no doubt made use of their ideas. Unfortunately, they probably made many bad forecasts, because Aristotle, who was the most influential, did not believe that wind is air in motion. He did believe, however, that west winds are cold because they blow from the sunset.

Key inventions and discoveries

The scientific study of meteorology did not develop until measuring instruments became available. Its beginning is commonly associated with the invention of the mercury barometer by Evangelista Torricelli, an Italian physicist-mathematician, in the mid-17th century and the nearly concurrent development of a reliable thermometer. (Galileo had constructed an elementary form of gas thermometer in 1607, but it was defective; the efforts of many others finally resulted in a reasonably accurate liquid-in-glass device.)

A succession of notable achievements by chemists and physicists of the 17th and 18th centuries contributed significantly to meteorological research. The formulation of the laws of gas pressure, temperature, and density by Robert Boyle and Jacques-Alexandre-César Charles, the development of calculus by Isaac Newton and Gottfried Wilhelm Leibniz, the development of the law of partial pressures of mixed gases by John Dalton, and the formulation of the doctrine of latent heat (*i.e.,* heat release by condensation or freezing) by Joseph Black are just a few of the major scientific breakthroughs of the period that made it possible to measure and better understand hitherto unknown aspects of the atmosphere and its behaviour. During the 19th century, all of these brilliant ideas began to produce results in terms of useful weather forecasts.

### THE EMERGENCE OF SYNOPTIC FORECASTING METHODS

**Analysis of synoptic weather reports.** An observant person who has learned nature's signs can interpret the appearance of the sky, the wind, and other local effects and "foretell the weather." A scientist can use instruments at one location to do so even more effectively. The modern approach to weather forecasting, however, can only be realized when many such observations are exchanged quickly by experts at various weather stations and entered on a synoptic weather map to depict the patterns of pressure, wind, temperature, clouds, and precipitation at a specific time. Such a rapid exchange of weather data became feasible with the development of the electric telegraph in 1837 by Samuel F.B. Morse of the United States.

Importance of the telegraph

By 1849 Joseph Henry of the Smithsonian Institution in Washington, D.C., was plotting daily weather maps based on telegraphic reports, and in 1869 Cleveland Abbe at the Cincinnati Observatory began to provide regular weather forecasts using data received telegraphically.

Synoptic weather maps resolved one of the great controversies of meteorology—namely, the rotary storm dispute. By the early decades of the 19th century, it was known that storms were associated with low barometric readings, but the relation of the winds to low-pressure systems, called cyclones, remained unrecognized. William Redfield, a self-taught meteorologist from Middletown, Conn., noticed the pattern of fallen trees after a New England hurricane and suggested in 1831 that the wind flow was a rotary counterclockwise circulation around the centre of lowest pressure. The American meteorologist James P. Espy subsequently proposed in his *Philosophy of Storms* (1841) that air would flow toward the regions of lowest pressure and then would be forced upward, causing clouds and precipitation. Both Redfield and Espy proved to be right (see Figure 62). The air does spin around the cyclone, as Redfield believed, while the layers close to the ground flow inward and upward as well. The net result is a rotational wind circulation that is slightly modified at the Earth's surface to produce inflow toward the storm centre, just as Espy had proposed. Further, the inflow is associated with clouds and precipitation in regions of low pressure, though that is not the only cause of clouds there.

In Europe the writings of Heinrich Dove, a Polish scientist who directed the Prussian Meteorological Institute, greatly influenced views concerning wind behaviour in storms. Unlike the Americans, Dove did not focus on the pattern of the winds around the storm but rather on how the wind should change at one place as a storm passed. It was many years before his followers understood the complexity of the possible changes.

**Establishment of weather-station networks and services.** Routine production of synoptic weather maps became possible after networks of stations were organized to take measurements and report them to some type of central observatory. As early as 1814, U.S. Army Medical Corps personnel were ordered to record weather data at their posts; this activity was subsequently expanded and made more systematic. Actual weather-station networks were established in the United States by New York University, the Franklin Institute, and the Smithsonian Institution during the early decades of the 19th century.

In Britain, James Glaisher organized a similar network, as did Christophorus H.D. Buys Ballot in The Netherlands. Other such networks of weather stations were developed near Vienna, Paris, and St. Petersburg (now Leningrad).

Establishment of national meteorological services

It was not long before national meteorological services were established on the Continent and in the United Kingdom. The first national weather service in the United States commenced operations in 1871, with responsibility assigned to the U.S. Army Signal Corps. The original purpose of the service was to provide storm warnings for the Atlantic and Gulf coasts and for the Great Lakes. Within the next few decades, national meteorological services were established in such countries as Japan, India, and Brazil. The importance of international cooperation in weather prognostication was recognized by the directors of such national services. By 1880 they had formed the International Meteorological Organization (IMO) for this end.

The proliferation of weather-station networks linked by telegraphy made synoptic forecasting a reality by the close of the 19th century. Yet, the daily weather forecasts generated left much to be desired. Many errors occurred as predictions were largely based on the experience that each individual forecaster had accumulated over several years of practice, vaguely formulated rules of thumb (*e.g.,* of how pressure systems move from one region to another), and associations that were poorly understood, if at all.

### PROGRESS DURING THE EARLY 20TH CENTURY

An important aspect of weather prediction is to calculate the atmospheric pressure pattern—the positions of the highs and lows and their changes. Modern research has shown that sea-level pressure patterns respond to the motions of the upper-atmospheric winds, with their narrow, fast-moving jet streams and remarkable waves that flow

through the air and pass air through themselves (see above *Upper-air waves*).

Early attempts to study the upper atmosphere

Frequent surprises and errors in estimating surface atmospheric pressure patterns undoubtedly caused 19th-century forecasters to seek information about the upper atmosphere for possible explanations. The British meteorologist Glaisher made a series of ascents by balloon during the 1860s, reaching an unprecedented height of nine kilometres. At about this time investigators on the Continent began using unmanned balloons to carry recording barographs, thermographs, and hygrographs to high altitudes. During the late 1890s meteorologists in both the United States and Europe used kites equipped with instruments to probe the atmosphere up to altitudes of about three kilometres. Notwithstanding these efforts, knowledge about the upper atmosphere remained very limited at the turn of the century. The situation was aggravated by the confusion created by observations from weather stations located on mountains or hilltops. Such observations often did not show what was expected, partly because so little was known about the upper atmosphere and partly because the mountains themselves affect measurements, producing results that are not representative of what would be found in the free atmosphere at the same altitude.

Fortunately, a large enough number of scientists had already put forth ideas that would make it possible for weather forecasters to think three-dimensionally, even if sufficient meteorological measurements were lacking. Henrik Mohn, the first of a long line of highly creative Norwegian meteorologists, Wladimir Köppen, the noted German climatologist (see above), and Max Margules, an influential Russian-born meteorologist, all contributed to the view that mechanisms of the upper air generate the energy of storms.

In 1911 William H. Dines, a British meteorologist, published data that showed how the upper atmosphere compensates for the fact that the low-level winds carry air toward low-pressure centres. Dines recognized that the inflow near the ground is more or less balanced by a circulation upward and outward aloft. Indeed, for a cyclone to intensify, which would require a lowering of central pressure, the outflow must exceed the inflow; the surface winds can converge quite strongly toward the cyclone, but sufficient outflow aloft can produce falling pressure at the centre.

Meteorologists of the time were now aware that vertical circulations and upper-air phenomena were important,

After J. Bjerknes (1919) in S. Petterssen, *Introduction to Meteorology*, 2nd ed. (1958), McGraw-Hill Book Company; reproduced with permission

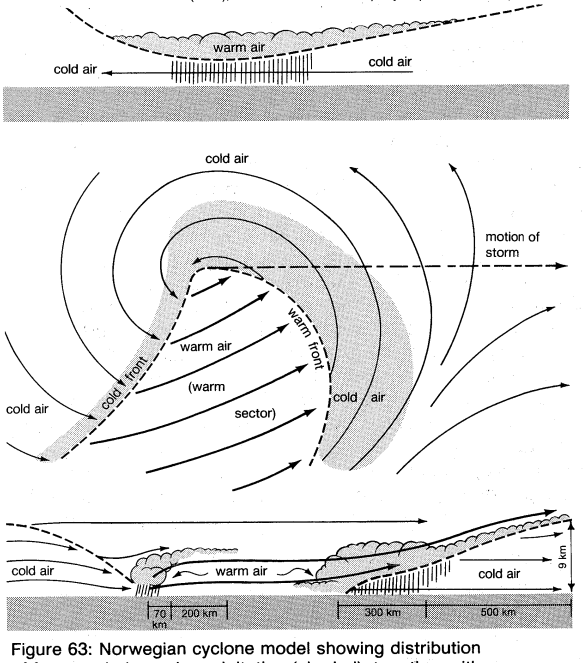



Figure 63: Norwegian cyclone model showing distribution of fronts, winds, and precipitation (shaded), together with sectional views to the (top) north and (bottom) south of clouds and precipitation.

but they still had not determined how such knowledge could improve weather forecasting. Then, in 1919, the Norwegian meteorologist Jacob Bjerknes introduced what has been referred to as the Norwegian cyclone model. This theory (depicted in Figure 63) pulled together many earlier ideas and related the patterns of wind and weather to a low-pressure system that exhibited fronts—rather sharp sloping boundaries between cold and warm air masses. Bjerknes pointed out the rainfall/snowfall patterns that are characteristically associated with the fronts in cyclones: the rain or snow occurs over large areas on the cold side of an advancing warm front ahead of advancing low pressure. Here, the winds are from the south and the warm air, being light, glides up over a large region of cold air. Widespread, sloping clouds spread ahead of the cyclone; barometers fall as the storm approaches, and precipitation from the rising warm air falls through the cold air below. Where the cold air advances to the rear of the storm, squalls and showers mark the abrupt lifting of the warm air being displaced. Thus, the concept of fronts focused attention on the action at air mass boundaries. The Norwegian cyclone model could be called the frontal model, for the idea of warm air masses being lifted over cold air along their edges (fronts) became a major forecasting tool. The model not only emphasized the idea but it also showed how and where to apply it.

In later work, Bjerknes and several other members of the so-called Bergen school of meteorology expanded the model to show that cyclones grow from weak disturbances on fronts, pass through a regular life cycle, and ultimately die by the inflow filling them (see above *Extratropical cyclones* and Figure 22). Both the Norwegian cyclone model and the associated life-cycle concept are still used today by weather forecasters.

While Bjerknes and his Bergen colleagues refined the cyclone model, other Scandinavian meteorologists provided much of the theoretical basis for modern weather prediction. Foremost among them were Vilhelm Bjerknes, Jacob's father, and Carl-Gustaf Rossby. Their ideas helped make it possible to understand and carefully calculate the changes in atmospheric circulation and the motion of the upper-air waves that control the behaviour of cyclones.

## MODERN TRENDS AND DEVELOPMENTS

**Upper-air observations by means of balloon-borne sounding equipment.** Once again technology provided the means with which to test the new scientific ideas and stimulate yet newer ones. During the late 1920s and '30s, several groups of investigators (those headed by Yrjö Väisälä of Finland and Pavel Aleksandrovich Malchanov of the Soviet Union, for example) began using small radio transmitters with balloon-borne instruments, eliminating the need to recover the instruments and speeding up access to the upper-air data. These radiosondes, as they came to be called, gave rise to the upper-air observation networks that still exist today. Approximately 75 stations in the United States and more than 500 worldwide release, twice daily, balloons that reach heights of 30,000 metres or more. Observations of temperature and relative humidity at various pressures are radioed back to the station from which the balloons are released as they ascend at a predetermined rate. The balloons also are tracked by radar in order to ascertain winds from their drift.

Use of radiosondes

Forecasters are able to produce synoptic weather maps of the upper atmosphere twice each day on the basis of radiosonde observations. While new methods of upper-air measurement have been developed, the primary synoptic clock times for producing upper-air maps are still the radiosonde-observation times—namely, 0000 (midnight) and 1200 (noon) Greenwich Mean Time (GMT). Furthermore, modern computer-based forecasts use 0000 and 1200 GMT as the starting times from which they calculate the changes that are at the heart of modern forecasts. It is, in effect, the synoptic approach carried out in a different way, intimately linked to the radiosonde networks developed during the 1930s and '40s.

**Application of radar.** As in many fields of endeavour, weather prediction experienced several breakthroughs during and immediately after World War II. The British

began using microwave radar in the late 1930s to monitor enemy aircraft, but it was soon learned that radar gave excellent returns from raindrops at certain wavelengths (five to 10 centimetres). As a result it became possible to track and study the evolution of individual showers or thunderstorms, as well as to "see" the precipitation structure of larger storms. Figure 64 shows an image of the rain bands (not clouds) in a hurricane.

By courtesy of National Oceanic and Atmospheric Administration; photograph, AP/Wide World Photos

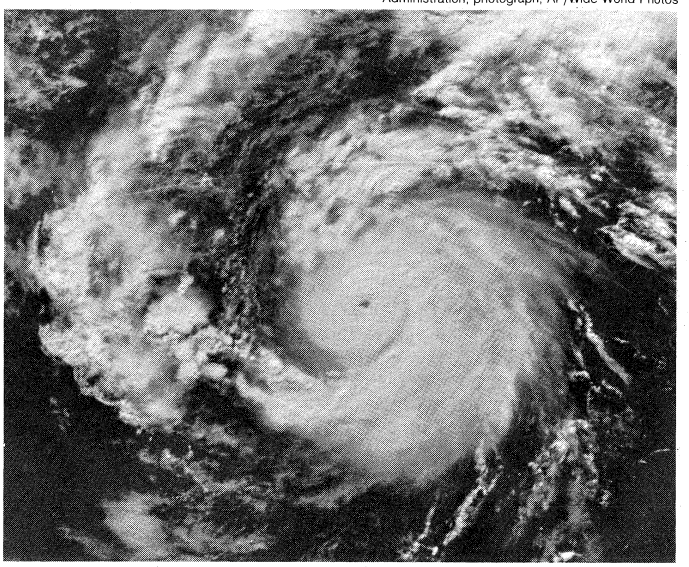


Figure 64: The well-defined eye and the rain bands of Hurricane Hyacinth about 805 kilometres (500 miles) south of the southern tip of Baja California, Mexico, photographed from an Earth-orbiting satellite on Aug. 9, 1976.

Radar surveillance of tornadoes and thunderstorms

Since its initial application in meteorological work, radar has grown as a forecaster's tool. Virtually all tornadoes and severe thunderstorms over the United States and in some other parts of the world are monitored by radar. Radar observation of the growth, motion, and characteristics of such storms provide clues as to their severity. Modern radar systems use the Doppler principle of frequency shift associated with movement toward or away from the radar transmitter/receiver to determine wind speeds as well as storm motions.

Using radar and other observations, the Japanese-American meteorologist Tetsuya T. Fujita discovered many details of severe thunderstorm behaviour and of the structure of the violent local storms common to the Midwest region of the United States. His Doppler-radar analyses of winds revealed "microburst" gusts. These gusts cause the large wind shears (differences) associated with strong rains that have been responsible for some plane crashes. Figure 65 shows how such a microburst caused the crash of an airliner at Kennedy Airport in New York on June 24, 1975.

Other types of radar have been used increasingly for detecting winds continuously, as opposed to twice a day. These wind-profiling radar systems actually pick up signals "reflected" by clear air and so can function even when no clouds or rain are present.

**Meteorological measurements from satellites and aircraft.** A major breakthrough in meteorological measurement came with the launching of the first meteorological satellite, the TIROS (Television and Infrared Observation Satellite), by the United States on April 1, 1960. The impact of global quantitative views of temperature, cloud, and moisture distributions, as well as of surface properties (*e.g.,* ice cover and soil moisture), has already been substantial. Furthermore, new ideas and new methods may very well make the 21st century the "age of the satellite" in weather prediction.

Medium-range forecasts that provide information five to seven days in advance were impossible before satellites began making global observations—particularly over the ocean waters of the Southern Hemisphere—routinely available in real time. Global forecasting models developed at the U.S. National Center for Atmospheric Research (NCAR), the European Centre for Medium Range Weather Forecasts (ECMWF), and the U.S. National Meteorological Center (NMC) became the standard during the 1980s, making medium-range forecasting a reality.

Meteorological satellites travel in various orbits and carry a wide variety of sensors. They are of two principal types: the low-flying polar orbiter, and the geostationary orbiter.

Polar orbiter

The first type circle the Earth at altitudes of 500–1,000 kilometres and in roughly north–south orbits. They appear overhead at any one locality twice a day and provide very high-resolution data because they fly close to the Earth. Such satellites are vitally necessary for much of Europe and other high-latitude locations because they orbit near the poles. These satellites do, however, suffer from one major limitation: they can provide a sampling of atmospheric conditions only twice daily.

The geostationary satellite is made to orbit the Earth along its equatorial plane at an altitude of about 36,000 kilometres. At that height the eastward motion of the satellite coincides exactly with the Earth's rotation, so that the satellite remains in one position above the Equator. Satellites of this type are able to provide an almost continuous view of a wide area (see above Figure 24). Because of this capability, geostationary satellites have yielded new information about the rapid changes that occur in thunderstorms, hurricanes, and certain types of fronts, making them invaluable to weather forecasting as well as meteorological research.

One weakness common to virtually all satellite-borne sensors and to some ground-based radars that use UHF/VHF waves is an inability to measure thin layers of the atmosphere. One such layer is the tropopause, the boundary between the relatively dry stratosphere and the meteorologically active layer below. This is often the region of the jet streams. Important information about these kinds of high-speed air currents is obtained with sensors mounted on high-flying commercial aircraft and is routinely included in global weather analyses.

**Numerical weather prediction (NWP) models.** Thinkers frequently advance ideas long before the technology exists to implement them. Few better examples exist than that of numerical weather forecasting. Instead of mental estimates or rules of thumb about the movement of storms, numerical forecasts are objective calculations of changes to the weather map based on sets of equations called models. Shortly after World War I a British scientist named Lewis F. Richardson completed such a forecast that he had been working on for years by tedious and difficult hand

From Theodore Fujita, *The Downburst* (1985)

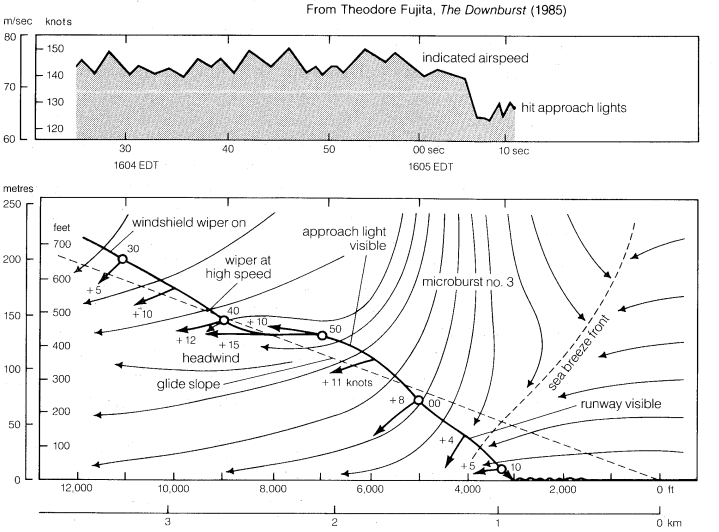



Figure 65: *The flight path of Eastern Air Lines flight 66 at Kennedy Airport, New York City, on June 24, 1975.* The flight ended in a crash caused by "microburst" gusts, the downburst of air from a thunderstorm. As the airplane approached the downburst, the head wind lifted it above the glide slope. After the pilot corrected for this, the airplane lost the head wind and received a downward push from the downburst. Both effects pushed the airplane far below the glide slope.

calculations. Although the forecast proved to be incorrect, Richardson's general approach was accepted decades later when the electronic computer became available. In fact, it has become the basis for nearly all present-day weather forecasts. Human forecasters may interpret or even modify the results of the computer models, but there are few forecasts that do not begin with numerical-model calculations of pressure, temperature, wind, and humidity for some future time.

The method is closely related to the synoptic approach (see above). Data are collected rapidly by a Global Telecommunications System for 0000 or 1200 GMT to specify the initial conditions. The model equations are then solved for various segments of the weather map—often a global map—to calculate how much it will change in a given time, say, 10 minutes. With such changes added to the initial conditions, a new map is generated (in the computer's memory) valid for 0010 or 1210 GMT. This map is treated as a new set of initial conditions, probably not quite as accurate as the measurements for 0000 and 1200 GMT but still very accurate. A new step is undertaken to generate a forecast for 0020 or 1220. This process is repeated step after step. In principle, the process could continue indefinitely. In practice, small errors creep into the calculations, and they accumulate. Eventually, the errors become so large by this cumulative process that there is no point in continuing.

Global numerical forecasts are produced regularly (once or twice daily) at the ECMWF, the NMC, and the U.S. military facilities in Omaha, Neb., and Monterey, Calif., as well as in Tokyo, Moscow, and Melbourne. In addition, specialized numerical forecasts designed to predict more details of the weather are made for many smaller regions of the world by various national weather services, military organizations, and even a few private companies. Finally, research versions of numerical weather prediction models are constantly under review, development, and testing at NCAR and at the Goddard Space Flight Center in the United States and at universities in several nations.

The capacity and complexity of numerical weather prediction models have increased dramatically since the mid-1940s when the earliest modeling work was done by the mathematician John von Neumann and the meteorologist Jule Charney at the Institute for Advanced Study in Princeton, N.J. Because of their pioneering work and the discovery of important simplifying relationships by other scientists (notably Arnt Eliassen of Norway and Reginald Sutcliffe of Britain), a joint U.S. Weather Bureau, Navy, and Air Force numerical forecasting unit was formed in 1954 in Washington, D.C. Referred to as JNWP, this unit was charged with producing operational numerical forecasts on a daily basis.

The era of numerical weather prediction thus really began in the 1950s. As computing power grew, so did the complexity, speed, and capacity for detail of the models. And as new observations became available from such sources as Earth-orbiting satellites, radar systems, and drifting weather balloons, so too did methods sophisticated enough to ingest the data into the models as improved initial synoptic maps.

Numerical forecasts have improved steadily over the years. The vast Global Weather Experiment, first conceived by Charney, was carried out by many nations in 1979 under the leadership of the World Meteorological Organization to demonstrate what high-quality global observations could do to improve forecasting by numerical prediction models. The results of that effort continue to effect further improvement.

Mesoscale numerical prediction models

A relatively recent development has been the construction of mesoscale numerical prediction models. The prefix meso- means "middle" and here refers to middle-sized features in the atmosphere, between large cyclonic storms and individual clouds. Fronts, clusters of thunderstorms, hurricane bands, and jet streams are mesoscale structures, and their evolution and behaviour are crucial forecasting problems that only recently have been dealt with in numerical prediction. Richard A. Anthes developed a widely used American research version, the NCAR/Penn State Mesoscale Model, which shows the potential for detailed predictions of the location and evolution of structures such as fronts, rainbands, and ice storms. By the late 1980s the national weather services of several countries were producing numerical forecasts of considerable detail by means of limited-area mesoscale models.

## Principles and methodology of weather forecasting

### SHORT-RANGE FORECASTING

**Objective predictions.** When people wait under a shelter for a downpour to end, they are making a very-short-range weather forecast. They are assuming, based on past experience, that such hard rain usually does not last very long. In short-term predictions the challenge for the forecaster is to improve on what the layperson can do. For years the type of situation represented in the above example proved particularly vexing for forecasters, but since the mid-1980s they have been developing a method called nowcasting to meet precisely this sort of challenge. In this method, radar and satellite observations of local atmospheric conditions are processed and displayed rapidly by computers to project weather several hours in advance. The U.S. National Oceanic and Atmospheric Administration operates a facility known as PROFS (Program for Regional Observing and Forecasting Services) in Boulder, Colo., specially equipped for nowcasting.

Nowcasting

Meteorologists can make somewhat longer-term forecasts (those for six, 12, 24, or even 48 hours) with considerable skill because they are able to measure and predict atmospheric conditions for large areas by computer. Using models that apply their accumulated expert knowledge quickly, accurately, and in a statistically valid form, meteorologists are now capable of making forecasts objectively (see above). As a consequence, the same results are produced time after time from the same data inputs, with all analysis accomplished mathematically. Unlike the prognostications of the past made with subjective methods, objective forecasts are consistent and can be studied, reevaluated, and improved.

Model Output Statistics

Another technique for objective short-range forecasting is called MOS (for Model Output Statistics). Conceived by Harry R. Glahn and D.A. Lowry of the U.S. National Weather Service, this method involves the use of data relating to past weather phenomena and developments to extrapolate the values of certain weather elements, usually for a specific location and time period. It overcomes the weaknesses of numerical models by developing statistical relations between model forecasts and observed weather. These relations are then used to translate the model forecasts directly to specific weather forecasts. For example, a numerical model might not predict the occurrence of surface winds at all, and whatever winds it did predict might always be too strong. MOS relations can automatically correct for errors in wind speed and produce quite accurate forecasts of wind occurrence at a specific point, such as Heathrow Airport near London. As long as numerical weather prediction models are imperfect, there may be many uses for the MOS technique.

**Predictive skills and procedures.** Short-range weather forecasts generally tend to lose accuracy as forecasters attempt to look farther ahead in time. Predictive skill is greatest for periods of about 12 hours and is still quite substantial for 48-hour predictions (Figure 66). An increasingly important group of short-range forecasts are economically motivated. Their reliability is determined in the marketplace by the economic gains they produce (or the losses they avert).

Weather warnings

Weather warnings are a special kind of short-range forecast; the protection of human life is the forecaster's greatest challenge and source of pride. The first national weather forecasting service in the United States (the predecessor of the Weather Bureau) was in fact formed, in 1870, in response to the need for storm warnings on the Great Lakes. Increase Lapham of Milwaukee urged Congress to take action to reduce the loss of hundreds of lives incurred each year by Great Lakes shipping during the 1860s. The effectiveness of the warnings and other forecasts assured the future of the American public weather service.

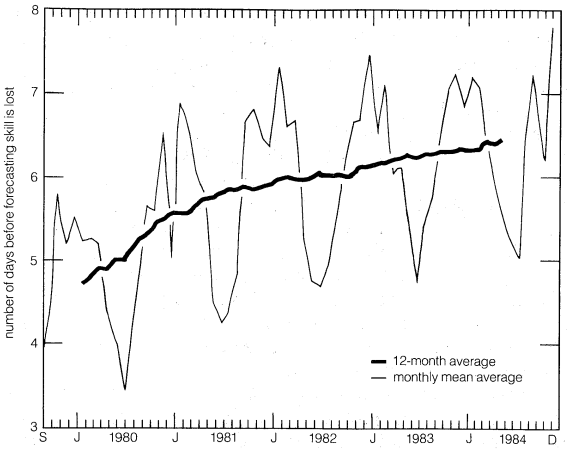



Figure 66: Predictive skill of the European Centre for Medium Range Weather Forecasting (ECMWF) model, September 1979–December 1984, shown in terms of the number of days into the forecast before accuracy is lost.

From L.O. Bengtsson, "Medium-Range Forecasting—The Experience at ECMWF," *Bulletin of the American Meteorological Society* (September 1985)

Weather warnings are issued by government and military organizations throughout the world for all kinds of threatening weather events: tropical storms variously called hurricanes, typhoons, or tropical cyclones, depending on location (see above); great oceanic gales outside the tropics spanning hundreds of kilometres and at times packing winds comparable to those of tropical storms; and, on land, flash floods, high winds, fog, blizzards, ice, and snowstorms.

A particular effort is made to warn of hail, lightning, and wind gusts associated with severe thunderstorms, sometimes called severe local storms (SELS) or simply severe weather. Forecasts and warnings also are made for tornadoes, those intense, rotating windstorms that represent the most violent end of the weather scale (see above). Destruction of property and the risk of injury and death are extremely high in the path of a tornado, especially in the case of the largest systems (sometimes called maxi-tornadoes).

Because tornadoes are so uniquely life-threatening and because they are so common in various regions of the United States, the National Weather Service operates a National Severe Storms Forecasting Center (NSSFC) in Kansas City, Mo., where SELS forecasters survey the atmosphere for the conditions that can spawn tornadoes or severe thunderstorms. This group of SELS forecasters, assembled in 1952, monitors temperature and water vapour in an effort to identify the warm, moist regions where thunderstorms may form and studies maps of pressure and winds to find regions where the storms may organize into mesoscale structures. The group also monitors jet streams and dry air aloft that can combine to distort ordinary thunderstorms into rare rotating ones with tilted chimneys of upward rushing air that, because of the tilt, are unimpeded by heavy falling rain. These high-speed updrafts can quickly transport vast quantities of moisture to the cold upper regions of the storms, thereby promoting the formation of large hailstones. The hail and rain drag down air from aloft to complete a circuit of violent, cooperating updrafts and downdrafts.

By correctly anticipating such conditions, SELS forecasters are able to provide time for the mobilization of special observing networks and personnel. If the storms actually develop, specific warnings are issued based on direct observations. This two-step process consists of the tornado or severe thunderstorm watch, which is the forecast prepared by the SELS forecaster, and the warning, which is usually released by a local observing facility. The watch may be issued when the skies are clear, and it usually covers a number of counties. It alerts the affected area to the threat but does not attempt to pinpoint which communities will be affected.

Severe local storm watch

By contrast, the warning is very specific to a locality and calls for immediate action. Radar of various types can be used to detect the large hailstones, the heavy load of raindrops, the relatively clear region of rapid updraft, and even the rotation in a tornado. These indicators, or an actual sighting, often trigger the tornado warning. In effect, a warning is a specific statement that danger is imminent, whereas a watch is a forecast that warnings may be necessary later in a given region.

## LONG-RANGE FORECASTING

**Techniques.** Extended-range, or long-range, weather forecasting has had a different history and a different approach from short- or medium-range forecasting. In most cases, it has not applied the synoptic method of going forward in time from a specific initial map. Instead, long-range forecasters have tended to use the climatological approach, often concerning themselves with the broad weather picture over a period of time rather than attempting to forecast day-to-day details.

Application of the climatological approach

There is good reason to believe that the limit of day-to-day forecasts based on the "initial map" approach is about two weeks. Most long-range forecasts thus attempt to predict the departures from normal conditions for a given month or season. Such departures are called anomalies. A forecast might state that "spring temperatures in Minneapolis have a 65 percent probability of being above normal." It would likely be based on a forecast anomaly map such as that shown in Figure 67, which shows positive temperature anomalies in the northern United States and negative anomalies to the south. No attempt is made to specify which spring days will be warmer than normal. Moreover, there is no implication that, even if the forecast is correct, no very cold weather will occur that spring. The only implication is that, when the season is over, in two

From (top) Extended Forecast Group, U.S. National Weather Service, National Oceanic and Atmospheric Administration; (bottom) J. Namias, "Remarks on the Potential for Long-Range Forecasting," *Bulletin of the American Meteorological Society* (February 1985)

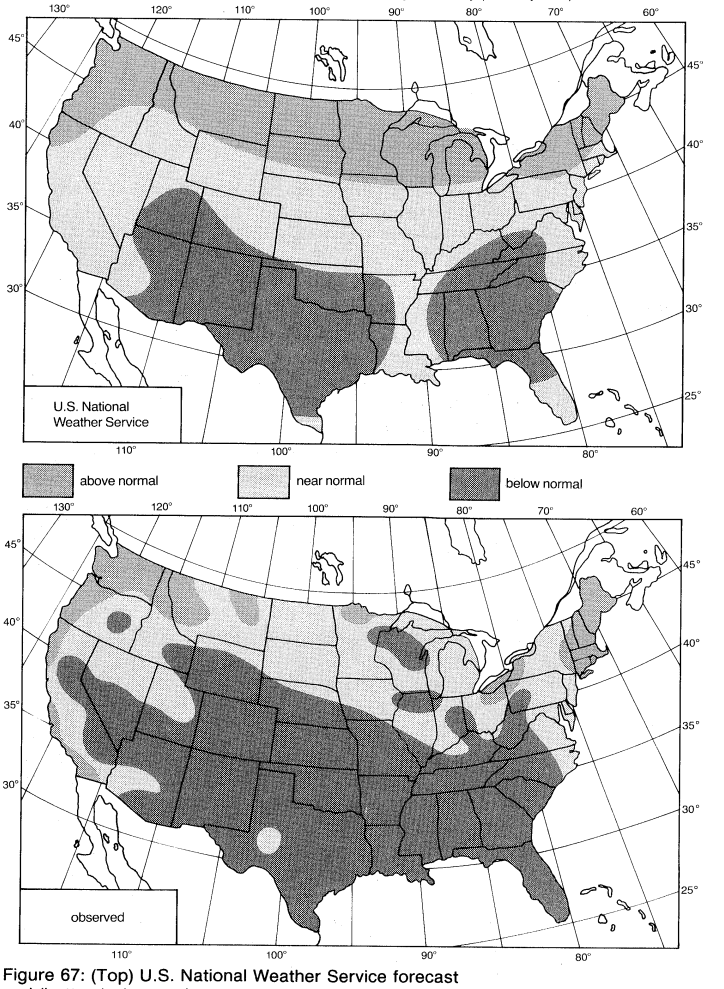



Figure 67: (Top) U.S. National Weather Service forecast and (bottom) observed temperature anomaly pattern for the spring of 1983.

of the three years for which such a forecast is made, the season will have been warmer than usual.

The U.S. Weather Bureau began making experimental long-range forecasts just before the beginning of World War II, and its successor, the National Weather Service, continues to express such predictions in probabilistic terms, making it clear that they are subject to uncertainty. Verification shows that forecasts of temperature anomalies are more reliable than those of precipitation, that monthly forecasts are better than seasonal ones, and that winter months are predicted somewhat more accurately than other seasons.

Extensive use of the analog method

Prior to the 1980s the technique commonly used in long-range forecasting relied heavily on the analog method, in which groups of weather situations (maps) from previous years were compared to those of the current year to determine similarities with the atmosphere's present patterns (or "habits"). An association was then made between what had happened subsequently in those "similar" years and what was going to happen in the current year. Most of the techniques were quite subjective, and there were often disagreements of interpretation and consequently uneven quality and marginal reliability.

Persistence (warm summers follow warm springs) or anti-persistence (cold springs follow warm winters) also were used, even though, strictly speaking, most forecasters consider persistence forecasts "no-skill" forecasts. Yet, they too have had limited success.

**Prospects for new procedures.** In the last quarter of the 20th century the approach of and prospects for long-range weather forecasting have changed significantly. Stimulated by the work of Jerome Namias, who headed the U.S. Weather Bureau's Long-Range Forecast Division for 30 years, scientists began to look at ocean-surface temperature anomalies as a potential cause for the temperature anomalies of the atmosphere in succeeding seasons and at distant locations. At the same time, other American meteorologists, most notably John M. Wallace, showed how certain repetitive patterns of atmospheric flow were related to each other in different parts of the world. With satellite-based observations available, investigators began to study the El Niño phenomenon (see above *El Niño/Southern Oscillation*). Atmospheric scientists also revived the work of Gilbert Walker, an early 20th-century British climatologist who had studied the Southern Oscillation, the aforementioned up-and-down fluctuation of atmospheric pressure in the Southern Hemisphere. Walker had investigated related air circulations (later called the Walker Circulation) that resulted from abnormally high pressures in Australia and low pressures in Argentina or vice versa.

All of this led to new knowledge about how the occurrence of abnormally warm or cold ocean waters and of abnormally high or low atmospheric pressures could be interrelated in vast global connections. Knowledge about these links—El Niño/Southern Oscillation (ENSO)—and about the behaviour of parts of these vast systems enables forecasters to make better long-range predictions, at least in part, because the ENSO features change slowly and somewhat regularly. This approach of studying interconnections between the atmosphere and the ocean may represent the beginning of a revolutionary stage in long-range forecasting.

Since the mid-1980s, interest has grown in applying numerical weather prediction models to long-range forecasting. In this case, the concern is not with the details of weather predicted 20 or 30 days in advance but rather with objectively predicted anomalies. The reliability of long-range forecasts, like that of short- and medium-range projections, has improved substantially in recent years. Yet, many significant problems remain unsolved, posing interesting challenges for all those engaged in the field.

(J.J.Ca.)

# SCIENTIFIC WEATHER MODIFICATION

## General considerations

Humans have long sought to purposefully alter such atmospheric phenomena as clouds, rain, snow, hail, lightning, thunderstorms, tornadoes, hurricanes, and cyclones. The modern era of scientific weather modification began in 1946 with work by Vincent J. Schaefer and Irving Langmuir at the General Electric Research Laboratories in Schenectady, N.Y. Schaefer discovered that when Dry Ice pellets were dropped into a cloud composed of water droplets in a deep-freeze box, the droplets rapidly were replaced by ice crystals, which increased in size and then fell to the bottom of the box.

Schaefer–Langmuir experiments

The Schaefer–Langmuir experiments in the laboratory and the atmosphere demonstrated that so-called supercooled clouds—namely those composed of water droplets at temperatures below freezing—could be dissipated. When the supercooled clouds were seeded with grains of Dry Ice, ice crystals formed and grew large enough to fall out of the clouds.

Certain substances other than Dry Ice can be used to seed clouds. For example, when silver iodide and lead iodide are burned, they create a smoke of tiny particles. These particles produce ice crystals in supercooled clouds below temperatures of about −5° C. Although many other materials can cause ice crystals to form, the above-mentioned are the most widely used. For the most part, Dry Ice is dispersed from airplanes, but silver iodide nuclei may be generated on the ground and carried upward by air currents, introduced from airplanes, or produced by pyrotechnic devices such as rockets or exploding artillery shells.

A wide variety of scientific tests and operational weather modification projects have been performed in many countries. The largest programs have been in the United States, the Soviet Union, Australia, and France.

Most weather modification programs in the United States have dealt with rain or snow. Although there is still considerable debate about the effectiveness of cloud seeding, the evidence indicates that under certain meteorological conditions, ice nuclei seeding may increase precipitation by amounts ranging up to some tens of percent. In other circumstances decreases may occur, and, in still others, seeding has no effect.

In the Soviet Union the major programs of weather modification have been aimed at reducing the fall of damaging hail. Experimentation in this area has apparently been successful. The Soviet procedures involve the introduction of ice nuclei into potential hail clouds by means of artillery or rockets. Attempts to modify hailstorms in other countries have had mixed success.

A number of hurricanes have been seeded with ice nuclei by American scientists. Although results have been inconclusive, some meteorologists believe it is possible to reduce hurricane intensities in some instances. Before asserting that hurricanes can be beneficially modified, however, more research is needed. Limited attempts to modify mid-latitude cyclones in the past have not been successful. In the case of tornadoes, knowledge of their dynamic structure remains limited, and no attempts have been made to control them.

The effective extent of weather modification

There are many misunderstandings about the present status of weather modification. Present techniques are concerned mostly with the influence of existing cloud systems. There is no evidence and no reason for believing, at this stage, that cloud seeding may cause or end droughts. Such dry periods result from peculiarities in the general circulation of the atmosphere that lead to sinking air and cloud-free skies in areas accustomed to precipitation. When there are no clouds present, there can be no cloud seeding.

As meteorologists have developed schemes for changing the weather, the ecological, social, and legal problems have become more serious. Many U.S. states have laws governing weather modification activities. Lawsuits have been filed in which the parties have contested ownership of the clouds and the precipitation therein.

## Methods of modifying atmospheric phenomena

### CLOUD SEEDING

As previously discussed, clouds form when atmospheric moisture condenses on small particles in the atmosphere called condensation nuclei. Typically, a cloud is composed of tiny spheres of water that range in diameter from a few micrometres to a few tens of micrometres. The number of cloud droplets per cubic centimetre ranges from less than 100 to more than 1,000; 200 droplets per cubic centimetre is approximately an average value.

An important characteristic of a cloud is its temperature. When it is everywhere above 0° C, the cloud is said to be warm. Often, clouds develop at altitudes where temperatures are below 0° C, but the droplets do not freeze because of the purity of the water. Such clouds are said to be supercooled. In the atmosphere, supercooling to temperatures of −10° C or even −20° C is not unusual. The lower the temperature, the greater the likelihood that the droplets will intercept so-called ice nuclei, which cause them to freeze. At temperatures below about −40° C, virtually all clouds are composed of ice crystals.

Many all-liquid clouds, whether warm or supercooled, are stable in the sense that the droplets are limited in size to a few tens of micrometres, and the clouds may last for some time without yielding rain or snow.

Sometimes nature is deficient in ice nuclei, and as a result supercooled clouds may persist for many hours. When this is the case, the addition of ice nuclei can upset the cloud stability by causing ice crystals to form; these can then grow and result in precipitation.

Substances used in cloud seeding

The introduction of any substance into clouds for the purpose of changing them is called cloud seeding. Warm clouds (above 0° C) have been seeded with materials such as sodium chloride or calcium chloride particles or by a water spray. The objective of these procedures is to produce giant cloud droplets that will grow by coalescence, fall, and sweep out smaller cloud droplets. Fogs over airports have been seeded in order to reduce the density of the cloud and to improve the visibility and ceiling conditions. Warm convective clouds have been seeded in an effort to increase rainfall.

Aircraft have been used to dispense a water spray or salt particles. In some cases, salt-water sprays have been dispersed. Unfortunately, such solutions tend to be corrosive to aircraft surfaces and have to be handled carefully. In some programs sodium chloride particles in powdered form have been blown up from the ground.

Most cloud-modification activities have been concerned with supercooled clouds and have involved seeding with ice nuclei. As noted above, the first substance found to be effective as a cloud-seeding agent was Dry Ice. Its temperature is so low (about −78° C) that it causes ice crystals to form spontaneously from water vapour. It has been estimated that a gram of Dry Ice will produce at least $3 \times 10^{10}$ ice crystals. The most common procedure for seeding with Dry Ice is to fly over a cloud and disperse crushed pellets, less than a millimetre to a few millimetres in diameter, along the path of flight. A typical seeding rate might be several kilograms of Dry Ice per kilometre of flight.

Dry Ice is no longer widely employed as a cloud-seeding agent because it suffers from the disadvantage of having to be delivered to the supercooled regions of the cloud and from the fact that, once a pellet of Dry Ice has evaporated, it can no longer affect the cloud. Supercooled clouds are now most commonly seeded with tiny particles of silver iodide.

Seeding with silver iodide

There are many techniques for seeding with silver iodide. All of them produce large numbers of minute particles that range in diameter from about 0.01 to 0.1 micrometre. A common procedure is to dissolve silver iodide in a solution of sodium iodide in acetone. The concentration of silver iodide may range from 1 to 10 percent. When the solution is burned in a well-ventilated chamber at a temperature of about 1,100° C, a very large number of ice nuclei are produced. The concentration increases rapidly as the temperature decreases. A typical quantity at −10° C is $10^{13}$ ice nuclei per gram of silver iodide. Exposure to ultraviolet light causes rapid deactivation of the silver iodide nuclei. The concentrations of nuclei may decrease by perhaps a factor of 10 for each hour of exposure.

In the United States, France, Switzerland, and Argentina, a great deal of the silver iodide seeding has been done by means of ground generators. When this is the case, air currents are expected to transport the nuclei into the supercooled parts of the cloud.

Experimenters in the United States, Australia, Israel, and other countries have used airplanes for dispersing silver iodide particles. For the most part, silver iodide in acetone has been burned in generators suspended from the wings of one or more airplanes. In some tests pyrotechnical devices have been employed for this purpose. When they are dropped from above into clouds, a solid mixture of silver iodide and an inflammable substance is ignited.

In the Soviet Union there has been widespread use of rockets fired from the ground. This procedure was first employed in Italy, but the Soviets refined it by employing a superior rocket in which the pyrotechnic mixture could be ignited anywhere along the path of flight. Also in the Soviet Union a great deal of ice-nuclei seeding has been done by means of 70-millimetre artillery guns that fire projectiles containing 100 to 200 grams of lead iodide or silver iodide. At a predetermined position the projectile explodes and disperses the ice nuclei.

Other cloud-modification methods

Various other schemes for seeding clouds have been employed. One that is used at Orly Airport in France involves tanks of propane (see below). When propane gas is released, it expands and cools the air, and, as a result, ice crystals are produced. Some tests have been made of this procedure in the United States.

Most cloud-modification work has been concerned with changing the size of cloud particles or the buoyancy of the cloud air. A number of scientists have been interested in developing procedures for changing the electrical structure of clouds. One practical goal has been the reduction of the number of forest fires caused by lightning. Small clouds have been modified electrically by releasing large quantities of ions from a long piece of wire mounted near the ground.

Attempts to change the electrical nature of large thunderstorms have involved seeding with ice nuclei and, in another set of experiments, with large numbers of short metal strips. The object has been to prevent the electrical charge of clouds from becoming so concentrated that a lightning stroke would occur.

### FOG DISSIPATION

In order for aircraft to take off and land, it is necessary that the ceiling (the height of the cloud base above the ground) and visibility be above certain minimum values. It has been estimated that, in the United States alone, airport shutdowns by fog were costing the airlines many millions of dollars annually. The vital effect of low ceilings and visibilities on military aircraft operation was forcefully emphasized during World War II when Allied aircraft flew out of foggy England.

During the late 1930s attempts were made to dissipate fogs by seeding them with salt particles, in particular calcium chloride. Some success was experienced, but this technique did not appear to be practical. During the mid-1940s large quantities of heat were used to clear airport runways. The scheme called FIDO (Fog Investigation Dispersal Operations) employed kerosene burners along the runways. The heat they released decreased the relative humidity of the air and caused droplet evaporation and a sufficient improvement in ceiling and visibility to allow aircraft to land or take off.

The dissipation of supercooled fogs by means of ice nuclei has been going on for many years. Tables have been prepared that specify the quantities of Dry Ice to be dispersed, depending on such factors as wind speed, cloud thickness, and temperature. A typical seeding rate might be about two kilograms per kilometre of flight. Special equipment has been developed for the purpose of dispensing Dry Ice flakes or pellets from an airplane or from the ground.

A network of propane gas sprayers and associated tanks of liquid propane are installed around the periphery of

Orly Airport. The particular sprayers to be activated on any foggy occasion are determined automatically by a computer that is fed meteorological data on visibility, temperature, and wind velocity.

Investigations have been made of the value of acoustical techniques for clearing fogs. Such schemes work well in a cloud chamber where standing sound waves can be set up, but there is no evidence that reasonable sound sources can effectively change the characteristics of fogs in the free atmosphere.

### PRECIPITATION MODIFICATION

Shortly after Schaefer's proof that Dry Ice seeding could modify supercooled stratus clouds, there were many projects aimed at increasing rain or snow by economically important amounts. The first cloud-seeding tests demonstrated that, in the course of dissipating stratiform clouds, some small amounts of snow fell that would not have fallen had there been no seeding. Certain meteorologists hypothesized that by seeding thick clouds it should be possible to cause substantial increases of rain or snow. Unfortunately, from the point of view of designing a scientific experiment that would put this hypothesis to the test, the thicker the cloud, the better the chance of natural precipitation. Cloud thickness alone does not allow a unique specification of the quantity of rain or snow. Other factors such as the strength and persistence of the cloud updraft, cloud-top temperature, the horizontal dimensions of the cloud, and its microphysical properties influence the amount of precipitation.

Problems of analyzing results

The complex nature of clouds has so far thwarted attempts to develop quantitative forecasts of rainfall of sufficient accuracy to be used to evaluate a cloud-seeding scheme. Also, as is well known, precipitation is highly variable in space and time. As a result, it is not possible, on the basis of a physical theory, to answer satisfactorily the question, "How much rain or snow would have fallen if there had been no cloud seeding?"

The most reliable evidence concerning the effects of cloud seeding has come from programs in which statistical techniques were employed to design experiments and test hypotheses dealing with the effectiveness of any particular cloud-seeding scheme. Many experimental designs and evaluating procedures have been used since the late 1940s. There have been many disagreements among scientists and statisticians about the interpretation of programs that have been conducted in the past.

A question of fundamental importance has been raised about the scientific value of precipitation-augmentation projects conducted by private or commercial interests. Such operations usually have been based on the assumption that seeding would increase precipitation. They have not been conducted as experiments designed to test whether or not such would be the result. Certain prominent statisticians have taken the position that because these projects have not purposefully incorporated "randomized" or other control procedures to reduce the effects of bias by the operators, the data they have yielded cannot be used to test the efficacy of cloud seeding. On these grounds, in 1957 optimistic conclusions by the U.S. Advisory Committee on Weather Control on the effectiveness of cloud seeding were rejected by various statisticians. The final report of the committee concluded that precipitation from winter supercooled clouds over the mountainous western United States was increased by some 10 to 15 percent as a result of silver iodide seeding. In 1966 a special panel of the National Academy of Sciences, again employing data mostly derived from private or commercial operators, arrived at almost the same conclusion.

The most recent evidence indicates that sometimes ice-nuclei seeding may increase precipitation from certain supercooled clouds by some tens of percent. In other circumstances the seeding may lead to decreases of about the same magnitude. In still other meteorological situations seeding is ineffective. With a few exceptions it still is not possible to specify the conditions under which positive or negative effects would be expected to occur. It appears that in certain types of supercooled clouds the temperature at the upper boundary of the cloud is an important but not sole indicator of the most likely effects of ice-nuclei seeding.

Use of mathematical models of clouds and cloud systems

Since the late 1960s, increasing effort has been made to develop mathematical models of clouds and cloud systems. Once an accurate model exists, it is possible to calculate the expected results of ice-nuclei seeding by means of a computer. This approach was employed by Joanne Simpson of the Environmental Science Services Administration and others to test the effects of heavy doses of silver iodide on cumulonimbus clouds. She found that the effects of ice nuclei on large convective clouds conformed closely with theoretical predictions. Certain specified clouds were caused to grow and to produce more rain than they would have if they had not been seeded (Figure 68).

An important but still unresolved question deals with effects of cloud seeding on precipitation downwind from the target area. For the most part studies have shown excesses of precipitation, but there still exists the possibility of decreases not only far downwind but in all other directions as well.

A number of tests have been made to stimulate rainfall from warm cumulus clouds by seeding them with sodium chloride particles. Experiments in India and certain other countries were reported to have increased the amount of rainfall successfully.

### MODIFICATION OF OTHER WEATHER PHENOMENA

**Electricity in clouds.** Various schemes have been employed to modify the electrical nature of clouds and the occurrence of cloud-to-ground lightning. Research has shown that by releasing large quantities of ions near the ground, it is possible to influence the electric properties of small cumulus clouds. This does not mean, however, that large clouds could be influenced in this manner.

There is no convincing evidence supporting the assertion that the release of electrically charged particles will influence the precipitation from fogs or clouds.

An extensive program dealing with the modification of lightning storms has been conducted by the U.S. Forest Service. Potential lightning storms were seeded heavily with silver iodide nuclei. The lightning characteristics of some traveling thunderstorms apparently were changed, but it still was not demonstrated conclusively that cloud-to-ground lightning can be reduced.

By courtesy of Joanne Simpson, Experimental Meteorology Laboratory, NOAA

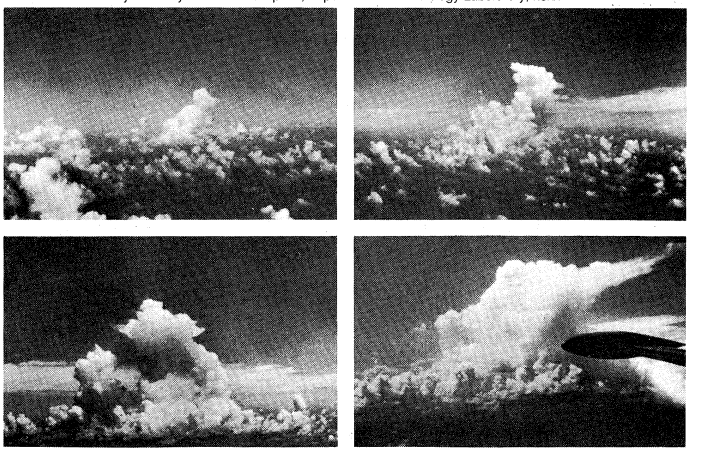


Figure 68: *Cloud seeding with silver iodide.*
(Top left) The cloud at the time of seeding. (Top right and bottom) The cloud at nine, 19, and 36 minutes after seeding.

**Hail suppression.** In many areas of the world hail does enormous destruction to agriculture, particularly fruit orchards and grain fields. There have been cloud-seeding projects aimed at reducing hail damage. Some operations have attempted to put so many nuclei into the supercooled parts of cumulonimbus that they would be almost totally converted to ice crystals. Such a procedure, called overseeding, is not considered practical because of the large quantities of material needed to seed the clouds over an area great enough to have an appreciable effect.

Most hail-suppression attempts have been based on the

Reducing the size of hailstones

concept that damage will be reduced if the hailstone sizes are reduced. This does not require overseeding. Consider, for example, an unseeded cloud that produces one hailstone having a two-centimetre diameter in each cubic metre of air. If ice-nuclei seeding could cause 100 uniform hailstones in the same volume from the same available quantity of supercooled water, their diameters would be about 0.4 centimetre. The small stones would melt as they fell through the layer of warm air below the freezing level. Even if they did not melt completely to form rain, by the time the hailstones reached the ground they would be too small to do any serious damage.

Silver iodide seeding of potential hailstorms has been carried out in many countries. Most of the ice nuclei have been dispersed from ground-based or aircraft-mounted generators. In Switzerland it appeared that there may have been more hail produced by seeding. In Argentina the results seemed to depend on the type of weather situation. In the United States varying results have been reported.

In the Soviet Union, ice nuclei have been inserted directly into the supercooled parts of clouds by means of rockets or artillery. In the latter technique a projectile explodes and disperses the nuclei. The rockets carry a cylinder of a pyrotechnic substance impregnated with silver iodide or lead iodide. It passes through the cloud while burning for a period of 45 seconds. Spectacular success in hail reduction has been reported by Soviet scientists. The benefit-to-cost ratios cited range from 4 to 1 up to 17 to 1. There have been no independent tests of the Soviet procedures, and as a result many other atmospheric scientists have been hesitant to accept the claims of success at face value.

**Severe storms.** Hurricanes can cause widespread destruction and human misery. An average hurricane has tremendous energy. In one day the energy released is about $1.6 \times 10^{13}$ kilowatt-hours, or at least 8,000 times more than the electrical power generated each day in the United States. This quantity is equivalent to a daily explosion of 500,000 atomic bombs of the 20-kiloton Nagasaki variety. These numbers should make it clear that it would be impractical to attempt to modify hurricanes by a brute force approach. It is necessary to find a means whereby a small input of energy may upset a natural instability and lead to large results. Ice-nuclei seeding is the one such approach now under investigation.

Hurricane-seeding experiment

The first hurricane-seeding test was carried out in 1947 by Irving Langmuir and his colleagues, who distributed about 91 kilograms of crushed Dry Ice in a storm. They apparently were convinced that the seeding caused a change in the track followed by the storm.

On Aug. 18 and 20, 1969, Hurricane Debbie was seeded as a part of Project Stormfury, a series of hurricane-modification experiments conducted by the Environmental Science Services Administration and the U.S. Navy. Heavy doses of silver iodide were dropped into the hurricane clouds from airplanes. The maximum measured wind speeds in the hurricane decreased by 31 and 15 percent on the two seeded days. On August 19, the day between the two flights, the storm reportedly reintensified.

The results of this experiment were in the direction predicted by a mathematical model of a hypothetical hurricane. Because the measured winds in two hurricanes seeded in earlier years also decreased, Project Stormfury scientists were optimistic that hurricanes could be modified beneficially, but many more experiments are needed before attempting the seeding of hurricanes close to land.

The violent nature of tornadoes would appear to dictate substantial programs of research to increase our understanding and control of these storms. In fact, very little scientific attention has been devoted to attempts to modify tornadoes. It has been speculated that they might be influenced by firing rockets into them and distributing materials to modify their temperature structure or electrical properties. Unfortunately, so little is known about the tornadoes that few scientists have confidence that such schemes would be effective.

## CHANGES IN THE RADIATION BALANCE NEAR THE GROUND

When air moves against a mountain slope it is forced to rise, and in so doing clouds and precipitation often are produced. Air moving over heated islands in the tropical oceans acts in a similar fashion even when the terrain is flat. The ground and the air over it warm up more than does the adjacent water. A rising convection current develops over the island. The result is that air moving over the island rises in a manner similar to that over a mountain. This is known as the "thermal mountain effect."

Scientists of the Esso Research and Engineering Company proposed that by covering large areas with asphalt it might be possible, over land, to simulate the results observed over tropical islands. The black asphalt would become hotter than surrounding lighter soil. It has been speculated that this would lead to convective currents, a thermal mountain, and more precipitation.

It has been proposed that cloud seeding be used to open large holes (about 100 kilometres in diameter) in supercooled layer clouds in order to change the radiation balance of a region. Clouds reflect perhaps 70 percent of the Sun's rays, while the ground reflects about 20 percent. If a deck of supercooled cloud could be dissipated during the daytime, there would be more solar energy absorbed, local warming, and possibly changes in atmospheric motions and behaviour. If large areas of clouds were dissipated at night, there would be excessive outgoing radiation and cooling. The effects of the resultant temperature drop over a large area still have not been evaluated thoroughly.

(L.J.B./Ed.)

### BIBLIOGRAPHY

*General works:* Introductory works include A.S. MONIN, *An Introduction to the Theory of Climate* (1986; originally published in Russian, 1982); PAUL E. LYDOLPH, *Weather and Climate* (1985); JOHN T. HOUGHTON (ed.), *The Global Climate* (1984); and LOUIS J. BATTAN, *Weather in Your Life* (1983). Two excellent comprehensive reference works are JOHN E. OLIVER and RHODES W. FAIRBRIDGE (eds.), *The Encyclopedia of Climatology* (1987); and DAVID D. HOUGHTON (ed.), *Handbook of Applied Meteorology* (1985). JOHN E. HOBBS, *Applied Climatology: A Study of Atmospheric Resources* (1980); and JOHN F. GRIFFITHS, *Applied Climatology: An Introduction,* 2nd ed. (1976), are also useful. Definitions of meteorological terms are provided in RALPH E. HUSCHKE (ed.), *Glossary of Meteorology* (1959, reprinted 1970); and WORLD METEOROLOGICAL ORGANIZATION, *International Meteorological Vocabulary* (1966), including nomenclature in English, French, Russian, and Spanish.

Current research is reported in the following journals: *Bulletin of the American Meteorological Society* (monthly); *Climate Monitor* (quarterly); *Climatic Change* (6/yr.); *International Journal of Biometeorology* (quarterly); *Journal of Climate and Applied Meteorology* (monthly); *Journal of Climatology* (bimonthly); *Journal of Meteorological Research* (bi-monthly); *Journal of Meteorology* (10/yr.); *Monthly Weather Review; Quarterly Journal of the Royal Meteorological Society; Soviet Meteorology and Hydrology* (monthly); *Weather* (monthly); *Weatherwise* (bimonthly); and *W.M.O. Bulletin* (quarterly). (Ed.)

*Solar radiation and temperature:* Introductory discussions of these basic elements of climate can be found in GLENN T. TREWARTHA and LYLE H. HORN, *An Introduction to Climate,* 5th ed. (1980); JOHN F. GRIFFITHS and DENNIS M. DRISCOLL, *Survey of Climatology* (1982); and STANLEY DAVID GEDZELMAN, *The Science and Wonders of the Atmosphere* (1980). G.W. PALTRIDGE and C.M.R. PLATT, *Radiative Processes in Meteorology and Climatology* (1976), provides more advanced treatment. (Ro.D.)

*Atmospheric humidity and precipitation:* Discussions of water vapour in the atmosphere and global water budgets are found in NEIL WELLS, *The Atmosphere and Ocean: A Physical Introduction* (1986); and F.H. LUDLAM, *Clouds and Storms: The Behavior and Effect of Water in the Atmosphere* (1980). Forms of precipitation are surveyed in B.J. MASON, *Clouds, Rain and Rainmaking,* 2nd ed. (1975); W.E. KNOWLES MIDDLETON, *A History of the Theories of Rain and Other Forms of Precipitation* (1965); and D.M. GRAY and D.H. MALE (eds.), *Handbook of Snow: Principles, Processes, Management & Use* (1981). B.J. MASON, *The Physics of Clouds,* 2nd ed. (1971), is an authoritative text. See also the cloud atlas by RICHARD SCORER, *Clouds of the World: A Complete Color Encyclopedia* (1972). (Ed.)

*Atmospheric pressure and wind:* General textbooks with effective discussions of wind and pressure are FREDERICK K. LUTGENS and EDWARD J. TARBUCK, *The Atmosphere: An Introduction to Meteorology,* 3rd ed. (1986); C. DONALD AHRENS, *Meteorology Today: An Introduction to Weather, Climate, and the Environment,* 2nd ed. (1985); LOUIS J. BATTAN, *Fundamentals of Meteorology,* 2nd ed. (1984); and STANLEY DAVID

GEDZELMAN, *The Science and Wonders of the Atmosphere* (1980). More sophisticated treatment of the wind/pressure relationship is provided by JOHN A. DUTTON, *The Ceaseless Wind: An Introduction to the Theory of Atmospheric Motion,* enl. ed. (1986); JAMES R. HOLTON, *An Introduction to Dynamic Meteorology,* 2nd ed. (1979); JOHN M. WALLACE and PETER V. HOBBS, *Atmospheric Science: An Introductory Survey* (1977); HORACE R. BYERS, *General Meteorology,* 4th ed. (1974); and E. PALMÉN and C.W. NEWTON, *Atmospheric Circulation Systems: Their Structure and Physical Interpretation* (1969). (P.J.S.)

*Major forms of weather disturbances:* JOE R. EAGLEMAN, *Severe and Unusual Weather* (1983), presents an overview of storms. Types of storms, their attendant phenomena, and their effects are treated in EDWIN KESSLER (ed.), *Thunderstorms—A Social, Scientific, & Technological Documentary,* 3 vol. (1981–82), including studies of tornadoes; R.H. GOLDE (ed.), *Lightning,* 2 vol. (1977); MARTIN A. UMAN, *Lightning* (1969, reissued 1984), a fundamental work; NARAYAN R. GOKHALE, *Hailstorms and Hailstone Growth* (1975); RICHARD A. ANTHES, *Tropical Cyclones: Their Evolution, Structure and Effects* (1982); ROBERT H. SIMPSON and HERBERT RIEHL, *The Hurricane and Its Impact* (1981); and SNOWDEN D. FLORA, *Tornadoes of the United States,* 2nd ed. (1954), a standard work. (L.J.B./N.R.W./Ed.)

*Climatic variations and changes:* Overviews are given by M.I. BUDYKO, *The Earth's Climate, Past and Future* (1982; originally published in Russian, 1980), a masterly account by the founder of contemporary climatology; H.H. LAMB, *Climate, History, and the Modern World* (1982); T.M.L. WIGLEY, M.J. INGRAM, and G. FARMER, *Climate and History: Studies in Past Climates and Their Impact on Man* (1981); and MICHAEL R. RAMPINO *et al., Climate, History, Periodicity, and Predictability* (1987), on the relationship between climate cycles and their causes, with an extensive bibliography. A.B. PITTOCK *et al.* (eds.), *Climatic Change and Variability: A Southern Perspective* (1978), offers an excellent antipodean analysis. EMMANUEL LE ROY LADURIE, *Times of Feast, Times of Famine: A History of Climate Since the Year 1000* (1971, reissued 1988; originally published in French, 1967), is an interdisciplinary study using many data sources to document climatic variations. R.S. BRADLEY, *Quaternary Paleoclimatology: Methods of Paleoclimatic Reconstruction* (1985), summarizes research techniques. (F.K.H.)

*Climatic classification:* Useful introductory discussions can be found in ARTHUR N. STRAHLER and ALAN H. STRAHLER, *Modern Physical Geography,* 3rd ed. (1987); and HERMANN FLOHN, *Climate and Weather* (1969; originally published in German, 1968). More advanced treatment is provided by A. HENDERSON-SELLERS and P.J. ROBINSON, *Contemporary Climatology* (1986); JOHN G. LOCKWOOD, *World Climatic Systems* (1985); ROGER G. BARRY and RICHARD J. CHORLEY, *Atmosphere, Weather, and Climate,* 4th ed. (1982); B.W. ATKINSON, *Dynamical Meteorology: An Introductory Selection* (1981); and R.G. BARRY and A.H. PERRY, *Synoptic Climatology: Methods and Applications* (1973). Specific treatment of the topic is given by JOHN E. OLIVER and L. WILSON, "Climatic Classification," in JOHN E. OLIVER and RHODES W. FAIRBRIDGE (eds.), *The Encyclopedia of Climatology* (1987), pp. 221–237; and JOHN E. OLIVER, *Climate and Man's Environment: An Introduction to Applied Climatology* (1973). The global distribution of major climate types is the subject of GLENN T. TREWARTHA and LYLE H. HORN, *An Introduction to Climate,* 5th ed. (1980); JOHN F. GRIFFITHS and DENNIS M. DRISCOLL, *Survey of Climatology* (1982); HOWARD J. CRITCHFIELD, *General Climatology,* 4th ed. (1983); JOHN G. LOCKWOOD, *World Climatology: An Environmental Approach* (1974); S. NIEUWOLT, *Tropical Climatology: An Introduction to the Climates of the Low Latitudes* (1977); and HERBERT RIEHL, *Climate and Weather in the Tropics* (1979). Particular regions are examined in H.E. LANDSBERG (ed.), *World Survey of Climatology* (1969– )—15 vol. have appeared to 1987; and GLENN T. TREWARTHA, *The Earth's Problem Climates,* 2nd ed. (1981). Climatic data are covered in HOWARD J. CRITCHFIELD, "Climatic Data, Sources of," in OLIVER and FAIRBRIDGE (*op. cit.*), pp. 272–276. Discussion of meso- and microclimates are found in T.R. OKE, *Boundary Layer Climates* (1978); RUDOLF GEIGER, *The Climate near the Ground* (1965; originally published in German, 1961); MASATOSHI M. YOSHINO, *Climate in a Small Area: An Introduction to Local Meteorology* (1975); and HELMUT E. LANDSBERG, *The Urban Climate* (1981). (A.J.A.)

*Climate and life:* Comprehensive introductions include STEPHEN H. SCHNEIDER and RANDI LONDER, *The Coevolution of Climate and Life* (1984); and JAMES LOVELOCK, *The Ages of Gaia: A Biography of Our Living Earth* (1988), a restatement of a controversial view of interactions between the living and inorganic parts of Earth. Causes of climatic change are explained in detail in B. BOLIN and R.B. COOK (ed.), *The Major Biogeochemical Cycles and Their Interactions* (1983); B. BOLIN *et al., The Greenhouse Effect, Climatic Change, and Ecosystems* (1986); NORMAN MYERS, *Conversion of Tropical Moist Forests* (1980), a clearly argued but controversial account; and ROBERT E. DICKINSON (ed.), *The Geophysiology of Amazonia: Vegetation and Climate Interactions* (1987). The impact of climate on human life is treated in MICHAEL GLANTZ, RICHARD KATZ, and MARIA KRENZ (eds.), *The Societal Impacts Associated with the 1982–83 Worldwide Climate Anomalies* (1987), a report on the effects of the 1982–83 El Niño/Southern Oscillation, published by the National Center for Atmospheric Research; WILFRID BACH, JÜRGEN PANKRATH, and STEPHEN H. SCHNEIDER (eds.), *Food–Climate Interactions* (1981); ROBERT W. KATES, JESSE H. AUSUBEL, and MIMI BERBERIAN, *Climate Impact Assessment: Studies of the Interaction of Climate and Society* (1985); and WILLIAM W. KELLOGG and ROBERT SCHWARE, *Climate Change and Society: Consequences of Increasing Atmospheric Carbon Dioxide* (1981). The impact of human activities on the climate is presented in WILLIAM C. CLARK and R.E. MUNN (eds.), *Sustainable Development of the Biosphere* (1986); MICHAEL C. MACCRACKEN and FREDERICK M. LUTHER (eds.), *Projecting the Climatic Effects of Increasing Carbon Dioxide* (1985); NATIONAL RESEARCH COUNCIL (U.S.), CARBON DIOXIDE ASSESSMENT COMMITTEE, *Changing Climate* (1983); and P.S. LISS and A.J. CRANE, *Man-Made Carbon Dioxide and Climatic Change: A Review of Scientific Problems* (1983). Climate model predictions are explored in STEPHEN H. SCHNEIDER, "Climate Modeling," *Scientific American,* 256(5):72–80 (May 1987); and A. BERGER *et al., Milankovitch and Climate: Understanding the Response to Astronomical Forcing,* 2 vol. (1984). (S.H.S./A.B.Ri.)

*Meteorological measurement and weather forecasting:* RICHARD A. ANTHES *et al., The Atmosphere,* 3rd ed. (1981), contains general discussions. The history of weather forecasting is recounted in GISELA KUTZBACH, *The Thermal Theory of Cyclones: A History of Meteorological Thought in the Nineteenth Century* (1979); A. KH. KHRGIAN, *Meteorology: A Historical Survey,* 2nd ed. rev. (1970; originally published in Russian, 2nd ed. rev., 1959); PATRICK HUGHES, "American Weather Services," *Weatherwise,* 33(3):100–111 (June 1980); and FREDERICK G. SHUMAN, "Numerical Weather Prediction," *Bulletin of the American Meteorological Society,* 59(1):5–17 (January 1978). Instruments used in weather analysis are examined in the classic texts by W.E. KNOWLES MIDDLETON and ATHELSTAN F. SPILHAUS, *Meteorological Instruments,* 3rd ed. rev. (1953, reprinted 1960); and W.E. KNOWLES MIDDLETON, *History of the Barometer* (1964), *A History of the Thermometer and Its Uses in Meteorology* (1966), and *Invention of the Meteorological Instruments* (1969); as well as in LEO J. FRITSCHEN and LLOYD W. GAY, *Environmental Instrumentation* (1979), with coverage limited to measurements at the Earth's surface; STUART G. BIGLER "Radar: A Short History," *Weatherwise,* 34(4):158–163 (August 1981); LOUIS J. BATTAN, *Radar Observation of the Atmosphere,* rev. ed. (1973); R.S. SCORER, *Cloud Investigation by Satellite* (1986); VINCENT J. OLIVER, "A Primer: Using Satellites to Study the Weather," *Weatherwise,* 34(4):164–170 (August 1981); ERIC C. BARRETT and DAVID W. MARTIN, *The Use of Satellite Data in Rainfall Monitoring* (1981); and ERIC C. BARRETT, *Climatology from Satellites* (1974). A good overview of forecasting is found in LANCE F. BOSART, "Weather Forecasting," ch. 4 in DAVID D. HOUGHTON (ed.), *Handbook of Applied Meteorology* (1985), pp. 205–279. Methods and problems of forecasting are presented in JAROMIR NEMEC, *Hydrological Forecasting: Design and Operation of Hydrological Forecasting Systems* (1986); D.M. BURRIDGE and E. KÄLLÉN (eds.), *Problems and Prospects in Long and Medium Range Weather Forecasting* (1984); K.A. BROWNING (ed.), *Nowcasting* (1982); and three articles from *Bulletin of the American Meteorological Society:* DONALD L. GILMAN, "Long-Range Forecasting: The Present and the Future," 66(2):159–164 (February 1985); JEROME NAMIAS, "Remarks on the Potential for Long-Range Forecasting," 66(2):165–173 (February 1985); and L. BENGTSSON, "Medium-Range Forecasting—The Experience at ECMWF," 66(9):1133–46 (September 1985). (J.J.Ca.)

*Scientific weather modification:* NATIONAL RESEARCH COUNCIL (U.S.) COMMITTEE ON ATMOSPHERIC SCIENCES, *Weather and Climate Modification: Problems and Prospects,* 2 vol. (1966), an authoritative survey, supplemented by *Weather and Climate Modification: Problems and Progress* (1973); GEORG BREUER, *Weather Modification: Prospects and Problems* (1979; originally published in German, 1976), a nontechnical examination of the subject; W.R.D. SEWELL *et al., Modifying the Weather: A Social Assessment* (1973), a collection of symposium papers; WILLIAM A. THOMAS (ed.), *Legal and Scientific Uncertainties of Weather Modification* (1977); ROBERT G. FLEAGLE *et al., Weather Modification in the Public Interest* (1974), a review of weather modification research and activities and related public policy issues; G. BRANT FOOTE and CHARLES A. KNIGHT (eds.), *Hail: A Review of Hail Science and Hail Suppression* (1977); ARNETT S. DENNIS, *Weather Modification by Cloud Seeding* (1980); and W.N. HESS (ed.), *Weather and Climate Modification* (1974), a comprehensive account of all aspects of weather modification. Appropriate articles can be found in *The Journal of Weather Modification* (annual). (L.J.B./Ed.)

# Cnidarians

The phylum Cnidaria (cnidarians) comprises a diverse group of aquatic animals. The phylum is made up of four classes: Hydrozoa (hydrozoans); Scyphozoa (scyphozoans); Anthozoa (anthozoans); and Cubozoa (cubozoans). The nearly 9,000 species share several attributes, supporting the theory that they had a single origin. Variety and symmetry of body forms, varied coloration, and the sometimes complex life histories of cnidarians fascinate layperson and scientist alike. Inhabiting all marine and some freshwater environments, these animals are most abundant and diverse in tropical waters. Their calcareous skeletons form the frameworks of the reefs and atolls in most tropical seas, including the Great Barrier Reef that extends more than 2,000 kilometres along the northeastern coast of Australia.

Only cnidarians manufacture microscopic intracellular stinging capsules, known as nematocysts or cnidae, which give the phylum its name. An alternative name for the phylum is Coelenterata, which refers to their simple organization around a central body cavity (the coelenteron). As first defined, Coelenterata included not only the animals now designated cnidarians but also sponges (phylum Porifera) and comb jellies (phylum Ctenophora). In contemporary usage, "coelenterate" generally refers only to cnidarians, but the latter term is used in order to avoid ambiguity.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* especially section 313.

This article is divided into the following sections:


General features 523
  Size range and diversity of structure 523
  Distribution and abundance 523
  Importance 524
Natural history 524
  Reproduction and life cycles 524
  Ecology and habitats 525
  Locomotion 525
  Food and feeding 525
  Associations 525
Form and function 526
  Tissues and muscles 526
  Support mechanisms and skeletons 526
  Nervous system and organs of sensation 526
  Digestion, respiration, and excretion 527
  Defense and aggression: nematocysts 527
Evolution 527
Classification 527
  Annotated classification 527
  Critical appraisal 528
Bibliography 528


## GENERAL FEATURES

**Size range and diversity of structure.** Cnidarians are considered primitive because they are radially symmetrical (*i.e.,* similar parts are arranged symmetrically around a central axis), they lack cephalization (concentration of sensory organs in a head), their bodies have two cell layers rather than the three of so-called higher animals, and the saclike coelenteron has one opening (the mouth). They are the most primitive of animals whose cells are organized into distinct tissues, but they lack organs. Cnidarians have two body forms—polyp and medusa (Figures 1 and 2).

Medusae and polyps

The body of a medusa, commonly called a jellyfish, usually has the shape of a bell or an umbrella, with tentacles hanging downward at the margin. The tubelike manubrium hangs from the centre of the bell, connecting the mouth at the lower end of the manubrium to the coelenteron within the bell. Most medusae are slow-swimming, planktonic animals. In a polyp, the mouth and surrounding tentacles face upward, and the cylindrical body is generally attached by its opposite end to a firm substratum. The mouth is at the end of a manubrium in many hydrozoan polyps. Anthozoan polyps have an internal pharynx, or stomodaeum, connecting the mouth to the coelenteron.

Most species of cubozoans, hydrozoans, and scyphozoans alternate in their life cycles between medusoid and polypoid body forms, with medusae giving rise sexually to larvae that metamorphose into polyps, while polyps produce medusae asexually. Anthozoans are polypoid cnidarians and do not have a medusa stage. Commonly polyps, and in some species medusae too, can produce more of their own kind asexually.

One body form may be more conspicuous than the other. For example, scyphozoans are commonly known as true jellyfishes, for the medusa form is larger and better known than the polyp form. In hydrozoans, the polyp phase is more conspicuous than the medusa phase in groups such as hydroids and hydrocorals. Hydromedusae are smaller and more delicate than scyphomedusae or cubomedusae, and may be completely absent from the life cycle of some hydrozoan species. Cubozoans have medusae commonly known as box jellyfish, from their shape. Some of these are responsible for human fatalities, mostly in tropical Australia and Southeast Asia, and include the so-called sea wasps. The polyp is tiny and inconspicuous.

Many cnidarian polyps are individually no more than a millimetre or so across. Polyps of most hydroids, hydrocorals, and soft and hard corals, however, proliferate asexually into colonies, which can attain much greater size and longevity than their component polyps. Certain tropical sea anemones may be a metre in diameter, and some temperate ones are nearly that tall. Anthozoans are long-lived, both individually and as colonies; the life span of both large and small sea anemones has been estimated on the order of centuries. All medusae and sea anemones occur only as solitary individuals. Scyphomedusae can weigh more than a ton, whereas hydromedusae are, at most, a few centimetres across. Tentacles of medusae, however, may be numerous and extensible, which allows the animals to influence a considerably greater range than their body size might suggest. Large populations of hydroids can build up on docks, boats, and rocks. Similarly, some medusae attain remarkable densities—up to thousands per litre of water—but only for relatively brief periods.

Individuals and colonies

**Distribution and abundance.** Many of the world's benthic (bottom-dwelling) ecosystems are dominated by anthozoans. Although soft and hard corals coexist in virtually all tropical areas appropriate for either, coral reefs of the tropical Indo-Pacific are built mainly by members of the anthozoan order Scleractinia (hard corals); whereas on coral reefs of the Caribbean members of the anthozoan subclass Alcyonaria (soft corals) are much more prominent. Aside from being the most numerous and covering the greatest area of any animals on the reef, the corals structure their environment, even after death. Soft corals contribute greatly to reef construction by the cementing action of the skeletal debris (spicules), filling in spaces between hard coral skeletons.

Coral reefs

Soft-bodied anthozoans are similarly dominant in other seas. Temperate rocky intertidal zones in many parts of the world are carpeted with sea anemones. They sequester the space that is therefore made unavailable to other or-

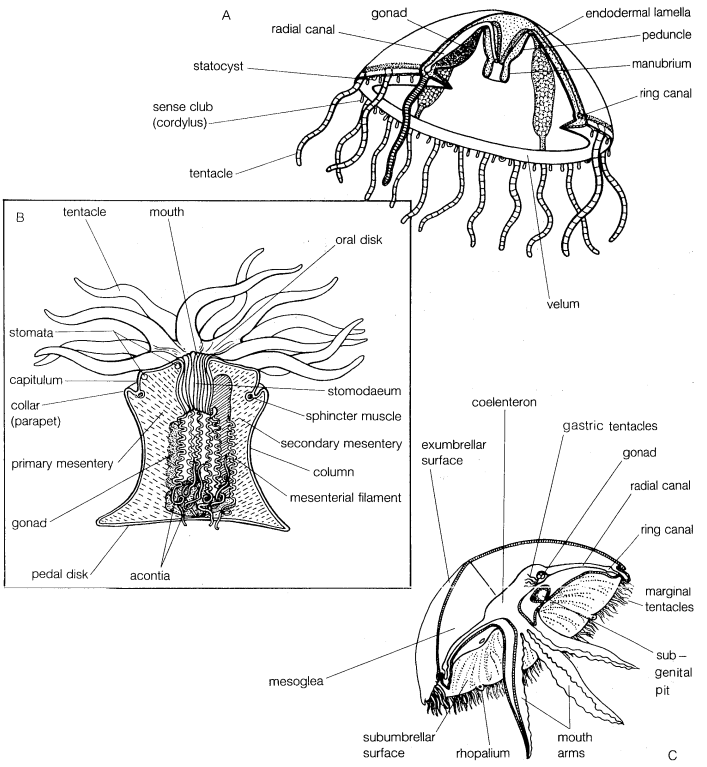



Figure 1: *Structure of cnidarians.*
(A) Hydromedusa with a section of the bell removed.
(B) Sea anemone in cross-section. (C) Scyphomedusa in cross-section.

From R.I. Smith and J.R. Carlton (eds.), *Light's Manual: Intertidal Invertebrates of the Central California Coast* (1975); University of California Press

ganisms, thus having a profound impact on community structure. The curious hemispherical anemone *Liponema* is the most abundant benthic invertebrate in the Gulf of Alaska, in terms of numbers and biomass. Parts of the Antarctic seabed are covered by anemones, and they occur near the deep-sea hot vents.

**Importance.** Prominent among organisms that foul water-borne vessels are sedentary cnidarians, especially hydroids. The muscles that make scyphomedusae strong swimmers are dried for human consumption in Asia. Sea anemones are eaten in some areas of Asia and North America.

Uses of corals

Throughout the tropics where reefs are accessible, coral skeletons are used as building material, either in blocks or slaked to create cement. Another use for cnidarian skeletons is in jewelry. The pink colour known as "coral" is the hue of the skeleton of a species of hydrocoral. Other hydrocorals have purplish skeletons. Skeletons vary in hue, and those considered most desirable command a high price. The core of some sea fans, sea whips, and black corals are cut or bent into beads, bracelets, and cameos.

All cnidarians have the potential to affect human physiology owing to the toxicity of their nematocysts. Most are not harmful to humans, but some can impart a painful sting—such as *Physalia,* the Portuguese man-of-war, and sea anemones of the genus *Actinodendron.* These, and even normally innocuous species, can be deadly in a massive dose or to a sensitive person, but the only cnidarians commonly fatal to humans are the cubomedusae, or box jellyfish. Anaphylaxis (hypersensitivity due to prior exposure and subsequent sensitization) was discovered with experiments on *Physalia* toxin. Extracts of many cnidarians, mostly anthozoans, have heart-stimulant, antitumour, and anti-inflammatory properties.

## NATURAL HISTORY

**Reproduction and life cycles.** All species of cnidarians are capable of sexual reproduction, which occurs in only one phase of the life cycle, usually the medusa. Many cnidarians also reproduce asexually, which may occur in both phases. In asexual reproduction, new individuals arise from bits of tissue that are budded off from a parent, or by a parent dividing lengthwise or crosswise into two smaller individuals. Polyps that remain physically attached to one another or embedded in a common mass of tissue constitute a colony. In some colonies, polyps share a common coelenteron through which food captured by any member is distributed to others. Hydrozoan polyp colonies, called hydroids, are prostrate, bushy, or feathery in form. Examples of other colonies are anthozoan soft corals and most reef-forming hard corals. Polyps that are produced asexually and then physically separate are called clones.

Budding

Although genetically identical, colony members of many hydrozoans and some anthozoans are polymorphic, differing in morphology (form and structure) and/or physiology. Each zooid within the colony has a specific function and varies somewhat in form. For example, gastrozooids bear tentacles and are specialized for feeding. Some colonies possess dactylozooids, tentacleless polyps heavily armed with nematocysts that seem primarily concerned with defense. Gonozooids develop reproductive structures called gonophores. Members of the order Siphonophora, free-floating colonial hydrozoans, display an even greater variety of polymorphs. These include gas-filled floats called pneumatophores, pulsating, locomotory structures called nectophores, and flattened, protective individuals called bracts or phyllozooids.

Although the medusa stage is absent in anthozoans, polyps produce additional polyps sexually and, in some species, asexually as well. Hydromedusae are budded from polyps that, in some colonial species, are specialized for this function. The major distinguishing feature of the cubozoans is that each polyp transforms entirely into a medusa. In most scyphozoans, a scyphistoma (scyphopolyp) produces immature medusae (ephyrae) by asexual fission at its oral end. This process, called strobilation, results in eight-armed, free-swimming ephyrae.

Differentiation of gametes

Gametes differentiate in parts of the body referred to as gonads, despite the fact that cnidarians cannot be said to have true ovaries and testes because they lack organs. In anthozoans, cubozoans, and scyphozoans, gametes develop in the endoderm, whereas in hydrozoans they ripen in the ectoderm, although they do not necessarily originate there. Sexes are commonly separate, but hermaphroditism

From R.D. Campbell, "Cnidaria" in A.C. Giese and J.S. Pearse (eds.), *Reproduction of Marine Invertebrates vol. 1: Acoelomate and Pseudocoelomate Metazoans* (1974); Academic Press

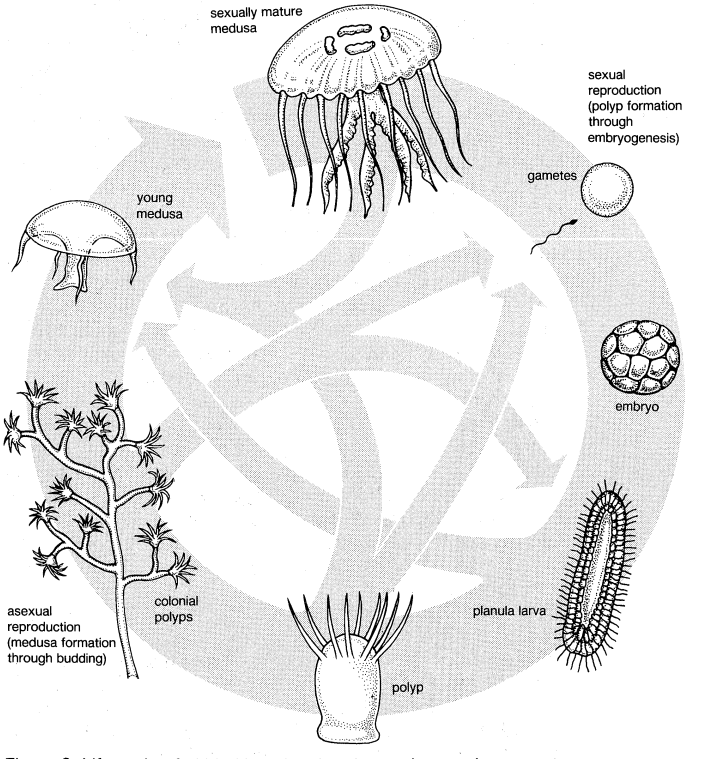



Figure 2: Life cycle of cnidarians showing the medusa and polyp stages. Not all cnidarians follow the life cycle exactly, and variation for some species is indicated by broken lines.

is known. Some hermaphroditic species are capable of self-fertilization. Gametes are generally shed into the sea, where the eggs are fertilized. Cleavage produces a ciliated ball of cells that elongates and develops a tuft of cilia at one end to become a planula larva, which may be free-swimming and planktonic, or crawling and benthic. Its ciliated tuft, which may have sensory abilities, is directed forward in locomotion. After a motile period, the planula attaches by its forward end to a solid object and develops tentacles around its posterior end, thereby transforming into a polyp. In some anthozoans and a few scyphomedusae, eggs are fertilized without being released. Embryonic development passes either partly or entirely within the mother's coelenteron or, as in the case of some anemones and some members of the anthozoan subclass Alcyonaria (octocorals), attached to the outside of her body. In some species of hydroids that lack a free medusa stage, eggs are fertilized and the embryo develops in specialized zooids that are essentially attached medusae. Such brooding species may release offspring as very advanced larvae or as miniature adults, so that a planktonic stage is absent from the life cycle.

**Ecology and habitats.** Most polyps require solid substrata for attachment, although a few burrow into soft sediments, extending only their tentacular crowns above the surface (Figure 3). Polyps are abundant in shallow waters, but sea anemones can also occur in the deepest parts of the oceans. Medusae maintain a favoured depth in the water and are carried about by currents. Most hydromedusae and scyphomedusae live in surface waters, generally in bays and along coasts, but certain species are abundant in the open ocean.

The nematocysts of cnidarians restrict potential predators to a limited array of specialists. Nematocysts also allow cnidarians to prey on a variety of would-be competitors for space and food. Most species that are capable of monopolizing space reproduce predominantly asexually. A coral, for example, can cover an area rapidly and commonly has the ability to overgrow other organisms, including corals of other species. Clone-forming sea anemones of several species actively compete for space by killing others, primarily those of their own species. When members of one clone encounter those of another, the two combatants inflate and slap one another with nematocyst-studded fighting structures (acrorhagi) located below the tentacles. Attacks may result in the death of one of the anemones, or both may retreat. Tentacle touching is involved in the recognition of non-clonemates, which presumably is chemically mediated.

Competition for space

Cnidarians are not immune from predation. Hydroids are victimized by nudibranchs that bite through the chitinous skeleton or crawl into its openings. The crown of thorns starfish, *Acanthaster planci,* extrudes its stomach over a coral colony, releases digestive enzymes, and then absorbs the liquified tissue. Butterfly and parrot fishes eat corals, being insensitive to the effects of nematocysts, which is also true of marine turtles that feed on pelagic scyphomedusae.

**Locomotion.** Medusae swim by jet propulsion (see below *Tissues and muscles*); however, most do so weakly and are carried passively by currents over long distances. Polyps are generally sedentary. Pennatulacean colonies move slowly across soft substrata by action of their inflatable peduncle (a stalk that attaches to the strata in the lower end and to the polyp body on the higher end). Sea anemones that are attached to firm substrata can creep slowly on their pedal disks or detach altogether, often in response to unfavourable physical conditions or to attack by predators. When provoked by certain starfish and nudibranchs, individuals of a few anemone species swim by paddling their tentacles or flexing their columns.

**Food and feeding.** All cnidarians are carnivores. Most use their cnidae and associated toxin to capture food, although none is known actually to pursue prey. Sessile polyps depend for food on organisms that come into contact with their tentacles. Some, such as colonial corals with minute polyps, feed on particulate material gathered in mucus impelled to the mouth by cilia (microscopic hairlike projections of cells capable of beating or waving). A hydromedusa alternately swims upward and sinks: on the upward course, its trailing tentacles are not apt to encounter food organisms, but in sinking, the extended tentacles "fish" through the water, capturing food. Once a food item is captured, tentacles move it to the mouth, either by bending in that direction or by passing it to tentacles nearer the mouth. The mouth opens, the lips grasp the food, and muscular actions complete swallowing.

Carnivorous habits

The edges of the mouths of some scyphomedusae are elaborated into mouth arms that trail behind the slowly swimming jellyfish, presenting huge surfaces for food gathering. The mouth of a scyphomedusa of the order Rhizostomae is subdivided into thousands of minute pores that lead by tubes to the coelenteron. Each pore is associated with an external ciliated gutter that collects minute organisms and detrital material as the medusa rests mouth-upward on the sea bottom.

Pink, orange, red, and brown cnidarians are commonly pigmented by carotenoids derived from crustaceans in their diet. Endodermal cells of some corals, medusae, hydras, and sea anemones contain single-celled golden-brown algae (dinoflagellates), called zooxanthellae, or green algae, called zoochlorellae. The carnivorous cnidarians cannot digest these algae but do derive a variety of nutrients from them, including glucose and oxygen. Carbon dioxide produced in respiration may be used by the algae in photosynthesis.

Zooxanthellae and zoochlorellae

**Associations.** Cnidarians enter into complex associations with a variety of other organisms, including unicellular algae, fishes, and crustaceans. Many of these relationships, such as those with zooxanthellae and zoochlorellae, are mutualistic symbioses—*i.e.,* relationships benefiting both partners. Reef-forming corals, which possess zoox-

From T.I. Storer and R.L. Usinger, *General Zoology* (copyright 1957); used with permission of McGraw-Hill Book Co.

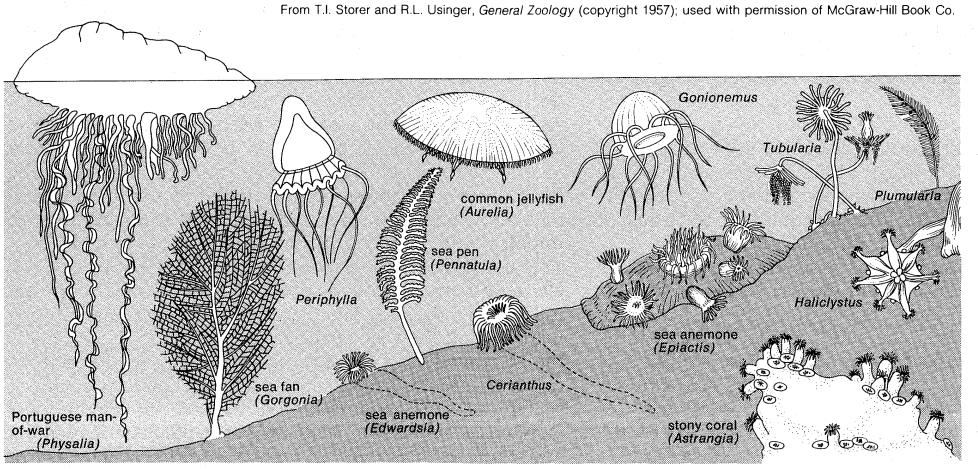



Figure 3: *Marine cnidarians in characteristic habitats.*
Hydrozoans: *Physalia, Gonionemus, Tubularia, Plumularia.* Anthozoans: *Gorgonia, Edwardsia, Pennatula, Cerianthus, Epiactis, Astrangia.* Scyphozoans: *Periphylla, Aurelia, Haliclystus.*

anthellae, form more substantial skeletons than do non-reef-forming corals, which lack zooxanthellae, for reasons that are not understood but are related to the algae. Many corals are so dependent on zooxanthellae that they cannot live in prolonged darkness, which is why coral reefs develop only in shallow, well-illuminated waters.

There are species of sea anemones that live on gastropod shells inhabited by hermit crabs, a type of crustacean that must change shells as it grows. Some hermit crabs move the anemones with them from the old shells to the new. In other cases, the anemones take the initiative, somersaulting from a now-empty shell onto the newly obtained one. A few deep-sea anemones form the shells in which their crabs dwell. This adaptation eliminates the need to change shells, but the death of one partner probably results in the death of the other. Certain true crabs carry anemones on their backs and legs, or even in their claws. These associations benefit the anemone by providing it with transport, and sometimes it can steal food from its crustacean partner. In turn, the sea anemone protects its host from predators such as octopuses and other crabs.

Gastropods (of the phylum Mollusca) also associate with cnidarians. Among the most remarkable are the nudibranchs that eat anemones and hydroids and then sequester certain types of immature, undischarged cnidae from the prey. Once the cnidae have matured within the nudibranch, they can be used in its own defense.

Symbiosis with fishes

One of the best known cnidarian symbioses is the mutualism between 10 species of tropical anemones and 26 species of anemone fish (such as the clown fish). These fishes live within the protective field of anemone tentacles, where they take refuge when a predator threatens. Immunity of the fishes to the stings of the nematocytes results from the thin layer of mucus that covers their bodies. It is unclear whether the mucous is made by the fishes themselves, or acquired by contact with the anemone's tentacles. Without its mucus, the clown fish, like any other small fish, may be stung to death and eaten by the anemone. Anemone fishes serve their hosts by driving away fishes that prey on anemones. Other fishes have a similar association with large medusae.

## FORM AND FUNCTION

**Tissues and muscles.** Cnidarians consist of two cell layers: an outer ectoderm and an inner endoderm that lines the coelenteron. Between these is sandwiched the mesoglea, a largely noncellular layer composed of a jellylike material permeated by a complex network of supporting fibres that may be microscopically thin or very thick. The fibres and jelly are elastic. In medusae, mesoglea comprises the bulk of the animal and forms a resilient skeleton. In polyps, the water-filled coelenteron acts as a hydrostatic skeleton, which, in concert with the mesoglea, maintains the form of these animals.

Muscles in cnidarians are extensions of the bases of ectodermal and endodermal cells. Individual muscle cells are relatively long and may occur in dense tracts in jellyfish or sea anemones. Most cnidarian muscles, however, are thin sheets at the base of ectodermal and endodermal layers.

Ectodermal and endodermal muscles

In polyps, ectodermal muscles are oriented lengthwise along the cylindrical body and tentacles; endodermal ones are usually circular. Contraction of circular muscles against coelenteric fluid causes the polyp's body to elongate; contraction of longitudinal muscles causes it to shorten. Similar layers of muscles extend and contract the tentacles. Bending results from unequal contractions of longitudinal muscles on opposite sides of the body.

In medusae, all muscles are ectodermal, restricted to the concave oral surface (subumbrellar surface), and organized into circular and radial tracts. Contraction of circular muscles squeezes the subumbrellar space, forcing out contained water and causing the medusa to move by jet propulsion. Recovery of the elastic mesoglea re-extends contracted muscle fibres. Radial muscle contraction distorts the bell, directing the water jet at angles to the longitudinal axis of the bell, allowing the medusa to steer.

**Support mechanisms and skeletons.** Most members of what is usually considered a soft-bodied group have some sort of skeleton aside from the hydrostatic system described above. Both external and internal skeletons occur in the phylum, but only among polyps.

External skeleton

Most hydroid polyps secrete a horny, chitinous external skeleton that is essentially a tube around the polyp and the network of stolons that interconnect members of a colony. As well as being protective, it confers stiffness for support and has joints for flexibility. A few scyphozoan polyps have comparable chitinous skeletons. Unlike those of hydroids, hydrocoral skeletons are composed of calcium carbonate and are internal by virtue of being shallowly penetrated by channels of living tissue. Hydrocorals, which include the order Milleporina (millepores), commonly called fire coral, and the precious red coral used for jewelry, form encrusting or branching skeletons similar to those of anthozoan corals.

An anthozoan coral polyp, which resembles a sea anemone, can nearly completely retract into the calcareous cup it secrets around itself. This external skeleton underlies a continuous, superficial layer of tissue. Non-reef-forming corals typically are solitary or form small, rather delicately branched colonies, their polyps being relatively large and widely spaced. In some species of reef-forming corals, polyps are so tightly packed that their individual units share common walls. Skeletons may be encrusting, massive, or arborescent (treelike). The latter type of skeleton is delicate and typical of quiet waters at depth or in lagoons, while the former two predominate where water motion is strong. Skeleton is laid down in massive corals at a rate of about one centimetre per year; branching corals may grow considerably more rapidly. The largest corals represent cooperative efforts of up to 1,000,000 tiny individuals precipitating calcium carbonate over centuries. Few attain such proportions, however, and even the largest are eventually broken down by boring organisms such as algae, worms, sponges, and barnacles, as well as by physical processes.

Internal skeleton

The last major category of cnidarian skeletons, formed by the anthozoan subclass Alcyonaria and the order Antipatharia, are internal. Sea fan and sea whip skeletons consist of the horny protein gorgonin with calcareous spicules fused to form a solid or jointed central rod. Soft coral spicules are discrete, mostly microscopic objects of diverse shapes that vary from needle-like to club- and anchor-shaped. Located in the ectoderm, spicules stiffen the colony. In some species the several spicules that form a protective cup around each polyp may be several millimetres long. The alcyonarian *Tubipora* is known as the organ-pipe coral after the form of its red calcareous skeleton. Blue corals (the order Helioporacea) have skeletons of crystalline calcareous fibres fused into sheets, which are used for jewelry. Colonies of black coral resemble bushes and may stand more than three metres tall. Their skeletons, made entirely of proteinaceous material similar to gorgonin, are likewise used for jewelry.

Sea anemones do not produce skeletons, although their close relatives in the order Zoanthinaria incorporate foreign objects (sand grains, sponge spicules) into their body walls, which gives them rigidity and toughness. Small anemones that live high in the intertidal zone commonly inhabit abandoned barnacle tests (shells), thereby acquiring some of the benefits of a skeleton.

**Nervous system and organs of sensation.** Medusae have a more highly developed nerve net than do polyps, a feature that is associated with the more active way of life of medusae. Swimming is coordinated by the nervous system. Nervous systems that are capable of conducting nerve impulses both quickly and slowly give these animals considerable behavioral responsiveness and flexibility. Ganglia or other accumulations of nerve cell bodies are not found in cnidarians, but there are gap junctions between neurons and between neurons and effectors, which allow the transmission of nerve impulses. Statocysts, located between the tentacles or near the tentacular base, inform the animal of its orientation with respect to gravitational forces. Light-sensitive ocelli (external patches of pigment and photoreceptor cells organized in either a flat disk or a pit) occur in some medusae of each of the three classes that possess this stage. Such sensory structures are closely associated with a nerve net.

Nerve net

In the past nematocysts were considered independent effectors; that is, they were thought to fire upon appropriate stimulation, without mediation by the nervous system. Evidence, however, favours there being some organismal control over their firing, which may consist only of adjusting the threshold for firing, or the selectivity. Some scientists believe that nervous complexes associated with batteries (groups) of cnidae are the mechanism of control.

Extracellular and intracellular digestion

**Digestion, respiration, and excretion.** Food is taken in and wastes are discharged through the mouth. Extracellular digestion occurs in the coelenteron, which has, in all except hydrozoans and some tiny members of the other classes, radial projections of the wall into the coelenteron that increase the surface area. Ingested material is broken down somewhat in the coelenteron and then taken up by endodermal cells for final intracellular digestion.

Respiration and excretion in cnidarians are carried on by individual cells that obtain their oxygen directly from water—either that in the coelenteron or that of the environment—and return metabolic wastes to it. Thus, all physiological functions are carried out at no more than the tissue level of differentiation.

**Defense and aggression: nematocysts.** Cnidae range from about 10 micrometres (.0004 inch) to 100 micrometres in length. Each consists of a spherical or cigar-shaped capsule with an eversible, hollow tubule extending from one end. In the unfired state, the tubule is coiled within the capsule. When a cnidarian contacts a predator or prey item, the capsule opens and the tubule everts. The tubule may be adhesive, or it may entangle the object. Both types serve to hold food items. A third type of tubule is armed with spines that penetrate predator or prey. Toxins contained in the capsule are injected through the tubule into the object being held. Each cnida can be fired only once. Undifferentiated interstitial cells of the ectoderm and endoderm appear to be the source of the cnidoblasts (cells that produce cnidae).

## EVOLUTION

Theories of evolution

Among theories proposed on the evolution of the phylum Cnidaria, most treat the radial symmetry and tissue level of organization as evidence that the group is primitive and hold that the medusa is the original body form, being the sexually reproductive phase of the life cycle. Another theory is that the original cnidarian was a planula-like organism that preceded both polyp and medusa. In either case, Hydrozoa is considered to be the most ancient of cnidarian classes, and Trachylina is thought to be the most primitive extant order of that group. An alternative view is that anthozoans are the stem of the phylum, which evolved from bilateral flatworms and is secondarily simplified. A corollary to this theory is that the polyp is the ancestral body form.

Speculations about the origin of the phylum are not easily resolved, for preservable skeletal structures developed relatively late in cnidarian evolution. The oldest fossilized cnidarians were soft-bodied. Representatives of all four modern classes have been identified in Ediacaran fauna of the Precambrian Period (570,000,000 years ago) known from more than 20 sites worldwide. As much as 70 percent of Ediacaran species have been considered to be cnidarians. Curiously, there are few fossil cnidarians of the Cambrian Period (570,000,000 to 500,000,000 years ago). The Conulariida, which existed from the Cambrian Period to the Triassic Period (225,000,000 to 190,000,000 years ago) are considered by some scientists to be skeletal remains of scyphopolyps, either ancestral to the coronates or without modern derivatives. Presumed fossil sea anemones are found in the early Cambrian Period. Colonies of Stromatoporoidea, considered to be an order of the class Hydrozoa that extended from the mid-Cambrian Period to the Cretaceous Period (136,000,000 to 65,000,000 years ago), produced massive skeletons. Although there were two groups of Paleozoic corals, neither of which has modern descendants, they were not great reef-builders during that era. Scleractinians arose in the mid-Triassic Period; blue corals, gorgonians, millepores, and hydrocorals have records from the Jurassic Period (190,000,000 to 136,000,000 years ago) or the Cretaceous Period to the present. Most other cnidarians are known only from the Holocene Epoch (within the last 10,000 years).

## CLASSIFICATION

**Annotated classification.** The following classification, limited to living cnidarians, generally follows that used by D.G. Fautin in S.P. Parker (ed.), *Synopsis and Classification of Living Organisms,* vol. 1 (1982), and L.H. Hyman, *The Invertebrates,* vol. 1, *Protozoa Through Ctenophora* (1940).

### PHYLUM CNIDARIA (COELENTERATA)

Nematocyst-bearing, radial metazoans without organs. Have a cellular inner endoderm and outer ectoderm, separated by noncellular mesoglea. Polyp and medusa forms; either or both may be present in one life history. Most polyps have tentacles around mouth; tentacles of medusae at bell margin. One internal cavity, the coelenteron, has 1 opening to exterior, the mouth. About 9,000 species.

#### Class Anthozoa

Exclusively polypoid with biradial symmetry. Oral end a disk with central mouth and hollow tentacles arising at margin and/or on surface. Mouth leads to coelenteron via stomodaeum that has ciliated troughs (siphonoglyphs) for water transport into and out of coelenteron. Coelenteron divided by radial mesenteries that extend inward and insert on the stomodaeum (complete mesenteries) or not (incomplete mesenteries). About 6,000 species.

*Subclass Alcyonaria*

Octocorals. Polyps with 8 pinnately branched tentacles, 8 mesenteries, and a single siphonoglyph. Nearly all colonial with internal skeletons.

*Order Stolonifera.* Polyps of colony connected by stolons. Skeletons of spicules or horny external cuticle. Shallow tropical and temperate seas.

*Order Telestacea.* Long axial polyps bear lateral polyps. Skeleton of spicules fused with a horny material. Tropical.

*Order Gorgonacea.* Sea fans and sea whips. Colonies commonly arborescent with axial skeleton of gorgonin and/or calcareous spicules. Polyps rarely dimorphic. Tropical and subtropical.

*Order Alcyonacea.* Soft corals. Small to massive colonial forms. Lower parts of polyps fused into a fleshy mass; oral ends protrude. Internal skeleton of isolated calcareous spicules. Primarily tropical.

*Order Helioporacea* (*Coenothecalia*). Blue coral. Massive lobed calcareous skeleton. Tropical; 1 Caribbean and 1 Indo-West Pacific species.

*Order Pennatulacea.* Sea pens and sea pansies. Fleshy, always dimorphic, unbranched colonies, with 1 axial polyp and many lateral ones. Polyp-free peduncle burrows into soft sediments; polyp-bearing distal end of the polyp (rachis) extends into water and may be completely retractile. Central skeleton a calcified axial rod; polyps and rachis have isolated calcareous spicules.

*Subclass Ceriantipatharia*

Black corals and tube anemones.

*Order Antipatharia.* Black coral. Large bushy colonies with thorny, hornlike axial skeleton formed by small polyps with 6 simple tentacles and 1 siphonoglyph. Mostly tropical and subtropical.

*Order Ceriantharia.* Tube anemones. Solitary polyps with 2 sets of tentacles (oral and marginal) that form feltlike tubes of specialized cnidae (ptychocysts) and burrow in soft sediments. Shallow waters worldwide.

*Subclass Zoantharia*

Sea anemones and corals. Six (or multiples of 6) tentacles (rarely branched). Mesenteries commonly arranged hexamerously. Solitary or colonial. Skeletons non-spicular calcareous, horny, or lacking. Usually 2 siphonoglyphs.

*Order Actiniaria.* Sea anemones. Solitary or clonal, never colonial; lacking skeleton; with or without basilar muscles. Mostly littoral or benthic, commonly attached to firm substrata but some burrow in soft sediments. Worldwide.

*Order Corallimorpharia.* Sea-anemone-like solitary or aggregated polyps lacking basilar muscles and skeleton. Coral-like muscles and nematocysts. Mostly tropical.

*Order Ptychodactiaria.* Sea-anemone-like, lacking ciliated tract on edge of mesenteries and basilar muscles. Both poles.

*Order Scleractinia* (*Madreporaria*). True or stony corals.

Mostly colonial; calcareous external skeleton; no basilar muscles or siphonoglyphs. Mostly tropical and subtropical.

*Order Zoanthinaria (Zoanthidea).* Solitary, clonal, or colonial polyps resembling sea anemones. Lack skeleton but may incorporate debris into body wall, commonly epizoic. One complete and 1 incomplete mesentery per pair. Mostly tropical.

**Class Cubozoa**

Tropical, cuboidal medusae that swim strongly; box jellyfishes. Margin simple with single or grouped tentacles arising above the 4 corners. Polypoid stage of most species unknown. Fiercely stinging members can cause human fatalities. Contains 1 order, Cubomedusae.

**Class Hydrozoa**

Life histories may involve both polypoid and medusoid stages, but either may be suppressed or absent. Tetramerous or radially symmetrical medusae small, with shelf of tissue (velum) across lower part of bell, which reduces diameter of subumbrellar aperture (condition known as craspedote). Colonial forms commonly polymorphic. Coelenteron undivided. Gametes ripen in ectoderm. Only class with some freshwater members. 2,700 species.

*Order Actinulida.* Curious group of solitary, motile cnidarians with features of both polyps and medusae. Europe; in marine sand.

*Order Chondrophora.* Floating polymorphic colonies supported by chitinous skeleton. Free medusae are produced; includes *Velella.* Oceanic; worldwide.

*Order Hydroida.* Hydroids. Usually colonial and polymorphic; release free medusae or retain modified medusoid reproductive structures on polyp colony. Polyps usually have a chitinous exoskeleton. Includes naked, solitary freshwater polyp *Hydra.* Largest order of Hydrozoa.

*Suborder Anthomedusae.* Medusae bell-shaped, with gonads on the stomach or sides of manubrium. Sensory structures consist of pigmented eyespots (ocelli). Skeleton, if present, lacks cup (hydrotheca) into which polyp may withdraw (a condition known as gymnoblastic); few species with calcareous exoskeleton. Most abundant in bays and shallow coastal waters.

*Suborder Leptomedusae.* Medusae saucer-shaped, but lacking in many species. Gonads on radial canals. Sensory structures usually statocysts. Hydroids with hydrothecae (condition known as calyptoblastic). All shallow marine waters.

*Suborder Limnomedusae.* Small medusae with gonads on stomach walls or radial canals. Polyps solitary or colonial, commonly with 1 or 2 tentacles, and no skeleton. Mostly freshwater.

*Order Milleporina.* Fire coral. Colonial forms producing massive calcareous skeletons. Gastrozooids and dactylozooids project through pores in surface of skeleton. Reduced, acraspedote (lacking a velum) nonfeeding medusae are released. Tropical.

*Order Siphonophora.* Pelagic polypoid colonies with greatest degree of polymorphism in phylum; lack medusae. Oceanic; worldwide. Includes Portuguese man-of-war, *Physalia.*

*Order Stylasterina.* Hydrocorals. Resembling millepores; colonies erect and branching or prostrate. Commonly yellow, red, or purple. Reduced medusae not freed; develop and produce gametes in cavities of skeleton (ampullae). Worldwide; includes precious red coral, *Corallium.*

*Order Trachylina.* Medusa dominant; reduced or no polyp stage. Statocysts and special sensory structures (tentaculocysts). Differ from other hydromedusae by having tentacles inserted above umbrellar margin. Oceanic, mostly warmer waters.

*Suborder Laingiomedusae.* Medusae with features of both Narcomedusae and Trachymedusae. Polyp unknown.

*Suborder Narcomedusae.* Scalloped margin; gonads on stomach walls. Manubrium lacking.

*Suborder Trachymedusae.* Smooth bell margin; gonads on radial canals arising from the stomach. Polyp and asexual reproduction absent.

**Class Scyphozoa**

Exclusively marine group in which acraspedote medusae predominate. Life histories commonly involve alternation of a very small polyp, the scyphistoma, with a medusa, which develops from an ephyra released by the polyp. Coelenteron of both divided by 4 longitudinal septa producing tetramerous radial symmetry. Gonads endodermal. Marginal sensory structures (rhopalia) with statocysts and/or ocelli. Most abundant in coastal waters, but oceanic species exist. About 200 species.

*Order Coronatae.* Large medusae that are conical, dome-shaped, or flattened, with furrow around bell above scalloped margin. Some species have scyphistoma stage with external chitinous skeleton. Oceanic, some species living at great depths.

*Order Rhizostomae.* Medusae like those of Semaeostomeae but with mouth subdivided into minute pores that connect with coelenteron. Mostly tropical. Deep-water species may lack polypoid stage.

*Order Semaeostomeae.* Most common and best known jellyfishes. Full alternation of polyp and medusa stages. Bell domed or flattened, with the margin scalloped into 8 or more sections. Edges of single mouth drawn out into 4 long arms. Most species in warm, coastal waters, a few in frigid waters; some oceanic. Includes the giant *Cyanea arctica,* which may attain 2 m in diameter.

*Order Stauromedusae.* Sessile jellyfish that are vase-, goblet-, or trumpet-shaped, and usually bear 8 groups of tentacles. No more than 2–3 cm long. Apparently lacking polypoid stage. Temperate and cold temperate waters worldwide.

**Critical appraisal.** The classification of living cnidarians is relatively stable and generally accepted. One unanswered question relates to their evolutionary position among the lower Metazoa (a division of the animal kingdom that includes all phyla except the Protozoa). Members of the small phylum Ctenophora (comb jellies, sea walnuts) are superficially similar to medusae, having a gelatinous body with one opening, and tentacles. The one species of Ctenophora possessing nematocysts had been thought to link the phyla. Mills and Miller have shown that members of the phylum Ctenophora obtain their cnidae by preying on a medusa, and the similarity in body form is considered convergence due to the pelagic way of life.

Anthozoa is a well-defined, coherent group, but relationships among its components are poorly understood, and the ranking of some of them is disputed. Some regard corallimorpharians as scleractinians that lack a skeleton. Similarity of larval ceriantharians to antipatharian polyps is the rationale for subclass Ceriantipatharia. Morphology of antipatharians is, however, in some ways, nearer that of alcyonarians than of zoantharians, and alternative schemes place Antipatharia in subclass Alcyonaria. Ceriantharia, too, sits uncomfortably in Zoantharia, although it bears no special relationship to the Alcyonaria. It is generally viewed as a divergent, early offshoot of the anthozoan line. Treatment of both Antipatharia and Ceriantharia as distinct subclasses of Anthozoa not particularly closely related to one another might better express their evolutionary relationships.

Many cnidarian biologists continue to regard Cubozoa as an order of Scyphozoa. Cubozoans have features of both Scyphozoa and Hydrozoa, but the complete metamorphosis of polyp into medusa supports its placement in a class intermediate between the other two.

Hydrozoan suborder Limnomedusae is not accepted by some workers, and the groups assigned to it are treated as members of the other suborders. In substantial ways, characters of many Limnomedusae bridge the differences between the Anthomedusae and the Leptomedusae. The hydrozoan order Chondrophora has morphological characteristics suggesting that the group might be better treated as part of Anthomedusae than as a discrete order.

BIBLIOGRAPHY. The standard technical treatment of the Cnidaria is LIBBIE HENRIETTA HYMAN, *The Invertebrates,* vol. 1, *Protozoa Through Ctenophora* (1940). The history of the phylum as deduced from the fossil record is covered in RAYMOND C. MOORE (ed.), *Treatise on Invertebrate Paleontology,* part F, *Coelenterata* (1956, reprinted 1967). C.E. MILLS and R.L. MILLER, "Ingestion of a Medusa (*Aegina citrea*) by the Nematocyst-Containing Ctenophore *Haeckelia rubra* (formerly *Euchlora rubra*): Phylogenetic Implications," *Marine Biology,* 78(2):215–221 (1984), offers information on the relationship between Ctenophora and Cnidaria. The following proceedings of international symposia on cnidarians provide useful information: W.J. REES (ed.), *The Cnidaria and Their Evolution* (1966); TAKASI TOKIOKA and SABURŌ NISHIMURA (eds.), *Recent Trends in Research in Coelenterate Biology* (1973); G.O. MACKIE (ed.), *Coelenterate Ecology and Behavior* (1976); and P. TARDENT and R. TARDENT (eds.), *Developmental and Cellular Biology of Coelenterates* (1980). See also the following proceedings of international coral reef symposia: *Proceedings of the Third International Coral Reef Symposium,* 2 vol. (1977); *The Reef and Man: Proceedings of the Fourth International Coral Reef Symposium,* 2 vol. (1982); and *French Polynesian Coral Reefs: Proceedings of the Fifth International Coral Reef Congress,* 6 vol. (1985). The major justification of recognizing class Cubozoa is B. WERNER, "Bau und Lebensgeschichte des Polypen von *Tripedalia cystophora* (Cubozoa, class. nov., Carybdeidae) und seine Bedeutung für die Evolution der Cnidaria," *Helgoländer wissenschaftliche Meeresuntersuchungen,* 27:461–504 (1975).

(C.H.Ha./D.G.Fa.)

# Coins and Coinage

The use of cast-metal pieces as a medium of exchange is very ancient and probably developed out of the use in commerce of ordinary ingots of bronze and other metals that possessed an intrinsic value. Until the development of bills of exchange in medieval Europe and paper currency in medieval China, metal coins were the only such medium. Despite their diminished use in most commercial transactions, coins are still indispensable to civilized economics; in fact, their importance is growing as the result of the widespread use of coin-operated machines. The term numismatics (from the Latin *numisma* and Greek *nomisma,* "coin") denotes the study of coins, including medals but excluding seals.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* Part Five, Division III, Section 533.

This article is divided into the following sections:


Numismatics 529
- Coins as historical data 529
- Coin collections and specialist societies 530

History of coinage 530
- Origins 530
- Ancient Greek coins 531
  - Early developments, *c.* 650–490 BC
  - From the Persian Wars to Alexander the Great, 490–336 BC
  - From Alexander the Great to the end of the Roman Republic, *c.* 336–31 BC
  - Subsidiary Greek silver coinages under the Roman Empire
  - Coinage in Judaea
  - Greek bronze imperial coinage, to AD 268
- Roman coins, Republic and Empire 534
  - The beginnings
  - Introduction of the denarius
  - Control and content of the coinage
  - Caesar and after
  - The 4th century and after
- Coinage in western continental Europe, Africa, and the Byzantine Empire (5th–10th century) 537
  - Post-Roman coinage in the West
  - Coinage in the Byzantine Empire
  - Charlemagne and the Carolingian coinages
- The later medieval and modern coinages of continental Europe 539
  - Portugal
  - Spain
  - France
  - The Low Countries
  - Switzerland
  - Italy and Sicily
  - Germany and central Europe
  - Scandinavia
  - Poland
  - Russia and the Balkans
  - The later Byzantine empires
- Coins of the British Isles, colonies, and Commonwealth 543
  - Ancient Britain
  - Roman Britain
  - Early Anglo-Saxon coins
  - Anglo-Saxon penny coinages
  - Post-Conquest coinage
  - Modern coinage
  - Scotland
  - Ireland
  - Isle of Man and Channel Islands
  - Colonies and Commonwealth
- Coins of Latin America 546
  - The colonial period
  - Hispanic-American colonial mints
  - Dissemination of Hispanic-American coinage
  - Emergency coinages in the era of independence
  - The independent countries
  - Brazil
- Coins of the United States 546
- Coins of Asia 547
  - Ancient Persia
  - Islāmic coins of the West and of western and Central Asia
  - India
  - China
  - Japan
  - Korea
  - Vietnam, Kampuchea (Cambodia), Laos
  - Burma and Thailand
- Coins of Africa 550
  - North Africa
  - Sub-Saharan Africa
- Medals and medallic art 551
  - Italy
  - France
  - Germany and Austria
  - The Netherlands
  - The Baroque period
- Techniques of production 552
  - Ancient minting
  - Medieval minting
  - Early modern minting
  - Contemporary mints

Bibliography 555


## Numismatics

### COINS AS HISTORICAL DATA

Being made in most ages of precious metal, or alternatively possessing a substantial token value, coins have always been prized, often hoarded, and, therefore, frequently buried for safety. The contents of such savings banks have been dug up in all ages, so that the coins of past civilizations continue to be found in vast numbers. Studied alongside literary or archaeological evidence, they yield a wide range of information that is especially valuable for chronology and economic history. Coins may reflect the wealth and power of cities and states, and study of their distribution may help to define the physical extent of territorial dominion or to illustrate major commercial connections. Thus, the popularity in ancient times of Athenian silver tetradrachms (coins worth four drachmas, the drachma having a unit weight of about 4.25 grams) in the Levant and of Corinthian silver staters (ancient Greek coins of various weight standards) in Magna Graecia (in southern Italy) testifies to established trade links; finds of early Roman imperial gold in India corroborate the reference of the Roman historian Pliny the Elder to the drain on Roman gold to pay for Indian and other Eastern luxuries; and huge finds of Arab silver coins in Scandinavia show the extent of trade, in particular the demand for furs by the ʿAbbāsid caliphs and the Sāmānid rulers of Iran. One result of such widespread commercial contacts is that certain currencies acquired special international preeminence. In ancient times, those of Athens, Corinth, and Philip II of Macedon were widely popular; so, in medieval times, were the gold dinars (a term derived from the Roman denarius) of the early caliphs and the gold ducats of Florence and Venice, while in modern times the silver dollars of Mexico and Maria Theresa of Austria and the gold sovereigns of Great Britain played a similar role. Moreover, the study of depreciation and debasement of coinage may illuminate past national financial distress; the heavily alloyed 3rd-century-AD Roman antoniniani (coins introduced by the Roman emperor Antoninus, originally

Internationally pre-eminent currencies

having a value of two denarii) tell their tale as clearly as the depreciating paper currency of Germany in and after 1919.

No less valuable than the economic evidence yielded by a comparative study of coins is their purely documentary importance. Together with medals, they present an unrivaled series of historical portraits from the 4th century BC to the present day, many of them otherwise unknown, like the Greco-Bactrian kings or certain usurpers during the Roman Empire. Greek coinage is a particularly notable contribution to the history of art, displaying not only the beauty and strength of many artistic traditions but also (like Roman coinage) the miniature likenesses of numerous large-scale sculptural and architectural works now lost. The imperial coinage of Rome, apart from its portraiture, is important above all for the remarkable detail of its chronological and political content; and from both Greek and Roman coins much can be learned of mythology and religion. The Christian influences active in medieval Europe can be similarly measured from medieval currencies.

The principal metals of which ancient coins were made were electrum (an alloy of silver and gold), gold, silver, copper, brass, and bronze—all of them more or less proof against decay. Their use at first was generally dictated by availability. The earliest coins of Asia Minor were of electrum, a natural alloy (later produced artificially) washed from Lydian rivers; gold became the major currency metal of southwestern Asia as a whole, being derived from Scythian, Pontic, and Bactrian sources. The city-states of the Greek mainland preferred the silver that adjacent mines supplied, and the mines of Italy led to the choice of bronze for the earliest coinage of Rome. With the development of internal economies and external trade, gold, silver, and copper or bronze quickly came to be used side by side; Philip II of Macedon popularized gold in Greece, and gold, together with silver, competed strongly with copper in the Roman imperial currency, becoming paramount in the Byzantine and Arab empires and in the great commercial currencies of the Italian republics of the 13th century onward. Silver, however, was nearly always powerful in Roman currency and was the major coinage metal of Europe from the 8th to the 13th century. Bronze or copper was first used for small change in Greece from the late 5th century BC and in the Roman and Byzantine systems as well; the vast currency of China consisted of base metals down to modern times.

The metals of currency

The foregoing metals furnished most currencies until the early 20th century, when the appreciation in value of gold and silver and the need to economize led to the general production of paper currencies for the higher units of value, accompanied by token units of lower value expressed in terms of nickel (used, exceptionally, in Bactria in the 2nd century BC), cupronickel, bronze, and, in times of postwar stress, aluminum and aluminum bronze. Lead, which may easily decay, has seldom been used for coinage, except by the Andhras, inhabitants of the Deccan in ancient India; in pre-Roman Gaul; and in the more recent coinages of the Malay states. Iron, very occasionally used in antiquity—*e.g.*, in Sparta—reappeared in German coins of World War I. Zinc was employed by Rome as a constituent of fine brass coins and as an element in the alloy of a few Chinese coins from the 15th to the 17th century. Base metals furnished the material for some Celtic coins in Gaul and Britain in the last century BC. In crises, currencies have been produced from leather, cloth, card, paper, and other materials.

#### COIN COLLECTIONS AND SPECIALIST SOCIETIES

The enormous number of coins produced from earliest times has resulted in organized collecting over a long period. The continuous history of coin collecting begins with the Italian Renaissance, and Petrarch was characteristic of his time in forming a classical series. During the 15th and 16th centuries many collections were made by princes or nobles, for whom Greco-Roman coins possessed both moral and aesthetic appeal. Among the more famous cabinets were those of Jean, duc de Berry, of the d'Este family, of the emperor Maximilian I, and of Matthias Corvinus. The two last collections became the nuclei, respectively, of the present national collections of Austria and Hungary; later the cabinet of Louis XIV was to serve France similarly, just as that of the Stuarts might have served England but for its dispersal in the Puritan revolution.

In the 17th century numismatic scholars began to catalog and document existing collections. Italy possessed more than 350, France and the Low Countries about 200 each, and Germany not many fewer. The advent of scholarly numismatic compilations had important results. Distinction between the genuine and the spurious became surer; analytical syntheses based on detailed catalogs began to teach the principles of scientific numismatics; the recording of new material was all the more keenly undertaken; and the part played by numismatic evidence in historical reconstruction was increasingly understood.

From the 18th century onward it was therefore all the more important to collect on a scale at once wide and discriminating; and whether in charge of a royal cabinet, like the eminent Joseph Eckhel (1737–98) at Vienna, or the possessor of a splendid private collection, like William Hunter (1718–83) at Glasgow, the 18th-century collectors made a great contribution. Lesser collectors could also advance the science; their numbers were to grow in the 19th century with the output of authoritative catalogs (including the British Museum series from 1873) and informed handbooks. This growth was reflected in the foundation in many European countries of specialist societies responsible for scholarly publications. But the day of the great private collection was not yet done: superb cabinets were formed in the late 19th century, and those of Richard Lockett, Virgil Brand, and Emily Norweb in the 20th bore comparison with all except the great museum collections. In general, however, museums have taken over the main task of forming large collections; those of London, Oxford, Cambridge, Glasgow, Paris, Berlin, Vienna, Munich, Boston, and New York City are among the richest.

Specialist societies

London emerged as, and still remains, the world's largest numismatic market (followed by Zurich), serving the interests of public collections and private collectors in many lands. Inasmuch as these interests are, jointly, directed increasingly toward the systematic elucidation of historical and economic problems, international cooperation has become more important, being exercised through the International Numismatic Commission, itself associated with the International Committee for Historical Sciences. But below this apex there spreads out a vast body of private collectors in many lands, whose interests may often have been stimulated in childhood by the chance gift or discovery of a handful of coins.

## History of coinage

#### ORIGINS

In both the East and the West, coinage proper was preceded by more primitive currencies, nonmonetary or semimonetary, which survived into the historic age of true coins, and may have derived originally from the barter of cattle, implements, and the like. The earliest currency of China of the 8th century BC consisted of miniature hoes and billhooks (pruning implements), with inscriptions indicating the authority. The small bronze celts (prehistoric tools resembling chisels) frequently found in hoards in western Europe probably played a monetary role. Even in modern times such mediums of exchange as fishhook currency have been known.

Metal has always achieved wide popularity as an exchange medium, being durable, divisible, and portable; and the origins of true coinage lie there. Ancient Egypt, which never developed a true coinage, was using gold bars of set weight from the 4th millennium BC; and a currency of gold rings was thereafter common. In the Middle East, also, gold rings long served the dual purposes of adornment and currency, supplemented by gold and silver bars from which segments could be cut. The choice of metal was, as usual, determined by availability. Around the Aegean Sea heavy talents (ancient units of weight and, later, of monetary value) of copper, ingots of 55 pounds (25 kilograms) or more, were in currency several centuries before true coinage, and the discovery of an iron bar with

a handful (*drachma*) of fractional iron spits (*obeloi*) dedicated in the Heraeum (a temple of the goddess Hera) at Argos, perhaps as part of King Pheidon of Argos' reforms of weights and measures in the 7th century BC, shows such currency continuing until historical times. Similar bundles of spits have been found elsewhere and are evidence of the desire to subdivide a cumbersome unit into smaller fractions for normal use. At the other end of the scale, there was, ultimately, the desire to express the value of a talent of copper or iron in terms of gold or silver; and Homer, who speaks of metal basins, tripods, and axes as gifts and prizes in a way that shows them as a recognized standard of wealth, also speaks of the talent of gold (*i.e.,* the value of a heavy base-metal talent expressed in a little pellet of gold). In Italy rough lumps of bronze (aes rude) formed a currency from early times, being succeeded by bars of regular weight; and Julius Caesar's record of the ancient British use of iron bars as currency (following his raids on Britain in 55 and 54 BC) is still borne out by not infrequent finds.

"Heavy" currencies

Such "heavy" currencies, mainly characteristic of European lands, show the employment of metals from which implements would normally be made. The impact upon this system of the gold of the East, and later of the silver of Greece, produced the need to value such metals in gold and silver, and this in turn resulted in the need to control and guarantee the quantity of gold and silver so used to avoid constant weighing. Once gold (and then silver) gained acceptance as conveniently small expressions of relatively high value, with a visible mark of guarantee, the stage of true coinage, as it first appeared in Asia Minor and India, had been reached. Not all lands, however, adopted true coinage: the easternmost fringes of the Greek world lacked it, and Carthage and Etruria were without coinage until the 5th century.

## ANCIENT GREEK COINS

**Early developments, c. 650–490 BC.** True coinage began soon after 650 BC. The 6th-century Greek poet Xenophanes, quoted by the historian Herodotus, ascribed its invention to the Lydians, "the first to strike and use coins of gold and silver." King Croesus of Lydia (reigned *c.* 560–546 BC) produced a bimetallic system of pure gold and pure silver coins, but the foundation deposit of the Artemisium (temple to Artemis) at Ephesus shows that electrum coins were in production before Croesus, possibly under King Gyges. Croesus' earliest coins were of electrum, which the Greeks called "white gold." They were stamped on one side with the facing heads of a lion and a bull; this type was later transferred to his bimetallic series of pure gold and pure silver. (Some recent scholarship, however, suggests that this latter series was struck, in fact, under Croesus' Persian successors.)

The early electrum coinage consisted of small, thick, bean-shaped pieces, with a device stamped in relief on one side, the other being roughly impressed. Their intrinsic value fluctuated according to their gold and silver content; but the weight of the unit was fairly steady at about seven to eight grams, and the types stamped on them were the guarantee of authority.

Croesus' relations with Greece were close, and his bimetallic system may have owed something to the fact that Greece had itself now produced its first silver coins. The oldest are of Aegina, with, obverse, a turtle—associated with Aphrodite—and, reverse, an incuse square. Tradition—*e.g.,* in Julius Pollux, the 2nd-century-AD Greek scholar, and elsewhere—regarded these as struck by Pheidon of Argos in virtue of his supremacy over Aegina; but the coins are too late to claim association with him in Aegina. They began no earlier than the late 7th century, when Aeginetan maritime ascendancy was growing, incidentally spreading the Aeginetan weight standard for coinage, based on a drachma of about six grams, over much of the Peloponnese and also the Aegean, where similar currency was produced in the islands. Ambition and pride stimulated two neighbouring powers to strike their own coins. Corinth with its pegasi (from their constant obverse type of a pegasus) was coining silver from *c.* 575 with a light drachma of about three grams, and it is reasonably certain that in Athens, in the first half of the 6th century, Attic coins, based on a drachma of about 4.25 grams derived from Euboea and with a variety of obverse types, including an owl (the reverses, like those of the Corinthian pegasi, were impressed with a die design), were supplanting the earlier coinage of Aegina. These early silver coins, while much less valuable intrinsically than the electrum and gold coins of Asia Minor, nevertheless possessed considerable purchasing power: the Aeginetan and Attic-Euboic didrachms and the Corinthian tridrachm were high denominations suitable for major commerce and not for everyday life. For intercity transactions these staters (*i.e.,* standard units) were conveniently linked by the mina weight (1/60 of a talent) of 425 grams, made up by 150 Corinthian, 100 Attic, and 70 Aeginetan drachmas. Fractional pieces developed only slowly.

Between 550 and 500 commerce and civic pride had spread coinage to many parts of the Greek world. From the Persian Empire, with its vast gold and silver coinage, successor to that of Croesus, to Magna Graecia and Sicily, and from the Dorian colony of Cyrene to the Greek or semi-Greek cities of Thrace, there was a network of varied and competitive currencies, generally of fine quality and steady weight. Improved minting techniques began to affect their appearance. A second type, in relief, was substituted gradually for the roughly impressed reverse punch. The important effect of this on the development of coin types is well seen in the reorganized coinage of Athens from *c.* 525, in which the obverse bears the Athena head and the reverse the owl of Athens—religious patron and civic device; the monarch's head on an English penny goes back, through Alexander's deified head, to the head of Athena, and the symbol of Britannia derives ultimately from such state badges as the owl. In certain cities of Italy and Sicily, however, including Tarentum and Metapontum, a different technique was popular, the obverse type in relief being repeated intaglio on the reverse, very probably with the object of concealing the older types of coins imported for restriking. For a long time the early coins of Greece carried no inscriptions or, at most, with very rare exceptions, a letter or two referring to the issuing city or state authority.

Athenian relief types

Greek coin types, early and even later, were simple in conception and often taken from the animal world. They include many kinds of animals (with the bull, symbol of a river, very common); birds (such as the owl of Athens, the eagle of Zeus at Olympia, the dove at Sicyon); insects (like the bee of Ephesus); fabulous creatures (like the griffin at Abdera); and vegetable objects. Not uncommonly such types were chosen as punning allusions to a city's name—the lion at Leontini; the goat at Aegae; the quince at Melos; the sickle-shaped harbour at Zancle; the *selinon* leaf at Selinus; the cock, harbinger of *hemera,* the day, at Himera. In others a city's staple product was proclaimed, like silphium at Cyrene, a silver-miner's pick at Damastium, a bunch of grapes at Naxos, a wine jar at Chios. Cult associations frequently dictated the choice of type. Tarentum showed its mythical founder, the dolphin-rider Taras; Knossos, the Minotaur (half man, half bull) or Labyrinth; Croton, the tripod of Apollo; Poseidonia, a statue of Poseidon, god of the sea. Human or anthropomorphic figures, however, were comparatively rare on early Greek coins, though the famous gold darics, a name derived from Darius I, and silver shekels of Persia showed the great king in an attitude of attack. Much more popular was the representation of idealized heads of deities, which, once established for the two Athenas, Parthenos and Chalinitis, at Athens and Corinth, quickly became the vogue elsewhere, encouraged by the development of double-relief coinage (*i.e.,* coinage with obverse and reverse in relief), which allowed the head of a civic deity to be paired on the other side by the city's symbol. The Greek tyrants, as a rule, chose to respect the theory of coinage as the corporate expression of state economy and thus regarded coinage as too important a matter for private production. The traditions that, rightly or wrongly, associated all the great lawgivers—Pheidon, Solon, and Lycurgus—with the institution of coinage as well as with reform of weights emphasize its position as a fundamentally corporate right.

The Sicilian decadrachm

In Sicily the defeat of Carthage in 480 BC may have been commemorated by the famous decadrachms (*Demareteia*) associated with Queen Demarete, wife of King Gelon. These superb and now very rare examples of early classical genius showed on the obverse the head of Arethusa (the fountain nymph of Syracusan Ortygia), wreathed (possibly for victory), and on the reverse a chariot above a fleeing lion.

**From the Persian Wars to Alexander the Great, 490–336 BC.** For a century and a half the previous pattern of Greek coinage spread widely all over the Greek world, its quantity stimulated by a growing sense of nationalism, its intrinsic quality kept high by commercial competition, and its technique raised to new and often superb levels in an age of self-confidence. In the last half of the period the designing and engraving of coin dies (punches) reached a standard rarely to be surpassed. The head of a patron deity was now generally established as the obverse type and was often shown in very high relief, sometimes indeed facing, as a tour de force. Engravers, especially in Sicily and Italy, began to sign their dies, thus preserving the names of masters such as the Syracusans Euainetos, Cimon, and Eukleidas, otherwise unknown. Reverse types, now more complex, increasingly showed groups or genre scenes—*e.g.*, the splendid frontally squatting Silenus (foster father of the wine god Dionysus) on the coinage for the refounding of Sicilian Naxos in 461, Dionysus seated backward on a donkey at Mende, or the many mythological compositions on Cretan coins—often diminishing the previous importance of the city badge. Inscriptions, though still often contracted, were in general use. The principal coinage metal was silver, of which the Attic weight standard gradually conquered the Aeginetan. Electrum was continued in the east—at Cyzicus, Lampsacus, Mytilene, and Phocaea—traveling thence mainly to the Black Sea; in the west it was coined at Carthage. In both areas it was produced as an artificial alloy. Gold was continued in the darics of the Persian kings and in the fine later series of Lampsacene staters; it was also struck at Panticapaeum on the Black Sea and on occasion at Syracuse, Tarentum, and Cyrene. Toward the end of the period, Philip II of Macedon instituted what was to be a world-famous gold coinage, undercutting and ousting that of Persia. Bronze made its appearance late in the 5th century, replacing the minute silver obol and other fractional silver coins that had hitherto been used as small change.

The currencies of the period included a few that were of world importance. The silver of Corinth and its Adriatic colonies was very numerous and was abundantly accepted, outside the Corinthian territories, by Italy and Sicily. The electrum of Cyzicus bore types that deliberately recommended it to many markets. Persian gold and silver coins enjoyed immense popularity in the 5th century. Metapontum, Tarentum, Thurium, Velia, and Syracuse were among the more prolific silver mints of the west. But the most famous commercial currency of all was that of Athens, the silver tetradrachms of which were struck in large numbers, fine quality, and obstinately unchanged appearance. These coins traveled widely in trade and were imitated as far afield as Egypt, Arabia, and Persia.

*Predominance of Athens.* Economic expansion and naval hegemony gave Athens near-imperial control over its allies in the 5th century. It may have been as early as 449 that Athenian edicts forbade the striking of silver coins by the allies or the use of currency, weights, and measures other than the Athenian and provided that previously minted local currencies should be handed in for exchange against that of Athens. The subjection of Aegina to Athens from 456 and the cessation of its famous and long-competitive "turtles" facilitated the monetary dominance of the "owls," which was carried further, stage by stage, as Athenian "allies" revolted, were reconquered, and lost their independence. But the embargo put by Athens on local silver coinage was not absolute and perhaps was not expected to be. Major allies such as Samos, Chios, and Lesbos continued their own currencies; Phocaea, Mytilene, and Cyzicus, though ceasing to coin in silver, continued with electrum. In other cities, small change in silver was issued. Beyond the effective range of Athenian power, cities in Pamphylia (*e.g.*, Aspendus) and in Thrace (*e.g.*, Abdera and Aenus) could continue silver coinage on a non-Attic standard, and the failure of Athenian control is seen in the sudden and often beautiful coinages of cities that threw aside its dominance, such as Olynthus from *c.* 430, or in the changed weight standard of others (*e.g.*, Acanthus). During the Peloponnesian War, Sparta cut off the supply of silver from the Laurium mines, and by 407 Athens was melting the gold Nikai (victory statues) from the Parthenon to make emergency coins, followed the next year by bronze small change—an unpopular substitute for the tiny silver coins previously carried in the mouth. City after city now rebelled against Athens, and there was a sudden burst of independent coinage.

Athenian coinage revived, with unchanged types, after the Athenian admiral Conon's successes against the Spartan fleet in 394. But wary former allies formed defensive leagues, as shown by current coinage, with a type of the Greek hero Heracles and the inscription ΣΥΝ ("the alliance"), by Cnidus, Ephesus, Samos, Byzantium, and other cities under Rhodian leadership. Rhodes spread its own coinage (with its head of the sun-god Helios and punning badge of a rose—Rhodon) widely in the eastern Mediterranean. Phocaea and Mytilene established a monetary union for their electrum. From 404 the Aeginetans were coining again, and on their former weight standard, though with a tortoise replacing the turtle. Corinthian coins continued to pour out. In the north a variety of important mints opened, and coins from mints in Asia Minor, notably Cnidus and Ephesus, testify to the prosperity brought by autonomy.

Monetary unions

*Artistic development.* In contrast to the deliberate archaism of Athenian types, a wide flowering was seen elsewhere. Sometimes this was the result of hybridizing influence, as when Greek artists rendered Scythian motifs at Panticapaeum or Punic ones for Carthage and such of its Sicilian colonies as Segesta and Eryx. Sometimes an artistic tradition was regional, harsh, and arresting, as in Crete or, as in Massilia and Emporion in the far west, a weak reflection of finer styles. Generally, however, there was an internationally high standard in coin design. Elis, guardian of the temple of Olympian Zeus and famous for its quadrennial Olympic Games, no doubt attempted to impress visitors with its superb coinage. On the coins issued from *c.* 500 to 322 the thunderbolt and eagle of Zeus were shown with Victory in various attitudes; later the heads of Zeus and Hera were nobly represented. In northern Greece brilliant artistry characterized the coins of Amphipolis, Acanthus, and Chalcidian Olynthus. The coins of Clazomenae and Cnidus in eastern Greece were also notable for their designs.

It was in Italy and Sicily that the finest work appeared. In Italy, Tarentine silver continued its type of Taras on a dolphin. In the middle of the 5th century the agonistic type showing a horseman appeared; the celebrated Tarentine cavalry was thus commemorated down to the middle of the 4th century. About 340 Tarentum issued very beautiful gold coins with a head of Persephone and, on the reverse, the infant Taras appealing to Zeus enthroned. Heraclea, founded in the middle of the 5th century, issued fine staters with a helmeted Athena and Heracles seated or strangling or wrestling with a lion. Metapontum introduced a most striking head of its founder, Leucippus. Other mints of the time were at Neapolis, with its types of the siren Parthenope and her father, the man-headed bull Achelous; at Velia, with its head of a nymph and, on the reverse, the eastern type of a lion attacking a bull; at Thurium, with its unusually fine head of Athena and the powerful bull on the reverse; and at Terina, remarkable for its beautiful treatment of the Victory type.

In Sicily, and particularly in Syracuse, the engraver's art reached perfection. The coins of Syracuse showed many varieties of the heads of Arethusa and Persephone, and the chariot of the reverse was found capable of varied treatment. After the middle of the 5th century, artists began to sign their work, and it is thus possible to prove that other towns engaged engravers from Syracuse. The Syracusan coinage was mainly silver. During the siege by the Athenians, beautiful little gold coins were struck with,

Zenith of Syracusan coinage

reverse, Heracles strangling a lion. With the prosperity following the enemy's defeat, Syracusan art reached its zenith. As the Demareteion commemorated the defeat of the Carthaginians, so the great series of decadrachms perpetuated the memory of the victory of 413 over the Athenians. The agonistic types and the word *athla* on some of them show that they were distributed at the games held to celebrate the victory; their types were widely copied, and their engravers, Cimon and Euainetos, gained a place among the world's greatest artists.

Among other cities of Sicily there was a notable series from Acragas in the 5th century, with its beautiful double-eagle type, seen most magnificently on the rare and famous decadrachms. Camarina showed fine types of the river god Hipparis and the nymph Camarina on a swan. Himera, before its destruction in 409, issued some very interesting types, such as the nymph Himera sacrificing while Silenus beside her bathes at the thermal spring for which Himera was noted; or Pelops (a grandson of Zeus) in his chariot, referring to a victory of a Himeran at the Olympic Games, which Pelops is said to have founded. Catana used the artist Heracleidas to design a splendid facing head of Apollo. Selinus abandoned its parsley leaf and issued some remarkable types, notably that of Apollo and Artemis in their quadriga and, on the reverse, the local hero sacrificing at an altar, alluding to the cessation of the plague as a result of appeals to Apollo as healer.



Silver tetradrachm from Syracuse, Italy, signed by the engraver Cimon above the headband of the nymph Arethusa, *c.* 410 BC. In the British Museum. Diameter 28 mm.

Reproduced with permission of the trustees of the British Museum; photograph, Ray Gardner for The Hamlyn Publishing Group Limited

**From Alexander the Great to the end of the Roman Republic, c. 336–31 BC.** Alexander introduced a new era in coinage, struck in vast quantities at a variety of mints from Macedonia to Babylon with uniform types and weights. After his death in 323 BC the Diadochi ("Successors"—a reference to the chief officers who partitioned his empire) were to reflect the importance of his coinage in their own differentiated issues—Seleucus in Syria, Philip Arrhidaeus in Macedonia, Lysimachus in Thrace, and Ptolemy in Egypt, where, except for tentative gold coined by Tachos and Nectanebo II between 361 and 343, no coinage had previously been struck. Alexander's influence on the Greek fringe was no less marked. The Arsacid kings of Parthia instituted a Greek style of coinage, as did Bactrian kings, culminating in the splendid portrait decadrachms of Amyntas *c.* 150 BC, while, even farther to the southeast, Indo-Greek kings struck coins, inscribed in both Greek and Prākrit, to the end of the 2nd century. The flood of coins of Philip II and Alexander, penetrating Europe from the Balkans, resulted in progressive imitations by Celtic peoples westward along the Danube until these imitations themselves influenced coins in Gaul and Britain in the 1st century BC. In the Mediterranean west, by contrast, Greek coinage yielded to the steady advance of Roman power; the late issues of Spain and Mauretania were of hybrid Greco-Roman origin.

*The coin portrait.* The coinage of Alexander established a new style: the coin portrait became an almost regular feature in Greek currency that was predominantly regal. The portrait, however, was not at first that of a living monarch. Philip II and Alexander were content with their names on their coins, of which the obverses showed, for Philip, Apollo and Zeus and, for Alexander, Heracles and Athena. Alexander added the title basileus (king) only after



Alexander the Great as Zeus Ammon on a silver tetradrachm of Lysimachus, 297–281 BC; thought to be a copy of a portrait by Lysippus. In the British Museum. Diameter 30 mm.

Reproduced with permission of the trustees of the British Museum; photograph, Ray Gardner for The Hamlyn Publishing Group Limited

his Persian conquest. After his death his deified portrait appeared on the coins of Lysimachus in Thrace and on the early coins of Ptolemy I in Egypt. It was not until 306 that a living king put his own portrait on his coins, when Ptolemy I appeared, still as god, with the aegis of Zeus. Seleucus I similarly put himself on his coins as Dionysus; in time the divine attribute was dropped, and the ruler appeared as a mortal wearing only the royal diadem. In Macedonia, Arrhidaeus, Cassander, and Antigonus still followed the types of Alexander; and the early coins of Demetrius I Poliorcetes (336–283) were without a portrait. Soon, however, his own portrait appeared, still with the horns that deify him. His successor had only types of deities. Pyrrhus did not appear on any of his extensive coinages, but the last two kings of Macedonia, Perseus and Philip V, left very fine portraits. The kings of Pontus, notably Mithradates VI, had a magnificent series of portraits. The kings of Pergamum used the same portrait throughout, that of the founder of the dynasty, Philetairus I, and the Ptolemies in Egypt throughout their long series used only the head and legend of Ptolemy I, except on certain special issues. Among the early Seleucids, Antiochus I was reluctant to drop the portrait of Seleucus I, but the portrait of the reigning monarch became the rule.

After the vast issues of gold by Philip II, Alexander (under whom its price in relation to silver cheapened to 1:10 from 1:13 or more), and Lysimachus, gold was but rarely struck. Silver was the general metal of coinage; the Attic standard, which Alexander had adopted for his tetradrachms, became the monetary standard of the Western world, and there was a great increase in the bronze coinage. Egypt, however, kept to its own standards and to gold.

The Attic standard

As the greater part of the Greek world was now ruled by the Diadochi, their various coinages naturally formed the main currencies of commerce. Third-century Athenian coinages were scarce except in bronze. In 229, however, Macedonia lost its supremacy over Athens, and friendly relations were established between Athens and Rome. Shortly after 200 the abundant issue of tetradrachms of the "new style" began, which went on for slightly more than a century, replacing the "archaic" Athena with a copy of the head of the Parthenos of the Athenian sculptor Phidias, and with an owl on the reverse perched on a Panathenaic amphora. Corinth went on striking its stater until 229, when, with its surrender to Antigonus III Doson, king of Macedonia from 227, the long series came to an end.

*Rise of Rome.* After the Roman conquest of Greece it is clear from the resumed activity of the mints that the Greek cities were autonomous in one respect at least, for the silver coinage required in Greek territory could be supplied only by Greek mints, the task being beyond the power of Rome at this time. The Thessalians issued silver coins of the type of Zeus and Athena and the legend *Thessalon;* a similar coinage was issued by the Boeotians. Maronea and Thasos issued tetradrachms that became a great commercial currency for trade across the Danube with the Scythians and Celts who imitated them. Macedonia itself issued tetradrachms bearing the names of Roman governors. In Asia, after the defeat of Antiochus III at Magnesia, there was an outburst of tetradrachms of Attic weight and local types at towns such as Lampsacus,

Smyrna, and Magnesia. Other cities resumed the issue of Alexander tetradrachms, continuing to the middle of the 2nd century, when the Roman province of Asia was set up and cistophori replaced them. These, so called from the Dionysian chest (the sacred box or basket carried in the worship of Dionysus, usually shown containing snakes), which formed the principal type, were first struck at Pergamum after 228 BC; the reverse is a bow in a case between two serpents.

Coinage of Etruria

In the west the rise of Rome in the 3rd century introduced a new factor into the history of Greek coinage. The first coinage to disappear was that of Etruria—a silver issue curiously always left blank on one side—after a life of two centuries. Rome's early intercourse with the Greek cities of Italy is reflected in the Romano-Campanian coinage. In the south the Italian campaign of Pyrrhus left its mark on various coinages, notably at Tarentum. The towns of Magna Graecia gradually lost their silver coinage under Roman influence, although Greek bronze coins lasted until the 1st century at Paestum.

In Sicily in the 3rd century, derivatives of earlier Syracusan coinage began to dominate the whole island. The Punic Wars brought the Romans to Sicily, where the Carthaginians had been established since the end of the 5th century and had struck coins of Syracusan and other Sicilian types with Punic legends and later with their own types. Sicily became a Roman province; henceforth, only bronze was struck in it, and these local coins continued into the first century, when the last trace of Greek coinage in the west disappeared.

**Subsidiary Greek silver coinages under the Roman Empire.** Although Greek coins under the Roman Empire were nearly all of bronze and intended for local circulation, exceptional coinages in silver were allowed by Rome as a continuation, for wider regional use, of important preconquest currencies. The largest of these, running from Augustus to Diocletian's coinage reform, was minted at Alexandria to supply the needs of Egypt and was generally of billon (an alloy of silver and base metals). Inscriptions were in Greek and obverses bore the emperor's portrait, while reverses (dated in regnal years by Greek numerals) showed a wide variety of types embracing Hellenistic, Roman, and Egyptian symbolism.

Alexandrian coinages

In Syria silver tetradrachms continued to be struck, mainly at Antioch but also at Tyre and some other mints. These gradually became baser in the course of the early 3rd century. Bronze was also struck by the Romans at these mints and frequently bears the letters *S C* (*Senatus consulto*), showing, like similar issues at Rome, imperial initiative exerted through senatorial agency. Of several other local silver coinages the large series of drachmas struck at Caesarea in Cappadocia from Tiberius to Commodus is the most important. The most usual type was a local one of Mount Argaeus.

A number of vassal states and protectorates continued to issue their own coinages in the precious metals until they became Roman provinces. The only gold coinage of this kind is that of the kings of the Bosporus, who struck coins from the time of Augustus to the beginning of the 4th century. This coinage became gradually debased. In Africa the kings of Mauretania issued their own gold and silver until AD 40.

**Coinage in Judaea.** Another pre-imperial series continued under the Roman Empire was that of Judaea. Except for rare silver coins of much earlier date, with types of Greek origin but marked with brief Hebrew inscriptions, there were no Judaean issues until *c.* 135 BC; the Seleucid coinage of Syria had in the meantime supplied the necessary currency. Antiochus VII, however, had granted to the Hasmonean high priest Simon Maccabeus the right of coinage, which enabled the natural resistance of the Maccabees to Greek polytheism to be satisfied by the representation of specifically Jewish symbols. These coins, like those of the rest of the dynasty, were of copper. Alexander Jannaeus (103–76 BC) was the first of the Maccabean priestly princes to style himself king on his coins, which bore his name and title in Greek as well as Hebrew, but Pompey's withdrawal of the kingly title was reflected in the coins of John Hyrcanus II. Antigonus Mattathias (40–37 BC), the last of the Maccabees, introduced the seven-branched candlestick as a type. Under the Herodian dynasty, from 37 BC, Greek alone was found on Judaean coins. Herod Philip (4 BC–AD 34) gravely infringed Jewish convention by showing the effigy of the Roman emperor; Herod Agrippa I (41–44) was more adroit, avoiding the imperial portrait in Judaea but introducing his own in Caesarea.

Coinages of Jewish revolt

From AD 66, silver shekels and halves were coined, with some bronze, at "Jerusalem the Holy" to mark the first revolt against Rome: issues of year 5 (AD 70–71), a precarious one for the insurgents, are very rare. After the Flavian conquest, there were no further Jewish coins until the second revolt (132–135), under Bar Kokhba, when silver and bronze briefly proclaimed "the redemption of Israel and the freedom of Jerusalem." Jewish coinage ceased with the revolt's collapse.

Greco-Roman bronze

**Greek bronze imperial coinage, to AD 268.** Under the Roman Republic many Greek cities and districts continued to issue their own bronze coins, and, particularly in Asia, these local Greek coinages went on under the empire down to Gallienus.

The right of coinage in Greece was sometimes continuous and sometimes intermittently permitted by the emperor or governor. Coins were struck not only by single towns but also jointly by alliances of towns (*homonoiai*). The general type is everywhere the same: obverse, a bust and, reverse, a type of local interest. Under the republic the Greek cities usually placed on the obverses of their coins an allegorical bust of some local hero, the local city goddess, or a personification of the people, the municipal council, or the senate. The Tyche, the titular goddess of the city, appears as a female bust wearing a mural crown. The goddess Roma is found as a helmeted female; *e.g.*, at Smyrna. Under the empire the usual obverse type is the head of the emperor, as on the imperial series proper. There are some notable exceptions. Macedonia, for example, had the head of Alexander the Great. Athens was privileged by Hadrian to use the head of Athena in place of the emperor's.

Representations of statues

It is the reverse types of this series of coins that give them their importance. The coins of Athens preserve representations of many statues famous in antiquity that have long since perished, such as the Athena Parthenos of Phidias; the great Athena Promachos on the Acropolis, visible far out at sea; or the Dionysus of Alcamenes, possibly a pupil of Phidias. A coin of Elis preserves the Olympian Zeus of Phidias, and one of Lacedaemon the Apollo of Amyclae, near Sparta. Local cults and incidents in the lives of the Greek divinities are common types. Local celebrities are also recorded, for example, Homer at several of the various towns that claimed him as a native (notably Smyrna), Anacreon at Teos, Sappho at Eresus in Lesbos, Herodotus at Halicarnassus, and Alcaeus at Mytilene, which recorded on its coins a whole series of its famous men, most otherwise unknown. Reverse types also represent many architectural views of great importance, and the celebration of games and festivals is frequently recorded on coins.

In conclusion, mention may be made of a notable example of the preservation of a local tradition on a Greek imperial coin. On a coin of Septimius at Apameia in Phrygia there appear as reverse type a man and woman in a chest or ark floating on water, with a raven on the top and a dove flying above with a branch in its beak; to remove any doubt about the scene represented, the ark is labeled NΩ (NO; Noah), and the coin is evidence of the local tradition that the ark rested on the mountain behind Apameia.

## ROMAN COINS, REPUBLIC AND EMPIRE

**The beginnings.** Although Roman coinage soon diverged from Greek conventions, its origins were similar. Rome, founded in the 8th century BC, had no true coinage until the 3rd. Roman historians later attributed coinage unhesitatingly to the much earlier regal period: some derived *nummus* ("coin") from Numa Pompilius, by tradition Rome's second king, and Servius Tullius was credited with silver coinage, as well as with bronze stamped with the device of cattle. Roman historical tradition, however,

seriously confused the elements of the true picture. Rough, unworked lumps of bronze (aes rude) were certainly used as a metal currency from the 6th century, if not much earlier, perhaps in rare conjunction with very small quantities of unworked gold and silver, themselves also passing by weight. Simultaneously, standards of value appear to have been expressed in terms of cattle and sheep, as is clear not only from the derivation of *pecunia* ("money") from *pecus* ("cattle," or "sheep") but also from the early assessment of fines in oxen and sheep. From this it was falsely concluded that bronze coins marked with the device of cattle existed from the 6th century. In fact, the expression of values in terms of cattle may have lasted, officially, into the 5th century, for it was not until the decemvirs (a legislative commission) codified the law and drew up the Twelve Tables (451–449 BC) that fines were fixed in bronze. This bronze still consisted of unworked lumps or, at most, rough bars of irregular weight.

During the 4th century BC, Roman contact with the Greek cities of southern Italy slowly increased; these included such prolific mint cities as Nola, Hyria, and Naples. The coinages of these cities consisted of silver didrachms, of which Rome presumably made use in any necessary dealings with them. A hint is given, however, of widening Roman monetary interests by two issues of bronze token coinage. These, though certainly not produced at Rome, may perhaps be regarded as the earliest coins in the name of the Romans, struck at Naples *c.* 325–285 within the terms of their alliance and intended for use in Campania, as distinct from Rome and Latium. It is unlikely, indeed, that a mint in the proper sense existed at Rome before 289, the year to which Pomponius assigned the establishment of *tresviri* (a board of three officials) who should be *aeris flatores* ("bronze melters"); and this mint (in the temple of Juno Moneta) did not yet produce true coins but aes signatum, bronze bars (of about six pounds) lacking a mark of value but bearing on each side a clearly recognizable type (including cattle) and perhaps equivalent in value to a Greek silver didrachm.

Bronze token coinage

These aes signatum bars were halfway between aes rude and true coinage. In 269 true coinage appeared. It consisted of aes grave, large circular cast coins of bronze all bearing marks of value, from the as (weighing one pound) down to its 12th, the uncia; the obverses showed the head of a deity, the reverses a ship's prow. These were paralleled at mints elsewhere by similar cast coins; their types showed not, as at Rome, Latin deities but rather Greek (in the south) or Umbrian and Oscan. At the same time, there appeared struck silver didrachms, on the standard of the Greek silver coins of Campania, bearing Greek types but marked *ROMANO* or *ROMA*. Accompanied by small struck bronze token coins, these were issued from Campanian mints, and they probably continued to the Second Punic War, terminating in a new issue of silver coins of Roman style and types (marked *ROMA*), including Jupiter in a quadriga (four-horse chariot) from which their name, quadrigati, derived; they were imitated in electrum by the Carthaginians in Capua. The quadrigati were of the weight of the lighter Romano-Campanian didrachms and reflected the rising cost of silver at a time of stress; concurrently the cast bronze coinage of Rome dropped steadily in weight from an as of one pound to one of three ounces or less. Financial stress is similarly to be seen in the exceptional issue of gold units and halves. Toward the end of the Second Punic War the quadrigati were replaced by silver coins of half their weight, with a Victory on the reverse, and hence called victoriates. By *c.* 190 a mainly silver coinage, Latin-inscribed, was in production at Rome and other authorized mints, accompanied by bronze coinage so greatly reduced in standard (and thus size) that it could at last be struck instead of being cast.

**Introduction of the denarius.** Adjustment of the previously fluctuating relationship between bronze and silver was first secured by the issue *c.* 211 BC of the silver denarius (marked *X—i.e.,* 10 bronze asses), together with fractional coins, also of silver (marked *V—i.e.,* five; and *IIS—i.e.,* 2½ asses—a sesterce, or sestertius). The denarii were lighter than the quadrigati; their types were a Roma head on the obverse, with the Dioscuri (the twin deities Castor and Pollux) and *ROMA* on the reverse. Their production came to be confined principally to the mint of Rome. The victoriates, again lighter (their weight standard had come from Illyria), were issued until *c.* 150 BC, being perhaps intended for principal circulation outside Italy. The denarius, however, quickly established itself as the major currency in the central and western Mediterranean. In its eastward expansion, Rome learned to make use of local currencies—gold staters of Macedonia and silver tetradrachms of Athens or Asia. Rome was also prepared to employ Macedonian gold in the west, as was shown by the release to western markets of large quantities of gold staters after *c.* 150 BC. In the 2nd century BC, Roman coinage in gold was exceptional. Coinage in bronze, however, continued, but further variation in silver–bronze values was seen in two developments. The as dropped in weight to that of an uncia and then less, becoming a token currency; together with its fractions, it was now always struck and not cast. The value of the denarius in terms of bronze was altered, being revalued *c.* 133 at 16 instead of 10 asses; the silver quinarius (now of eight asses and with the types of the victoriate) became rare; and the silver sesterce (now equal to four asses) virtually disappeared. After *c.* 80 BC the striking of bronze was discontinued until the time of Caesar.

Employment of local currencies

These developments mirrored the economic difficulties of the day. Reduction of the weight of the as from one to ½ ounce in 89 BC was accompanied temporarily by debasement of the denarius, resulting in the issue of denarii with serrated edges, intended to show that they were not plated.

**Control and content of the coinage.** The coinage was controlled by the Senate, acting for the sovereign people; and the conduct of the mints was in the hands of boards of junior magistrates, the *tresviri.* From about the mid-2nd century, each of a mint's three *tresviri* normally issued coins bearing his own name, and on special occasions these were supplemented by issues of quaestors, curule aediles, prefects, or praetors; these were distinguished by special inscriptions such as *ex s(enatus) c(onsulto)* and *ex a(rgento) p(ublico).* The function of all these officials was quantity and quality control.

The moneyers' names at first were shown as simple monograms. The Dioscuri reverse was followed by Diana or Victory in a biga (two-horse chariot), and these again by figures of Jupiter, Juno, or Apollo in a quadriga, with the moneyers' names in fuller form. In the mid-2nd century BC, however, newer tendencies appeared, as when Sextus Pompeius Fostlus paired the Roma obverse with a reverse showing his traditional ancestor Faustulus discovering the wolf and twins; the reference was to the greatness of Rome, but it was to be seen through the lineage of a moneyer. Later republican denarii gave keen expression to party politics, as when corn ears recorded *c.* 100 the purchase of grain by the quaestors Piso and Caepio, or the head of Ceres with ploughing oxen proclaimed the program of the Marians, or Sextus Nonius Sufenas advertised the games that he had staged as praetor.

Steps toward imperial coinage

It was in the provinces, however, that the republican coinage took the decisive steps toward its finally imperial character. Campaigning generals began, in the 1st century BC, to operate mints for paying their troops in the field. In Italy mint policy had usually looked beyond personal politics to the state. But the military coinages of the *imperatores* equated the state with the personalities of the generals. Such were the aurei and denarii struck from eastern mints *c.* 82–81 by Sulla, with, obverse, *L. SVLLA* and head of Venus (his family patroness) and, reverse, *IMPER(ator) ITERVM,* priestly jug between trophies. Pompey issued comparable aurei *c.* 61 (also in the east). From these precedents the earlier coinage of Julius Caesar followed naturally in the late 50s and early 40s, with, obverse, *CAESAR* and elephant (the family badge) and, reverse, priestly symbols, or obverse, head of Venus (his traditional ancestress) and, reverse, *CAESAR,* Gaulish trophy, and captives. Such coinages still avoided the portraiture of a living man, the only examples of which hitherto had been on provincially struck coins.

**Caesar and after.** In the last year of his life, Caesar developed personal control of the coinage to a point at

which it lay ready to hand for Augustus to use later as a fully imperial instrument. Already, from 46 BC, coinage in gold had been instituted in Rome by Caesar's lieutenant Hirtius. Caesar's seizure of the treasury and his expansion of the annual board of moneyers from three to four members indicated his intention to deal absolutely with the coinage. In 44, denarii were issued in considerable quantity by his *quattuorviri,* bearing the portrait of Caesar on the obverse, with such inscriptions as *DICT(ator) QVART(um)* or *DICT(ator) PERPETVO,* and Venus Victrix or other semipersonal reverse types. For the token coinage a new alloy was now first struck—yellow orichalcum or brass, a copper–zinc alloy. Caesar may have enjoyed a monopoly of zinc from mines in Cisalpine Gaul.

From 44 to 31, bronze coinages were struck at various non-Italian mints, notably in or around Sicily, by officials attached to the cause of one or other of the members of the second triumvirate—Antony, Octavian, and Lepidus. But the principal issues of these years were of gold and silver. The mint of Rome continued its regular series until *c.* 37 and then ceased. Antony's coinage emanated at first from Gaul, then increasingly from eastern mints, including his cistophori and denarii (some showing his head conjoined with Cleopatra's) struck in Asia: his vast issue of often base denarii showing warships and military standards, shortly before the naval battle of Actium, was eastern. Octavian coined mainly in Gaul, Italy, and Africa. The piratical movement of Sextus Pompeius was reflected in the activity of a mint or mints in Sicily.

It was characteristic of most of the gold and silver after 44 that it showed portraits of the rival statesmen on the obverses, with reverses that alluded to their achievements or policies. This was true even of the "liberators" who murdered Caesar, for a famous eastern issue in the name of Brutus showed his portrait, with *BRVT(us) IMP(erator)* on the obverse, with reverse *EID(ibus) MAR(tiis)*—the fatal Ides of March—and daggers flanking a cap of liberty. By the close of the Roman Republic, three factors had entirely transformed the originally simple idiom of the early denarial coinage: gold was freely struck in addition to silver; the types of both were personal to military leaders and included living portraiture; and coinage could be produced elsewhere than at Rome.

*Early imperial mint policy.* Augustus (27 BC–AD 14) based the coinage on the aureus of $^1/_{42}$ of a pound of gold, equivalent to 25 denarii, each of $^1/_{84}$ of a pound of silver, the metals being struck almost pure. The denarius was valued at 16 asses. Token coinage consisted henceforth of brass sesterces and dupondii (equal to four and two asses, respectively), with copper asses, halves, and quarters, the as being the most common. Nero in AD 64 lightened aureus and denarius to $^1/_{45}$ and $^1/_{96}$, respectively, but debasement of silver subsequently took place. Under Septimius Severus it reached 40 percent, and Caracalla issued a debased double denarius of the weight of only 1½ denarii. Gallienus' double denarius of copper and silver, leached to give a more silver-rich surface, marked a monetary breakdown, only partially cured when Diocletian and Constantine again made gold the firm basis for supplementary pure silver and abundant copper coinage.

Early imperial mints

Augustus' earliest gold and silver were coined chiefly in the east—*e.g.,* at Ephesus and Pergamum—and more briefly at Emerita in Spain. Bronze also was mainly eastern, though some was struck at Nemausus (Nîmes). The Rome mint was reopened *c.* 20 BC for gold and silver and remained open for this purpose until *c.* 12 BC; its bronze continued irregularly. From 12 BC, Lugdunum (Lyon), with other mints of uncertain identity, undertook the main western coinages in gold, silver, and bronze. After 64 Rome was once more the chief mint for all metals. Official mintages were supplemented by a mass of regional or local coinages, while official coinages from eastern mints provided necessary currency for local Roman frontier forces.

The bronze of Rome was marked *S(enatus) C(onsulto)* and continued to bear the names of the *tresviri monetales*—masters of the mint, now reduced to their traditional number—until 4 BC. But *S C* also appeared on bronze from Lyon and Antioch in imperial provinces, showing that whatever nominal senatorial rights of coinage still lingered on—the *tresviri* are known until the 3rd century—the emperor wielded effective control over all metals everywhere. This was logical, since his economic powers were equally comprehensive. In fact, the old senatorial mint was transferred from the temple of Juno Moneta on Rome's Capitoline Hill and merged, probably after the fire of 64, with an imperial mint for gold and silver elsewhere in the capital. Henceforth, it worked in sections—six were normal later—controlled immediately by an imperial procurator and staffed by slaves or freedmen.



Portrait of Nero on a bronze sestertius struck at the mint in Lyon, Fr., *c.* AD 64–66. In the British Museum. Diameter 36 mm.

Reproduced with permission of the trustees of the British Museum; photograph, Ray Gardner for The Hamlyn Publishing Group Limited

*Portraits and types.* The use of Caesar's own portrait upon coinage set a precedent; although under Augustus and Tiberius token denominations occasionally lacked the imperial portrait, it was thereafter an essential element of virtually every gold, silver, and bronze coin of the official mints, as also of nearly all provincial and local coins. Emphasis on the personality of the emperor (extended sometimes to empresses, sons, or deceased members of the imperial house) was a powerful propaganda instrument in a coinage that circulated throughout a vast empire. The great series of imperial portraits, from Augustus to Romulus in AD 476, is artistically outstanding. Many of the finest appeared on the large brass sesterces down to the 3rd century and on the even larger bronze medallions produced for presentation; but particular care was taken over the portraits for gold, which, being softer, showed a beautiful and highly sensitive impression. Nothing is known of the portrait artists, though it is likely that they were often from the Greek East.

Imperial reverse types

Imperial reverse types, if artistically less remarkable, are uniquely important for the unparalleled fullness of the historical commentary that they supply. The major mints provided annual evidence of imperial interests: victories in war; frontier defense (*e.g., Rex Parthis datus*—"A king is given to the Parthians"—of Trajan); a well-earned peace (*e.g.,* the *Pace terra marique parta Ianum clusit*—"There being peace on land and sea, the doors of the Temple of Janus were closed"—of Nero); the birth of an heir or alternative provision for the succession; public shows; acts of social reform or public relief (*e.g., Civitatibus Asiae restitutis*—"For the restitution of the citizenries of Asia"); imperial journeys (*e.g., Adventus Augusti*—"The arrival of the emperor"); and religious or other anniversaries (*e.g.,* the *Felix temporum reparatio*—"Happy days are here again"—on Rome's 1,100th birthday). Their interpretation demands care, since, being selected by imperial officials, their tenor can conflict with the attitude of anti-imperial historians. But they show the efforts made by emperors, as the omnipotent semireligious heads of a huge and heterogeneous empire, to conciliate and inform. They contributed powerfully to the growing conception of an eternal Roman empire, seen no less in the special types of eagle (the soul flown heavenward) or funeral pyre or temple in honour of "good" emperors consecrated as *divi* than in the annual record of military victory, economic security, and provincial peace and implicit in the regularity of imperial succession. The normal colour given to this imperial program was religious, for the coinage types commonly embraced such characteristically Roman concepts as Aequitas (Justice), Fides (Faith), and Concor-

dia (Harmony)—social virtues operating in the guise of minor deities.

**The 4th century and after.** Diocletian's institution of the tetrarchy, by which the empire was divided administratively between two Augusti and two Caesars, brought fundamental changes in social and economic policy; the instability of prices called for complete renewal of the monetary system. His coinage reforms took place in stages from *c.* 286 to *c.* 296. First, new aurei were struck at 60 to the pound of gold. Then, *c.* 293–294, new silver coins, of good purity, were struck at the revised Neronian weight of 96 to the pound of silver. Finally, *c.* 294–296, new copper coins appeared that were larger and intrinsically more valuable than the small debased double denarii of previous reigns. The contemporary names of these silver and copper pieces are not known. This reformed coinage was struck at a variety of mints from Londinium (London) to Alexandria, most of which coined in all three metals. Types were closely controlled in the silver and copper coinage; in the latter the almost universal type was for some years that of the "Genius Populi Romani." The obverse bore the portrait of one or other of the tetrarchs, each of whom coined with portraits of all four.

The breakdown of the tetrarchy after 306 weakened the new system. Copper was quickly and steadily lightened, and silver struck very sparingly. Gold, however, continued in good supply; and though Constantine's solidus showed a reduced weight standard, there was no shortage of gold throughout the rest of the 4th century. In time, silver coinage increased, especially after *c.* 350, when the miliarense (1/1,000 of a gold pound) and smaller denominations appeared. By the end of the 4th century, however, the size of copper coins had dropped very sharply, and in the 5th, until the Western Empire collapsed in AD 476, the western coinage consisted finally of gold with a little silver, struck mainly from the mints of Ravenna and Rome.

Coinage under Constantine

From 312, when Constantine became emperor of the West, coin types began to show new tendencies. The imperial portrait was still the dominant feature. Reverses displayed complementary themes—the glory of the army, vows for continued imperial rule, the constant struggle against barbarian pressure on the frontiers. The old variety of pagan gods—Jupiter excepted—mainly disappeared, though Sol, popular from Aurelian onward, was used, especially by Constantine. Christian emblems did little to take their plaee, though the Christian monogram, the Greek letters chi and rho superimposed, sometimes on a standard, began to appear with Constantine and was combined with the alpha and omega under Constantius II and Magnentius. On the whole, however, there was an unavowed truce between Christianity and paganism, only occasionally broken, as when Julian revived a range of pagan types; the full development of the Christian tradition in coinage was reserved for Byzantium.

## COINAGE IN WESTERN CONTINENTAL EUROPE, AFRICA, AND THE BYZANTINE EMPIRE (5TH–10TH CENTURY)

The fall of Roman power in the West left the gold currency of the Byzantine Empire undisturbed; it was to become the most dominant single influence in European coinage for 1,000 years, competing at first with the gold of the Arab caliphates and later with that of the great Italian commercial republics as well. Byzantine coinage, in its continuity, contrasted strongly with the often erratic monetary systems from the 5th to the 7th century in western Europe, where Germanic invaders inherited the apparatus, money included, of the Roman Empire. In general, they took over the main features of late Roman coinage. Emphasis on gold continued, with silver and some bronze; gold chiefly served for the triens, or third (1/3 of the Constantinian solidus). The types of the gold coins for some time reflected Byzantine prestige, showing a formalized portrait obverse and titles of the reigning Byzantine emperor, toward whom widespread respect was paid even when Western kings began to add their personal monograms to the normal Victory reverses. Imperial prerogative, so powerful an influence on western gold, had less effect on silver, the types of which in the West became more flexible; in bronze, where obvious efforts were at times made to link with traditional Roman design, flexibility was greater still. In technique these coinages varied widely: that of Italy was not without elegance; that of Spain developed an elaborately stylized balance, depending largely on its bold letter forms; the highly abstracted figures of Gallic coins have found great favour among 20th-century artists, while those of Africa and Britain were in general considered artistically inferior. The weights of gold coinages were kept at a reasonably steady level, though fineness ultimately declined with the economic decline of the issuing kingdoms themselves.

**Post-Roman coinage in the West.** In Italy Odoacer (476–493) had coined in silver and bronze at Ravenna after setting up a Teutonic kingdom. The Ostrogothic coinage that followed, from Theodoric (493–526) onward, consisted of gold, mainly imitating current Byzantine issues and with the imperial portrait (Theodoric's fine portrait on a unique triple solidus is wholly exceptional). Silver and bronze were supplementary. The Lombards of Italy (568–774) had no distinctive coinage of their own until the gold struck in the name of Grimoald, duke of Beneventum (662–671), which was followed by gold and silver from a number of mints elsewhere. In Africa the Vandal kings Gunthamund (484–496) and Hilderic (523–?530) issued silver and bronze coinage, respectively, inscribed with their names; the types and denominations looked to imperial models and, in the case of the bronze, to those of Carthage especially. Vandal gold was perhaps struck by Gaiseric (428–477) or Huneric (477–484) in the Byzantine emperor's name, but in the absence of any royal monogram it cannot easily be attributed. The chief Spanish coinage was that of the Visigoths, who controlled southern Gaul also and—after Leovigild (568–586)—Suevia (modern Galicia), with its rich gold mines; hence the fact that of 79 Visigothic mints a high proportion was concentrated in northwestern Spain. Visigothic gold coinage was produced up to the Arab invasion in the 8th century and consisted almost entirely of thirds, at first imitating Byzantine models, and bearing kings' names and titles. The most prolific mints were Mérida, Toledo, Seville, Tarragona, and Córdoba.

Ostrogothic coinage



Gold triple solidus of Theodoric from the mint at Rome, *c.* AD 500. In the Bibliothèque Nationale, Paris. Diameter 34 mm.

By courtesy of the Bibliotheque Nationale, Paris

In Gaul the Burgundians struck their own imitative gold thirds, first, under Gundobad (473–516), inscribed with a royal monogram, though not yet displacing the imperial name and portrait. The largest of the Gaulish coinages, however, was that of the Merovingian Franks, beginning with Clovis I (481–511). The gold consisted mainly of thirds, at first with some subsidiary silver and copper, inscribed by Theodoric I (511–533/534) and Childebert I of Paris (511–558) with their own names. As elsewhere, the types of the gold borrowed steadily from the imperial series, either the former Roman or the current Byzantine. Reverses showed a Victory, though the theme of the "cross on steps" of Tiberius II (578–582) gradually displaced it, beginning in the south. Obverses generally showed a profile, and later sometimes a frontal, bust. A profound break with tradition came when Theodebert I (533/534–547/548) substituted his own name on his gold for that of the Byzantine emperor—a change that in turn was to influence Visigothic gold. The right of striking gold had meanwhile been widely extended, to mints presumably

Borrowings from imperial types

operated by royal permission and numbering nearly 500 in all. These were distributed over an area including not only what is now France but also the Low Countries, the Rhineland, and Switzerland. The types of Merovingian gold coins diverged increasingly from imperial models: nearly all of them were inscribed on the obverse with the name of the issuing authority, most often municipal, and on the reverse with that of the moneyer. As the Merovingian dynasty drew to a close in the 8th century, gold coinage became poorer in quality, and it gave way to the small silver denarius, of about 1.2 grams, struck in quantity. This change heralded the Carolingian revival of the denarius.

Coinage supply to Britain was interrupted when the mints of Roman Gaul were closed *c.* 395, and scarcely any gold or silver coin entered Britain during *c.* 450–550. Subsequent penetration of Merovingian gold encouraged a brief Anglo-Saxon coinage of gold thirds (see below).

**Coinage in the Byzantine Empire.** Inspiring many features of these transient coinages, but outliving them all, stood the currency of the Byzantine Empire. It was based on the gold solidus (1/72 of a pound) of Constantine—the bezant of 4.5 grams (about 70 grains) maximum, which dominated so much of European trade to the 13th century. Until the 10th century, halves and thirds were also used. This gold was proverbial for its purity until the 10th century. The fundamentally religious nature of the empire was fully reflected in the coinage: throughout 10 centuries there was scarcely a single issue that did not look directly to the Christian faith, since apart from reverse types and legends, which were purely religious, the obverses showed the emperors as specifically Christian rulers by the use of adjuncts or appropriate inscriptions.

Byzantine coinage began effectively with the reign (491–518) of Anastasius I. Thenceforth, it consisted, in addition to gold, of silver and bronze. Silver, always rather rare, consisted of the small siliqua (1/24 of a solidus) or keration, followed by the larger miliaresion and the still larger hexagram. Bronze was in most periods very commonly struck. Its appearance and tariffing were reformed by Anastasius, who issued large pieces marked *M, K, I,* and *E* (equal to 40, 20, 10, and five *nummi*); other multiples were found either later or locally, as *IB* (equal to 12 *nummi*) at Alexandria. Such marks of value continued until Basil I (867–886). Constantinople itself was the main mint in all three metals, which were coined also at Carthage and Ravenna. Thessalonica, Nicomedia, Cyzicus, Antioch, and Alexandria struck bronze only; at one time or another Rome struck gold and bronze, while Syracuse and Catana also contributed. The technique of gold and silver minting was generally high; that of bronze was coarse, and overstriking was common.

Major Byzantine mints

*Types and legends of Byzantine coins.* For gold, the earliest obverses were diademed profile busts or helmeted facing busts, both common on previous coins of eastern and western empires. The facing bust showed the emperor in military panoply with a cross in his hand or on his helmet, and, if the cross was lacking on the obverse, it appeared on the reverse. With Justin I (518–527) and Justinian I (527–565), the seated figures of the emperors were shown side by side (527). Thereafter, the facing head became more common: from the time of Phocas (602–610) it was increasingly formalized, a process that reached its climax in the 8th century. Under Heraclius (610–641) the habit began of showing the emperor with one or more of his sons; and, with figure types now more common, it was possible to show emperor and empress together or even, as with John I Tzimisces (969–976), the emperor being crowned by the Virgin, with the hand of God above. The reverses of the gold coins at first emphasized the Victory (doubtless regarded as an angel) of previous issues. Tiberius II introduced the cross potent on steps, a type destined to play a long and important part. Justinian II (685–711) was the first to use the haloed bust of Christ, who had previously been shown only on a coin of *c.* 450, in the act of marrying the empress Pulcheria to Marcian. The Iconoclasm of Leo III (717–741) and his successors banished such divine representations in favour either of the cross on steps or of imperial figures on the reverses, but with Michael III (842–867) the bust of Christ returned. From Basil I the throned Christ predominated.

The obverses of the silver coins, beginning with profile busts, thereafter included seated figures, facing busts, and purely epigraphic designs. The introduction of the larger hexagram by Heraclius in 615 allowed fuller scope for later designers, whose reverses often consisted of a cross on steps or a bust of Christ surrounded by inscriptions; from the 10th century the cross bore a central portrait medallion of the emperor himself.

In bronze coinage there was at first less flexibility. The earliest types were, obverse, a profile bust and, reverse, a cross and mark of value. Under Justinian I the facing bust prevailed, and in his 12th year he introduced the dating of his bronze coins on the reverse, in the form *Anno XII;* the inclusion of a regnal date was thereafter normal on bronze until Constans II (641–668). From the time of Justin II (565–578) the obverses showed two or more standing imperial figures combined (until Basil I) with the mark of value. From the 10th century the reverses were taken up wholly by three or four lines of inscription; and the anonymous bronze coins of John Tzimisces combined such a reverse, reading *Iesus Christus Basileu(s) Basile(on),* with a new obverse showing the facing bust of Christ designated *Emmanuel.*

Byzantine bronze types

The orthography of Byzantine coin legends became remarkably complex as the Latin and Greek alphabets were increasingly mingled and individual letters took on new or specialized forms and words were severely abbreviated. At first the inscriptions were purely Latin, the emperor's names and titles being in the conventional form *D(ominus) N(oster)—P(ius) F(elix) Aug(ustus).* Even before Anastasius, however, *Perpetuus* had been a variant for *P.F.*, and, abbreviated in the form *PP,* it finally prevailed. In the 7th century, Greek letters were more commonly mixed with the Latin in such legends as that of Justinian II, when he styled himself *Servus Christi;* and in the later 8th, the general shift to Greek from Latin conceptions was plain in the emperor's new title of *Basileus.* Comparatively long votive inscriptions, as "Lord, help thy servant," and metrical inscriptions (a practice more common in Asia than in Europe) began in the 10th century.

*Economic role of Byzantine coins.* Byzantine gold coinage, until its debasement from the 10th century, was immensely important in the economic life of the Levant and western Mediterranean. The total output of gold was great, and its influence can be judged partly from the distribution of the coins themselves and partly by the typological influence exerted by the Byzantine upon other coinages, from the first Arab-Sāsānian gold of the East to that of Italy and Gaul in the West. In the 5th and 6th centuries, Byzantine solidi accumulated in the Baltic area, doubtless in payment for furs; and, in the 6th and 7th, solidi of a slightly lighter weight were hoarded in France, the Low Countries, Scandinavia, Germany, the Balkans, Russia, the Levant, and northern Africa. In these last two regions Byzantine gold competed from the 7th century with the increasing output of Arab gold dinars.

**Charlemagne and the Carolingian coinages.** While the bezant and dinar maintained gold currency along the Mediterranean, northern Europe from the 8th century suffered a shortage of gold and turned its almost exclusive attention to silver, inherently more convenient as a unit of exchange. A previous Merovingian tendency to introduce silver alongside gold was carried much further when the Carolingian ruler Pepin the Short (751–768) replaced gold by silver, introducing the denier, which was to be the basis of all medieval coinage in the north. His new coin was wider and thinner than previous silver pieces. The normal types were simple—obverse *R P* (for *Rex Pepinus*), reverse *R F* (for *Rex Francorum*).

Coinage of Pepin the Short

Charlemagne (768–814) reorganized northern currency in a way that affected it permanently. Coining at first simply as *Carolus R F,* he defeated the Lombards in 774 and entered Rome, becoming king of Lombardy as well. His deniers were later made wider and still heavier (about 25 grains), and he introduced the smaller and subsidiary obole, or half-denier. The main types of his deniers were threefold: the monogram of his Latinized name, Carolus;

a temple (sometimes a gateway); and, more rarely, a portrait. Monogram deniers were coined in France, Germany, northern Italy, and northeastern Spain; temple deniers were also widely struck, often inscribed *XRISTIANA RELIGIO,* though this legend was sometimes replaced by the name of a major French mint city. On Christmas Day 800, Pope Leo III crowned Charlemagne as Roman emperor, and, thenceforth, his deniers, either with the temple type and "Christian" legend or with a mint name alone, styled him *Kar(o)lus Imp. Aug.,* sometimes adding *Rex F(rancorum) et L(angobardorum).* His mints lay mainly in France, the Rhineland, and the Low Countries.

Louis I (814–840) continued his father's monetary system with little essential change. But the infringement of his minting rights emphasized the economic importance of northern ports, especially the Frisian Duurstede, from the neighbourhood of which emanated large numbers of copies of his gold sous and half-sous. These portrait coins originally were designed presumably for presentation to the Holy See, since the reverse bore the inscription *MVNVS DIVINVM* around a cross. They were struck sparingly, and no Carolingian gold thereafter appeared. Charlemagne's pattern of coinage, sometimes varied, was extended to Lotharingia, with such powerful mints as Cologne, Metz, Trier, and Strasbourg. From the time of the French kings Louis II and III (877–882) the Carolingian currency pattern weakened, and feudal coinages made their first appearances. Louis IV d'Outremer granted coinage rights to the archbishop of Reims as early as 850, and the system was swiftly developed in the 10th century, concessions being made to a large number of ecclesiastical foundations and even in a few cases to lay lords as well. In Spain, Carolingian mints were established only in the extreme northeast, at Barcelona, Ampurias, and Gerona. The kingdom of Aquitaine, under Charlemagne, was reserved to the Frankish king's son, and its coins were modeled on the Carolingian pattern. Northern Italy was an integral part of the territories controlled by the earlier Carolingians, but from the mid-9th century changes began to show: the deniers of Pavia and Milan, though retaining Carolingian types, became broader and thinner, with wide rims like those of the later German bracteates (see below). Venice, a republic from the late 7th century, ruled by a doge under Byzantine protection, did not coin until the 9th, when it struck deniers for the Carolingians; but after Lothair I it omitted mention of the imperial name.

Papal coinage

At Rome papal coinage began with Adrian I (772–795), Byzantine in style and types, but after Charlemagne's visit in 774 all deniers (except during an imperial interregnum) were struck jointly with the pope's monogram and the emperor's name, until 904; thenceforth, the papal name appeared in full and alone. The principalities of Beneventum and Salerno and the duchies of Naples and Amalfi fell within the Byzantine–Arab orbit, and their gold, silver, and bronze showed these beside Carolingian influences; bronze coins in particular followed Byzantine models, while the gold tari of Salerno were curious fractional copies of Arab dinars. In central Europe, Carolingian coinage was not reflected east of the Rhineland, but in the north the imitation of Carolingian money in or around Duurstede bred more distantly derivative issues elsewhere, possibly even in Scandinavia; these were, in effect, silver deniers, but their types, with their emphasis on ships and animal designs, show them to belong to the Nordic, as opposed to the Teutonic, stream of monetary design.

### THE LATER MEDIEVAL AND MODERN COINAGES OF CONTINENTAL EUROPE

The change of power from Frankish to German emperors in the 10th century saw the silver denier extended into central and northern Europe. In the East the decay of the Byzantine Empire was reflected in the debasement of its gold coinage to electrum; after the temporary fall of Constantinople to Western crusaders in 1204, Byzantine tradition was carried on in the silver coinages of the derivative empires of Trebizond, Nicaea, and elsewhere. The revival of gold coinage in Italy in the 13th and 14th centuries, promptly copied elsewhere, led to the need for a silver denomination larger than the denier, and the grosso and its equivalents soon spread widely. From the 14th century coinage began to lose its Gothic stiffness: the Italian Renaissance pointed the way to naturalism in portraiture and to greater fluency of ornament. In the 15th century the first experiments were made with mechanical methods of coining, and by the 16th the new techniques were being generally adopted (see below *Techniques of production*). The traditionally privileged nonregal mints were incapable of producing the mechanical power needed for the intensive coinage not only of the large gold denominations resulting from the influx of Spanish-American treasure after 1493 but also with the equally large silver thalers, or dollars, beginning to be produced with silver from the German Joachimsthal mines. Multiplication of gold and silver coinages, and their larger denominational values, emphasized the need for token coinages, which were produced from the 17th century. Britain was effectively on the gold standard from the end of the 18th century, together with Portugal, but it was not until the second half of the 19th that continental Europe followed suit. Paper currencies of this period were fully redeemable in gold coin, but the gold standard was abandoned during World War I; since then, paper has been redeemable effectively only in base-metal alloys.

Renaissance developments

The coin types of the later medieval period were relatively crude. Portraiture, schematically stiff on later Byzantine money, was revived with striking realism most notably in Renaissance Italy and thereafter flourished. Reverses revealed feudal influence in shields of arms and civic emblems. These developments set the general pattern of modern coinage, usually with an obverse portrait and some form of national badge or arms on the reverse. From *c.* 1800 onward this pattern was standardized to a large degree.

**Portugal.** Coinage began in Portugal, after the expulsion of the Moors, with Afonso I (1128–85), whose gold maravedis, copied from the gold of the Berber Almoravids, retained certain Arab features in design. Some base silver was also struck. Rights of coinage were, from the start, reserved to the kings, almost exclusively. Peter I (1357–67) reformed the coinage on the basis of the gold dobra of about 4.9 grams, with types copied from those of contemporary France: obverse, king enthroned; reverse, ornamental cross. There was a similarly imitative silver gros tournois (based on the weight standard of Tours, Fr.). Peter's successors developed his system. Copper was struck from the 15th century. From the 16th to the 18th century, gold was coined in quantity and in denominations of handsome size down to the half-escudo. In the 19th century the basic gold denomination was the crown. In the 20th century token denominations (in terms of centavos) have prevailed in various alloys, though silver was introduced in 1954 for the 10-escudo piece and for certain purely commemorative issues.

**Spain.** As in Portugal, the coinage struck after the expulsion of the Moors was almost without exception regal. That of Navarre started under Sancho III Garcés (*c.* 1000–35) with deniers of Carolingian influence. The series of Castile and León began with similar pieces under Alfonso VI (1065–1109), and that of Aragon under Sancho Ramírez (1063–94). Among the earliest gold was that of Alfonso VIII of Castile (1158–1214), copying an Arab gold dinar but with Christian professions in its Arabic script. Gold portrait doblas appeared under Sancho IV of Castile and León in the 13th century, and the portraiture under Pedro I in the 14th was of high quality. Gold coinage multiplied in the 15th century, with Henry IV coining huge pieces of superb Gothic style; silver and billon were also in good supply. The union of the crowns of Castile and Aragon in 1479, and subsequently the influx of American precious metals, resulted in an abundant coinage in gold (the excelente and its multiples) and silver (the real and its multiples)—the silver piece of eight being the famous Spanish dollar. This last denomination enjoyed enormously wide currency, and its type (obverse, royal portrait; reverse, Pillars of Hercules with *PLVS VLTRA* on scroll) was universally known.

Influx of American precious metals

**France.** The dynasty of Hugh Capet (987–996) made no immediate change in the previous Carolingian coinage

system: deniers and their halves, the oboles, continued, but tended to decline in fineness. Feudal deniers began to appear in abundance beyond Capet's kingdom of north central France; the most important and numerous were issued from the 10th century by the abbey of St. Martin at Tours, with a "castle" type destined to exert wide influence. This monnaie tournois was lighter than the royal monnaie parisis (based on the Paris weight standard), generally in the ratio 4:5. Louis IX in and after 1262 reformed the coinage. The sou became in 1266 the silver gros tournois, $^{23}/_{24}$ fine and weighing about four grams; its types continued the "castle" of the denier tournois but with concentric inscription and ornament frequently imitated. With this there appeared a gold écu, with the royal lilies on a shield. Subsequent development down to the 15th century emphasized more and larger gold denominations; silver continued, often debased. Design reached magnificent heights of Gothic splendour, seen in the *masse d'or* ("sceptre of gold"), *mouton d'or* ("Paschal Lamb"), *ange d'or* ("angel of gold"), and *franc d'or* (franc ["free"], a term first applied to a coin of John II, minted in 1360 to commemorate his ransom from the English). The Anglo-Gallic issues of the time were comparably beautiful. Feudal coinage was severely limited, that of Brittany and (at



Louis XIII silver écu blanc (louis d'argent), Paris, 1643; the dies for the coin were engraved by Jean Warin. In the British Museum. Diameter 44 mm.

Reproduced with permission of the trustees of the British Museum; photograph, Ray Gardner for The Hamlyn Publishing Group Limited

first) Aquitaine being most important. Types in Aquitaine later showed some English influence, while in the gold of Provence that of the Florentine florin was noticeable. In the 16th century broad, thick silver coins were adopted, familiarized by the testons (from *testone,* which means "head") of Italy; these, together with the gold écus, set the general pattern. Early in the 17th century the use of machinery for coining was the subject of experiments by Nicolas Briot; both he and Jean Warin were famous for their technique and style under Louis XIII. The late 17th and 18th centuries, though their coinage was of considerable external magnificence, were not devoid of monetary difficulty. Louis XV suppressed independent local minting, Strasbourg being the last provincial mint to survive, though royal branch mints continued. Under the Revolution Louis XVI coined first as constitutional king, in gold, silver, and copper; but from 1793 the issues were wholly republican, with the inscriptions *République français, Liberté,* etc., and the symbols (cap of liberty, cock) that have survived in modern French coinage. The precious metals were in short supply; gold and silver were demonetized and paper took their place, together with copper. In 1793 the decimal system was adopted, in terms of francs, decimes, and centimes, coins now being dated by the Revolutionary era; gold coinage was effectively lacking until the time of Napoleon. From 1866 France was joined with Belgium, Italy, and Switzerland in a monetary convention defining the denominations, quality, and weight of gold and silver coinage in terms of francs. In the 20th century alloys were introduced, and the Vichy government of Henri-Philippe Pétain also used zinc, iron, and aluminum. From 1950 paper money was increasingly replaced by alloy coins, the "heavy" revalued franc being introduced in 1959.

20th-century materials

**The Low Countries.** In Merovingian and Carolingian



The gold vieil heaume of Louis de Mâle, count of Flanders, 1367, one of the largest Flemish gold coins. In the Fitzwilliam Museum, Cambridge. Diameter 35 mm.

By courtesy of the Fitzwilliam Museum, Cambridge, Collection of Professor P. Grierson

periods a few mints operated in the Low Countries. Subsequently the area was divided among a number of dukes, counts, seigneurs, and ecclesiastics. In the 16th century the Low Countries passed to the House of Austria, and the daalder (dollar) appeared. English military operations were accompanied by the issue of gold pieces. The 16th century produced some remarkable siege pieces from Amsterdam, Bergen op Zoom, and elsewhere. With the establishment of the Kingdom of Holland under Louis Napoleon in 1806, coinage began to conform with that of the Napoleonic Empire. Belgium emerged as an independent kingdom in 1831, and in 1860 adopted a cupronickel alloy for its French- or Flemish-inscribed or bilingual coinage.

**Switzerland.** The coinage of Switzerland illustrates its varying fortunes. First there was the gold money of the Merovingian kings, among whose mints were Basel, Lausanne, St. Maurice-en-Valais, and Sitten (Sion). The silver deniers that Charlemagne made the coinage of the empire were issued by fewer mints. The dukes of Swabia began to strike at Zürich in the 10th century, and the empire from the 10th to the 13th century granted the right of coinage to various ecclesiastical foundations. Bern was allowed a mint by the emperor Frederick II in 1218, and other towns and seigneurs subsequently gained the same right. The demi-bracteate appeared about the middle of the 11th century, and about 1125 it was superseded by the true bracteate, which lasted until about 1300. (Bracteates were lightweight silver coins so thin that they bore only a single type, repoussé [hammered into relief on the reverse], for which a special technique [including the use of wooden dies] was devised.) The Swiss Confederation developed in the 14th century, and by degrees the cantons struck their own money. These, together with the coins of some few sees and abbacies, formed the bulk of Swiss money of the medieval and modern periods. The cantonal coinage, interrupted by the French occupation, was suppressed in 1848, when a uniform currency was adopted.

**Italy and Sicily.** At the close of the Carolingian period the coinage of Italy fell into two main classes. In nearly all of the north, including Rome, it consisted of silver deniers of Carolingian derivation, mainly struck at Pavia, Milan, Lucca, and Verona. At Venice and over most of the south the dual influences of the Byzantine and Arab empires were prominent. Monetary fashions were shown in the coinage of Sicily struck by the Normans. Robert Guiscard in 1075–85 struck small gold coins called taris of almost wholly Arabic appearance, together with bronze of Byzantine style. Roger I of Sicily Latinized the bronze, and Roger II coined silver ducats of Byzantine type; Arab-style gold taris still continued for commercial reasons, since the great Fāṭimid coinage was then the currency of all western Muslims. After southern Italy and Sicily had fallen to German power, Frederick II (1212–50) restored a Latin coinage of gold, of splendid style and execution and good fineness, in proto-Renaissance style. His gold augustale (patterned after the aureus) and their halves, struck *c.* 1231 at Brindisi and Messina, were accompanied by billon deniers. Sicily soon passed to Charles I of Anjou (1266–85), and its Angevin coinage, like that of Naples,

assumed the French medieval style, succeeded in turn by that of the Aragonese kings.

Coinage of the northern Italian cities

In northern Italy leading cities were issuing silver with a free choice of types—portraits, badges, or figures of patron saints and others, with explanatory legends. Mantua celebrated Virgil; Florence from *c.* 1189 showed its lily with St. John the Baptist; and Genoa chose the *janua,* or eponymous gate. Venice, abandoning the imperial name early in the 12th century, set a precedent *c.* 1192 in the issue of the larger silver grosso or matapan, using the henceforth familiar types of Christ on the reverse and, obverse, St. Mark presenting the gonfalon (the banner of the republic) to the doge. The influence of the gold coinage of Frederick II on such cities was soon evident. Genoa was striking gold as early as 1252. Florence issued the first of its famous and profuse series of fiorini d'oro, or gold florins. The lily continued as the civic type, together with the standing figure of the Baptist. Regular weight (about 3.50 grams, 54 grains) and fineness won the fiorino universal fame and wide imitation; double florins were introduced in 1504. Venice in 1284 produced its gold ducat, or zecchino (sequin), of the same weight. Venetian ducats rivaled Florentine florins in commercial influence and were widely copied abroad. The series begun under Giovanni Dandolo continued with the names of the successive doges until the early 19th century.

At Rome no papal coins appeared from 984 until purely epigraphic types recorded the names of Leo IX and the emperor Henry III in 1049–54. Thereafter, there was a further gap until Urban V (1362–70). The Senate of Rome coined silver deniers from 1188, with the antique legend *Senatus Populus Q.R.* and figures of SS. Peter and Paul. In 1252 Brancaleone struck deniers with the seated figure of Rome and the legend *Roma Caput Mundi;* Charles of Anjou in the 13th century and Cola di Rienzo in the 14th also coined, as Roman senator and tribune, respectively. Senatorial gold ducats were introduced on the Venetian model in 1350. Papal coinage returned from Avignon in 1367 with Urban V, who assumed rights over the mint of Rome; gold, silver, and bronze later developed, with types (crossed keys, tiara, personal arms, and many different emblems) that, with few interruptions, have lasted ever since. Since 1869 papal coinage has been mainly of a commemorative nature, in silver, acmonital (stainless steel), and bronze, of denominations corresponding with the Italian state coinage.

Design under patronage

The patronage given by the popes to notable artists—*e.g.,* Francia and Benvenuto Cellini—resulted in a fine and often lavish standard of design in their coins and medals. Similar patronage was shown by the noble houses of Ferrara, Mantua, Milan, and elsewhere, whose coinages from the 15th century attained a splendid level. The size of gold and silver denominations was growing, as witness the silver teston of 1472; and the portraits made by Caradosso, Francia, and others of equal fame are among the finest small-scale Renaissance works. Later coins of still larger size of the duchies of Savoy and Florence are remarkable. Italian coinage continued to be divided among a number of kingdoms, principalities, and duchies until 1861, when Victor Emmanuel I first coined as king of all Italy. The metals were gold, silver, and bronze; alloys were introduced under Umberto I (1878–1900). Under Victor Emmanuel III (1900–46) reverse types borrowed heavily from the antique, and his later issues reflected the influence of the Fascist regime, being dated by the Fascist era from 1936 (year XIV) as well as by the Christian. From that same year, he appeared as emperor (of Ethiopia) as well as king of Italy, and after 1939 coins were struck for him, with a helmeted portrait, as king of Albania. After World War II the republican coinage of Italy, in aluminum and steel, concentrated mainly on symbols of agricultural fertility and national industry.

**Germany and central Europe.** Territorially, the German issues began and developed in an area that has since been many times divided and from which Austria, Hungary, Czechoslovakia, and Yugoslavia have emerged as separate states. Classification of these issues remains one of the most formidable numismatic problems.

From the 10th to the 12th century the Carolingian pattern of coinage was continued; but with the advent of the Swabians under Conrad III in 1138, unity disappeared. In the west the silver denier continued. In the east the coinage of very thin bracteates was developed. The western deniers were in part from imperial mints, scattered among a much larger number of feudal mints, representing ecclesiastical rather than lay authorities. Westphalia produced a profuse ecclesiastical coinage. That of Cologne was especially important, showing the former Carolingian "temple" combined with the linear inscription *S(ancta) Colonia A(grippinensis)*; and that of Münster was comparable in influence. This area was conservative and prosperous; the weight of its deniers was well maintained, and, although Anglo-Saxon and, later, English and Byzantine influences became noticeable, its types changed but little.

Westphalian ecclesiastical coinage

In the eastern region a sharp decline in weight led to the thin, single-type bracteates, and the designs quickly broke away from Carolingian tradition. Issued by a wide variety of authorities, many of them ecclesiastical, these coins showed a great range of human figures and portraits (saintly and secular) together with representations of churches, castles, and heraldic devices in an essentially medieval Germanic style. The difference between the heavier western deniers and the lighter eastern bracteates was perhaps partly responsible for the emergence of the Mark. This weight of solid silver, the mass of which varied from one time and area to another, stood at about $^2/_3$ of the gold pound, which equaled 240 western silver deniers.

Transition from medieval to modern coinage took place with the emperor Louis IV of Bavaria (1314–47), who introduced gold and multiplied the silver grossus already issued by Cologne under Henry VII (1308–13). Louis reduced the number of purely imperial mints. Many others operated by rights granted to the nobility, the churches, and certain municipalities, and from these henceforth appeared the bulk of German coinage, including from 1520 the large silver thalers (so called from the Joachimsthal mines in Bohemia and from which derived the word *dollar*). In the 16th and 17th centuries the thalers and their multiples, of handsome and even ornate appearance, dominated the silver currency of Germany. Thalers of Saxony and Brunswick are especially well known. The thaler continued as a unitary denomination to the 19th century in Germany proper, but in 1870 German adherence to the gold standard caused its abandonment. From 1870 the kings of Prussia as emperors coined for all Germany; henceforth, the innumerable local variations in coinage were subsumed under the gold Reichsmark of 100 pfennigs, the silver standard being abandoned. After World War I the rulers of German states abdicated or were deposed, and everywhere the value of the Mark declined to zero, its place being momentarily taken by inflated paper currencies. Silver was coined mainly for commemorative pieces between World Wars I and II, including the Hindenburg portrait pieces; zinc, aluminum, and alloys furnished the wartime currency of 1939–45. After 1948 the coins of West Germany were inscribed *Bundesrepublik Deutschland;* those of East Germany, *Deutschland* alone (with emblems of industry and fertility).

In Austria there was a ducal silver coinage in the 11th century. It remained crude until the 14th century, when Albert II (1330–58) introduced a gold florin of Florentine character. The gros appeared with Frederick III (1440–93): thereafter, development was parallel with that of Germany, with thalers taking a prominent place. Those with the portrait of Maria Theresa acquired wide popularity on either side of the Red Sea. They continued to be coined in large numbers at Vienna and London, with the date 1780, for circulation in those regions: 24,000,000 were struck in 1940–41 from British mints alone.

The Bohemian ducal coinage of deniers from the 10th to the 12th century showed Byzantine, Scandinavian, and even English influences; by the 12th century the Prague mint was developing its own style. Wenceslas II first produced the gros in 1300, and John of Luxembourg (1310–46) the first gold florins, with, obverse, crown and, reverse, rampant lion. The regal coinage of Hungary began with the deniers of Stephen I (St. Stephen; 1000–38), and the style remained crude until Charles I (1310–42) introduced

a lily-bearing gold florin and a silver gros modeled on those of Naples and Rome.

Austro-Hungarian imperial coinage

With the formation of the Austro-Hungarian Empire in 1848, the coinage of the two countries, including Bohemia, was unified. During 1857–68 the coinage conformed to the terms of a monetary convention with Germany. The coins of Austria and Hungary were differentiated from 1868: the former were inscribed in German or Latin, and the latter in Magyar. Since 1923 the republican coinage of Austria has been conspicuous for its commemorative silver coins. That of Hungary, under the regency of Adm. Miklós Horthy, emphasized the crown of St. Stephen; under Soviet domination types have symbolized revolution, peace, fertility, and industry, together with architectural motifs for silver.

Czechoslovak coinage from its inception in 1918 has shown the lion of Bohemia; special coinages have commemorated St. Wenceslas (in gold) and Tomáš Masaryk and—after the Soviet occupation of 1968—Stalin (in silver). Yugoslavia, similarly an offshoot of Austria-Hungary, has a currency based on paras and dinars. That of Albania, until its domination by the Soviet Union in the early 1950s, drew heavily on classical Greek types.

**Scandinavia.** The origin of Norwegian, Danish, and Swedish coinages is clearly the result of the Danish conquest of England. The Runic alphabet was employed, though not by any means exclusively, on many early coins of Denmark and Norway. The Norwegian series began with Haakon the Great (*c.* 970–995), who copied the pennies of Ethelred II. In the second half of the 11th century, a coinage of small, thin pennies began, which developed into bracteates. Magnus VI (1263–80) restored the coinage, more or less imitating the English sterlings of the time. The money of Denmark began with pennies of Sweyn I (*c.* 987–1014), also copied from the coinage of Ethelred II; the coins of Canute (Cnut) the Great (1016–35) and Hardecanute (Harthacnut; reign extended to England in 1040–42) were mainly English in character. With Magnus I (reign extended to Denmark in 1042–47) other influences, especially Byzantine, appeared, and the latter was very strong under Sweyn II Estridsen (1047–74). Bracteates came in in the second half of the 12th century. The coinage is very difficult to classify until the time of Eric of Pomerania (1397). There were important episcopal coinages at Roskilde and Lund in the 12th and 13th centuries. Sweden had very few early coins; they began with imitations by Olaf Skötkonung (995) of English pennies and included the usual bracteate coinage. The money was restored by Albert of Mecklenburg (1364–89). The thaler was introduced by Sten Sture the Younger (1512–20). The money of Gustav II Adolf (1611–32) is historically interesting. Under Charles XII (1697–1718) there was highly curious money of necessity (*i.e.,* a coinage struck to fulfill a need, usually in time of war and siege, but with inadequate technical means available). The small copper daler was struck, sometimes plated; types included Roman divinities. During the 17th and 18th centuries there was a large issue of enormous plates of copper, stamped with their full value in silver money as a countermark.

Modern Norwegian coinage, like that of Denmark, is remarkable in including certain denominations pierced with a central hole. That of Sweden has included some large commemorative pieces of silver. In Denmark the Copenhagen mint has produced a colonial coinage for Greenland. Iceland, formerly joined with the Danish crown, has struck republican coins since 1944.

**Poland.** After monetary beginnings derived from Germany, Poland developed a 16th-century coinage in gold, silver, and billon that reflected its status as the greatest power in eastern Europe; its thalers were especially remarkable for fine portraiture and decoration, including the superb pieces coined by Danzig after 1567, when this area sought Polish protection. Dismemberment of Poland in the 17th and 18th centuries was followed by fluctuations in status, which have continued ever since. The coinage of independence after World War I celebrated national symbolism and national heroes, such as Józef Klemens Piłsudski and John III Sobieski. On the coins produced during German occupation in World War II and under Soviet control thereafter, the Polish eagle has been a prominent emblem. Danzig struck its own coinage (in pfennig and gulden) while a free city (1920–39).

**Russia and the Balkans.** The earliest Russian coins were produced for the princes of Kiev in the 10th century and showed strong Byzantine influence. The staple coinage later came to consist of small silver kopecks and their halves (dengi) of Mongolian derivation. Ivan IV (1547–84) standardized the types of the dengi as "Tsar and Grand Prince of All Russia," showing a uniform design of a mounted lancer. From the 15th to the 17th century unstable social and economic conditions were reflected in clipping and counterfeiting, until reforms began in 1654. Peter the Great (1689–1725) reorganized the currency: gold was coined regularly from 1701, and silver rubles and billon kopecks also appeared, together with copper fractions. In 1725, after Peter's death, copper "plate money" was briefly produced (as in Scandinavia) at Ekaterinburg. Recoinage on a large scale occurred in 1741. Under Catherine II (1762–96) copper rubles of great size were briefly struck, and substantial five-kopeck pieces were in common production; Russian copper was also produced in Georgia. In the 19th century, Russian coinage followed conventional lines apart from the short-lived introduction in 1828 of platinum for pieces of three, six, and 12 rubles. The silver ruble, however, remained the monetary basis, worth 100 kopecks until a change to gold in 1897. Soviet issues since the Revolution have been mainly of alloys, with scarce silver and, very rarely, gold; types have usually included the hammer and sickle and the star, together with allusions to industry and agriculture, though the Russian eagle was used at first.

Currency of Peter the Great

Finland, as a Russian grand duchy from 1809, struck in gold, silver, and bronze until declaring independence in 1917; since then, its coins have shown the Finnish lion. Latvia coined as an independent state from 1918 to 1940. Lithuanian independence, similarly until 1940, was reflected in autonomous coinage, with the type of a galloping knight.

The medieval coinages of the northern Balkan states are of great morphological interest. They are chiefly silver grossi, showing a mixture of Byzantine and Venetian influences. The Bulgarians had a regular silver coinage from Ivan Asen I (1186–96) to Ivan Shishman (1371–93). Modern Bulgarian coinage began in 1879. The Serbian coinage lasted from Stephen Vladislav I (1234–43) to the mid-15th century. There was also a coinage of the bans (local officials) of Bosnia (late 13th to 15th century). The independent city of Ragusa is remarkable for the bold Roman style of its early copper (13th century) and for its rich and varied later issues.

In Romania a princely coinage from 1866 became a royal one, of orthodox pattern, from 1881; the 20th-century types, until the fall of the monarchy in 1947, were remarkably varied. That of Greece began with the republican government of 1828: the basis was the silver phoenix of 100 lepta. This was followed, under the monarchy from 1833, by the drachma of similar value. The 20th century emphasized the types of ancient Greece, though modern issues have broken from this tradition.

**The later Byzantine empires.** From the time of Basil II (976–1025) the fabric of the gold nomismata (successor of the solidus) and also of the silver began to change, from using a narrower, thicker blank (flan) to one wider and thinner, which was also given a curious cup shape, hence the name *nummi scyphati* (cup money); gold scyphati declined in purity until, under Nicephorus III (1078–81), they were very base. Silver remained generally scarce; the issue of bronze became uneven. New conventions in legends and types were introduced: Constantine IX (1042–55) showed on his silver an invocation to the Virgin in iambic trimeter; and an invocation used by Romanus IV (1068–71) took the form of a hexameter, carried over from obverse to reverse. Figures of the saints appeared in the 12th century. At the same time, the intrinsic quality of the coinage had sunk to a level of desperate confusion, seen most plainly under Alexius I Comnenus (1081–1118), whose "gold" was sometimes no more than billon or even bronze. The influence of Western types was seen

Legends and types

powerfully in the bronze struck by Andronicus II with, reverse, a cross pattée surrounded by a circular inscription within a double border. Western influence continued in the 15th century, especially under John VIII Palaeologus, whose visit to Italy in 1438 (when Pisanello made his splendid portrait medal) doubtless familiarized him with the designs of the grosso and gros, which were imitated unmistakably on John's silver and from which derived the English groat. By this time the Byzantine idiom in coinage was virtually dead.

With the capture of Constantinople by the crusaders in 1204, the power of the Byzantine Empire was split among a number of smaller authorities, of which the "empires" of Thessalonica and Nicaea were short-lived: in both, the coinage (where attributable) was of normal Byzantine character. The empire of Trebizond, however, continued a separate existence until 1461; its small silver pieces, called "Comnenian white money," were prized for their purity and enjoyed a wide currency. Through such means the influence of Byzantine types was exerted on the contemporary coinages of Armenia and elsewhere in Asia Minor.

(C.H.V.S.)

## COINS OF THE BRITISH ISLES, COLONIES, AND COMMONWEALTH

**Ancient Britain.** The earliest coinage of Britain consisted of small, cast pieces of speculum, a brittle bronze alloy with 20 percent tin. These coins copied the bronze of Massilia (Marseille) of the 2nd century BC and circulated, mainly in southeastern Britain, early in the 1st century BC; their relationship with contemporary iron currency bars is uncertain. At the same time, uninscribed gold coins of the Gaulish Bellovaci, a tribe located near Beauvais, imitated from the famous gold stater of Philip II of Macedon, were being introduced, probably by trade. The first Belgic invasion, *c.* 75 BC, brought variants of these, from which arose a complex family of uninscribed imitations. The study of distribution in Britain has ascribed them to fairly well-defined tribal areas in the south and east; some are crude, but the best illustrate the peak of Celtic art in Britain. The gold was of variable purity. After the second Belgic invasion (following Caesar's raid in 55 BC) the coinage entered a historical phase through the addition (in Latin, and with Roman titles, etc.) of the names of kings. Roman influence under Augustus prompted the introduction of silver and copper to reinforce the gold and the Romanization of previously "Celtic" types. Claudius' conquest in AD 43 ended native coinage except for crudely cast billon pieces long continued in the Hampshire and Dorset area; the gold of the tribe of the Brigantes in what is now Yorkshire was the last to disappear.

**Roman Britain.** Unofficial copies of Claudian bronze were produced in Britain to alleviate the shortage of official Roman coinage after the conquest. Thereafter, no coinage was produced until the reign of the usurper Carausius (AD 286–293), who coined profusely in orthodox Roman fashion at Londinium (London) and elsewhere in gold, silver, and copper; the same was done briefly by Allectus, his murderer (AD 293–296). Diocletian's London mint was continued under Constantine until AD 324; thereafter, except under Magnus Maximus (AD 383–388), whose usurpation was legitimized by the Eastern emperor Theodosius I, Britain lacked an official mint, being supplied with coinage mainly from Gaul. Imitative bronze pieces, however, appeared in the 3rd century and continued to be made in the 4th.

**Early Anglo-Saxon coins.** Infiltration of Merovingian gold from France in the 6th century prompted the issue of Anglo-Saxon gold "thirds" in the 7th; solidi were only very rarely struck, because of their high intrinsic value. Output, never great, was confined chiefly to the London–Kent area. The London mint, almost certainly episcopal, signed its coins with the name *LONDVNIV;* Kentish coinage was mainly regal. In addition, there were a perhaps small Mercian series and another from York. A further series, copied from late 4th-century Roman prototypes, was struck *c.* 650, when the gold content was fast diminishing. Gold coinage soon gave way to that of small thick silver sceats (meaning "a portion"; about 1.29 grams, or 20 grains) of essentially different style. Some had Runic legends, including the name Peada, supposedly a reference to the king (flourished 656) of Mercia; most, however, were nonregal, and their legends are Latinized. Types were varied, and some almost certainly originated in Frisia, where sceats are found in large quantities, denoting the trading connection that called for their use; these show animal and floral design. In the south the sceats lasted until *c.* 800. Small silver sceats were developed in the mid-8th century in Northumbria, where they quickly gave way to copper, which lasted until *c.* 850.

The London mint

**Anglo-Saxon penny coinages.** English coinage proper began with the silver penny of Offa, king of Mercia (757–796). It was first struck at around the weight of the sceat, from about 790, and its weight increased to about 22½ grains (equal to 240 to the Tower pound; the standard pound used by the Royal Mint until its replacement in 1526 by the troy pound, whose name derives from the French city of Troyes, site of a major medieval fair). The new pennies showed Carolingian influence in their broad, flat forms. Their designs, however, insofar as they were not influenced by late Roman coin portraiture, displayed a brilliant originality (both in Anglo-Saxon portraiture and also in pattern design and decorative lettering), which had no equal for some centuries. Offa's coinage, though minted expressly for him as *Rex Merciorum,* was mainly current in the southeast and was probably struck at Canterbury. Evidence for this lies in the fact that the moneyers of Offa were also those of the kings of Kent, and the coins of archbishops of Canterbury, including Jaenbeorht (died 792), bore the names of Offa and his successor Cenwulf: under Ceolwulf (821–823) the mint name of Canterbury appeared on the coins. Offa struck pennies with the portrait and name of his wife, Cynethryth, as *Regina M(erciorum)* and also issued gold pieces copying a 774 dinar of the caliph al-Manṣūr but with the addition of *OFFA REX.* Ceremonial gold coins, like Offa's, all now represented by unique examples, were minted, perhaps partly to pay the Romescot (an annual tax to the papal see), by Archbishop Wigmund of York (837–854?), Edward the Elder (899–924), Ethelred II (978–1016), and Edward the Confessor (1042–66).

Offa's coinage influenced design under the kings of Kent and East Anglia, as can be seen in the coinage of the Wessex kingdom, which was produced first at Winchester, then, after the Battle of Ellendun in 825, at Canterbury, still the only permanent mint south of the Humber. Under Aethelwulf (839–858) a uniform type of coinage was achieved for all of southern England except East Anglia. The Viking invaders, from *c.* 870, left an important mark on English coinage, with new designs of northern European derivation. York and Lincoln were their principal mints, from which numerous memorial coins of SS. Peter and Martin were issued. Meanwhile, Alfred (871–899) greatly extended the Wessex kingdom, as shown by his operation of mints: Canterbury (still much the largest), Gloucester, Exeter, Winchester, London, and possibly Oxford. His coins were of careful workmanship and showed Viking influence. In the 10th century the power of English kings spread quickly northward. Under Athelstan (924–939) there were nearly 30 mints at work, mainly southern and central but reaching to Chester; under Edgar (959–975) there was much more uniformity of type. The Council of Grateley under Athelstan had enacted that each permitted mint was to have but one moneyer, with specified exceptions; London, for example, had eight. By the time of Ethelred II more than 70 mints were at work; London, Winchester, Lincoln, and York were the largest and most profuse. From about this time the types were generally standardized: obverse, king's portrait, and, reverse, some cruciform design.

Viking influence

**Post-Conquest coinage.** The Norman Conquest of 1066 made little change in the mint system or in the coinage (though the facing portrait acquired great popularity); the pre-Conquest moneyers stayed in office and struck coins for William I. After his reign the number of mints tended to decline. The pennies of William II have nothing in their legend to distinguish them from his father's issues, but it is possible to allot eight types to William I and five to

his son. Forgery gave Henry I much trouble, and one step he took to prevent it was to issue his later coins with a snick in the edge to show that the silver was good. He also coined round halfpence; previously, silver pennies had had to be halved or quartered to produce a smaller denomination. The civil wars of Stephen's reign produced many interesting coins, such as those struck in the claimant Matilda's name as *Imperatrix* and the pennies of Eustace Fitzjohn and other barons, very much on the pattern of feudal issues in France.

Henry II ceased the regular change of types customary since William I's reign and struck one type until 1180. As a result the work of the English mints reached its lowest level. His short-cross penny, so called from its reverse design, first issued in 1180, remained unchanged—including the name *Henricus*—not only by Henry II but also by Richard and John and Henry III until 1247, when Henry III coined the long-cross penny with the arms of the cross extended to the edge of the coin to discourage clipping. He also reduced considerably the number of mints. Edward I subordinated all mints to the authority of the master worker in London, William de Turnemire. In 1279 he introduced a new type of penny, with, obverse, bust of the king and, reverse, long cross with three pellets in each angle, a type that was much imitated abroad and persisted on silver until Henry VII. The moneyers' names disappeared from the reverse legends, and their place was taken by the name of the mint (*e.g., CIVITAS LONDON*). Edward I also struck halfpennies and farthings to replace the cut pennies that had hitherto done duty for small change. He also introduced a groat, or fourpenny piece, but this larger coin did not establish itself until Edward III's reign. The coins of Edward I, II, and III can be distinguished only by a minute study of detail. Privileged ecclesiastical mints still continued active.

Henry III had attempted in 1257 to issue a gold coinage by striking the gold penny (45 grains) of the value of 20 pence silver, later raised to 24; but the difficulty of relating gold to silver proved insuperable, and the coinage was withdrawn. In 1344 Edward III issued his fine gold series—florin, leopard, and helm (1/2 and 1/4 florin)—but his attempt to introduce a gold currency failed. A gold coinage was finally established in currency in 1351 with a noble of 120 grains of gold and its subdivisions, the half- and quarter-noble. In the same year, the silver penny was reduced to 18 grains and the groat issued (on Flemish models). The noble was valued at six shillings and eightpence (1/2 mark). Its obverse, the king in a ship, is supposed to allude to the naval victory off the Flemish city of Sluis in June 1340. The reverse type is a floreate cross with considerable ornamentation. The weight of the noble was reduced by Henry IV in face of foreign competition. Edward IV distinguished his noble by a rose on the ship (rose noble, or ryal) and raised its value to 10 shillings, while a new gold coin, the angel, was introduced to replace the old value of the noble; the penny was reduced to 12 grains. The angel is so called from its type of St. Michael and Lucifer. The reverse is a ship with a cross in front of the mast. (In the 16th century the angel became the piece given to those touched for King's Evil [scrofula] in the belief that the king's touch could cure. It was struck for this purpose down to the reign of Charles I, and small versions were struck by the later Stuarts and pretenders, but it was not again issued as legal tender.)

Establishment of gold coinage

The next important change in the coinage was the introduction in 1489 of the sovereign, a splendid gold coin of 240 grains, current for 20 shillings, with, obverse, Henry VII seated on an elaborate throne and, reverse, a Tudor rose with central shield of arms. Henry also issued the first English shilling, a handsome, though scarce, coin with a fine portrait, probably by John Sharp, formally appointed engraver in 1510. Henry VII altered the types of the smaller silver coins by replacing the three-centuries-old cross and pellets by a long cross and shield, while the inscription *POSVI DEVM ADIVTOREM MEVM* ("I have made God my helper") took the place of the mint legend; the stereotyped bust was replaced on the groat by an excellent profile portrait and on the penny by the king seated. Henry VIII debased the gold coinage and reduced the weight of the sovereign, the reverse type of which was now the royal arms supported by a lion and dragon. He introduced the gold crown of five shillings, with its half, raised the angel to seven shillings and sixpence, and introduced the George noble—so called from its type of St. George and the Dragon—to take the angel's old value. In 1544 he issued the base shilling, or teston, of 12 pence and debased the silver coinage. When Edward VI again restored a coinage of fine silver, he introduced the silver crown of five shillings (the first English coin dated in Arabic numerals), which took the name of the gold piece of the same value introduced a few years earlier. The reign of Mary is notable for the appearance of the portrait of her husband, Philip II of Spain, on the shilling.

Coinage under Elizabeth I

Elizabeth I continued her father's denominations and restored the purity of the silver coinage. She soon discontinued the groat, Edward VI having introduced the silver sixpence and threepence, although she continued its half, the twopence. Her "portcullis," or trade coinage for use by the newly incorporated East India Company, appeared in 1600–01. She also experimented with machinery for coinage, although the insistence of the moneyers on their immemorial right to use manual methods delayed its establishment until after the Restoration. James I introduced a number of new gold coins, the most important being the "unite," or sovereign (20 shillings), so called from its legend (*Faciam eos in gentem unam* ["I will make them into one race"]) alluding to the union of the crowns of Scotland and England. Charles I made no changes in the coinage until the Civil War (when Parliament coined in London and the King's mint traveled with him); the King's financial difficulties added many new coins to the English series. These included 20-shilling and 10-shilling pieces in silver, the large gold £3 pieces of Oxford, and the fine Oxford silver crown, with a view of Oxford below the usual type of the king on horseback, made by the engraver Thomas Rawlins, employed at the Oxford Mint (1642–46) under its master, Thomas Bushell; the siege pieces rudely struck on silver plate at various Royalist strongholds show to what straits the King's party was reduced. Under James I and Charles I are found the first English copper coins, the "Harrington" farthings, which were struck under contract. From 1649, copper tokens, mainly of farthing value, were produced in large numbers by many municipalities and private traders. The coinage of the Commonwealth (1649–60) is remarkable for the simplicity of its types, and this is the only period of English coinage when the legends have been in English. Coins struck with the lord protector Cromwell's bust and superscription, although not uncommon, apart from the 1656 half crowns, seem never to have circulated.

**Modern coinage.** The modern coinage dates from the reign of Charles II. After issuing the old denomination of hammered money in the first two years of his reign, he replaced the unite, or broad, in 1662 by the guinea, so called from the provenance of its gold. This was a 20-shilling piece. It was not until 1717, after various oscillations, that its value was fixed at 21 shillings. His silver coins were the crown, half-crown, shilling, and so on, all regularly and beautifully struck on the new mill that was then established at London's Tower Mint. In 1672 he introduced the copper half-penny and farthing with the Britannia type. The finest coin of his reign is not a regular issue. It was the "Petition" crown made by Thomas Simon, engraver at the mint under the Commonwealth, and bears on the edge a petition to the King that he might be given the same office under the restored monarchy. For the great recoinage under William III, provincial mints were briefly opened at Bristol, Exeter, Chester, Norwich, and York. Of 18th-century coinage mention may be made of the practice of recording the provenance of the metal of particular issues, as in the *VIGO* issues of Anne struck from captured Spanish bullion in 1702, the *S(outh) S(ea) C(ompany)* silver of George I, and the *LIMA* coinage of George II made of bullion brought by Admiral Anson from his voyage around the world. Toward the end of the century the scarcity of government silver was largely made good by Spanish dollars, and by tokens issued by the Bank of England. The deficiency in copper, briefly remedied by

The "Petition" crown




George III "cartwheel" copper twopence struck not at the Royal Mint, London, but under contract at the Soho Mint, Birmingham, using steam-powered coining presses, 1797. In the British Museum. Diameter 40 mm.

Reproduced with permission of the trustees of the British Museum; photograph, Ray Gardner for The Hamlyn Publishing Group Limited

large "cartwheel" issues, was made up by private issues of vast numbers of tokens. In 1816 a great recoinage took place with the introduction of the sovereign and silver coins, each with Benedetto Pistrucci's design, St. George and the Dragon. Britain was on the gold standard from 1821. In 1848, the two-shilling piece (florin) was issued and proved a most popular coin. The double florin, which was first issued in 1887, did not take the public fancy; the practical disappearance of the crown piece from circulation also reflects the public dislike of large coins, though commemorative crowns have remained popular.

The gold sovereign disappeared from internal currency in 1914 after a career of 300 years, but it has continued to be struck in irregular quantities for export abroad. Silver was alloyed up to 50 percent in 1920 and in 1947 gave way to cupronickel: the silver threepence was replaced by an angular coin of nickel-brass.

Decimal currency

On Feb. 15, 1971, the U.K. currency was decimalized. With 100 "new pence" to the pound, the decimal coinage was based on denominations of these new pence (½ [since demonetized], one, two, five, 10, 50); the sixpence coin continued briefly as a 2½-pence coin before being phased out, as were the old penny and threepence coins. On April 21, 1983, three versions of a £1 coin were introduced bearing the head of Elizabeth II, obverse, and the royal coat of arms, or the emblematic leek (Wales), or thistle (Scotland), reverse. Versions bearing the designs of a flax plant (Northern Ireland) and an oak tree (England) were introduced in 1986 and 1987, respectively. Royal maundy money (since the 18th century) is maintained with the issue of silver pieces of onepenny, twopence, threepence, and fourpence: the silver penny and groat thus survive.

Wales has had no coinage, except perhaps for the penny of the chieftain Howel Dda (909–950).

**Scotland.** Coinage began by following English usage in regard to types and weights: the earliest silver pennies were those of David I (1124–53) and copied Stephen's, though the use of profile portraits in the 13th and 14th centuries showed an interesting divergence. Gold nobles and silver groats were issued by David II in 1356–57 on the standard of Edward III. From Robert III onward the French or Flemish standard for gold was preferred, and during the 16th century, especially under James V and Mary, a strong continental influence on design was apparent in a series of gold coins of originality and frequent beauty. Silver coins had begun to show debasement of metal, and as early as James III copper small change—"black" farthings—had been introduced. The Scottish coinage of James VI (James I of England) marked a peak in range and variety: after the union of the crowns in 1603, Scots coinage decreased in quantity and ceased in 1707 after the union of the Scottish and English parliaments. Hitherto the value of Scottish coinage in relation to English had been 12:1. Under George VI a shilling with a Scottish reverse was first coined as part of the general British series.

**Ireland.** Irish currency from the 8th to the 10th century consisted mainly of Anglo-Saxon, French, Viking, and Arabic silver; in the later 10th and 11th centuries silver pennies of the Norse kings of Dublin imitated Anglo-Saxon and Scandinavian models. There were also thin silver bracteates copying Norman types. Anglo-Irish coinage proper began with silver pennies and halfpence of John (the only coins to bear his name, which did not appear on English coins). Dublin, Waterford, Cork, Limerick, and Trim were striking silver (increasingly base) from groat to farthing from the 13th to the 15th century. The "three crowns" coinage with the saltire cross (the Fitzgerald arms) beside the shield came in with Edward IV and was continued by Richard III and Henry VII. "Harp" groats were struck at Dublin under Henry VIII, followed by his much baser issues. Gold was never coined, but copper was introduced quite early. In Ireland, as in England, the English Civil War produced a number of siege pieces, notably the money of the Irish peers Inchiquin and Ormonde. James II for his Irish campaign issued his "gun-money" series of brass (made partly from melted down old cannon), to be redeemed in silver when he should regain the throne. In 1723 there was uproar in Ireland over the monopoly granted to the English entrepreneur William Wood to mint a debased Irish coinage, derisorily termed "Wood's halfpence." Under pressure from, among others, Jonathan Swift, the satirist, whose "Drapier's Letters" furiously attacked him, Wood revoked his patent in 1725 (see below *Coins of the United States*). Irish coinage was discontinued in 1822; from then on, Irish needs were served by English coinage. In 1928 the coinage of the Irish Free State (later the Republic of Ireland) was introduced, with, obverse, harp on all denominations, and reverses bearing a range of animal, bird, and fish designs. In February 1971 the coinage of the republic was decimalized.

Siege pieces

**Isle of Man and Channel Islands.** The Isle of Man first had its own coinage, in bronze, from 1709 to 1733 under the earls of Derby; this was continued, in 1758 only, for the earls of Athol. Regal coinage in bronze appeared intermittently from 1786 to 1839. The characteristic badge was the "three-legs," or triskelis, forming the spokes of a wheel. Since 1840, English issues have been current.

Jersey and Guernsey have had their own bronze coinage for well over a century, showing the shield of three leopards proper to the Duchy of Normandy. Variation of types has occurred since World War II, the end of which prompted also the special issue of the Jersey "liberation penny" to mark the end of the German occupation. The coinage was decimalized in February 1971.

**Colonies and Commonwealth.** British colonial issues, begun under Elizabeth I with silver for the East India Company, were extended in the 17th century. New England colonists struck the silver "pine tree" and "oak tree" money from 1652; Charles II had silver rupees coined at Bombay for the East India Company with the company's arms; silver and copper "hog money" (obverse, boar; reverse, ship) was issued for Bermuda; and James II struck tin coins for American plantations.

There were few official attempts to provide colonial coinages in the 18th century; thus, currency in the British West Indies was based on Spanish, Portuguese, and Brazilian gold and especially on Spanish silver dollars, normally cut and counterstamped. Spanish dollars were similarly used in the early 19th century at Sierra Leone. In the 19th century, however, colonial issues proper multiplied. That of the Ionian Islands, from 1819, was among the earliest. In Malta one-third farthings were issued by William IV and Victoria. Gibraltar had copper from 1842. Farther afield, token bronze had been coined for Nova Scotia and New Brunswick from 1823; a general coinage for Canada appeared in 1858. Ceylon's coinage began with bronze half- and quarter-farthings and silver three-halfpence from 1838. The private *raj* of Sarawak had coinage from 1841. Australian coinage of sovereigns started as early as 1855; its unit was changed to that of the dollar on Feb. 14, 1966. In Cape Colony little coinage was produced until the Boer republic of South Africa had been incorporated in the Union. Silver and bronze for Hong Kong began in 1863, and in closely succeeding years colonial issues were

started for Jamaica, Cyprus, Mauritius, Zanzibar, North Borneo, Honduras, and elsewhere.

0th-entury sues

In the 20th century, coins of the colonies continued in general to show a crowned bust of the monarch; those of the self-governing Commonwealth powers exchanged a crowned for an uncrowned bust. New Zealand issues, with Maori designs prominent, began only in 1933. Indian and Pakistani coinages, each bilingual with English, grew out of the imperial Indian coinage, the British sovereign's head being replaced in India by pictorial designs and in Muslim Pakistan by calligraphic and symbolic devices.

Generally, Commonwealth and colonial coins alike have emphasized on their reverses either national symbols or national heraldic devices. Those U.K. dependencies that had not by February 1971 decimalized their currencies adopted the new decimal currency that the United Kingdom introduced at that time. The currencies of the many nations that achieved independence in the second half of the 20th century exhibited a variety of types, including portraits, traditional emblems, and renderings of indigenous flora and fauna. (C.H.V.S.)

## COINS OF LATIN AMERICA

**The colonial period.** Spanish colonists carried to the New World the Castilian currency system, which had been regulated as to standard, weight, and size of the coins within a bimetallic pattern by the ordinances of Ferdinand and Isabella issued in Medina del Campo in 1497. The double base of the system consisted of the gold excelente (replaced in 1535 by the escudo) and the silver real. The coins of Spanish America were specifically: in gold, the escudo (3.38 grams), two-escudos, four-escudos, eight-escudos, or onza (the famous gold ounce), and the half-escudo, or escudito; in silver, the real (3.43 and 3.38 grams), the half-real and the quarter-real, or cuartillo, and the two-reales, four-reales, and eight-reales (this last known also as the duro, or peso fuerte). During the 16th century, for a brief period, a coin of three reales was minted in Mexico. Gold was not minted in a uniform manner until after the second half of the 17th century; until then Hispanic-American currency had been almost exclusively silver coinage. Copper was rarely minted in Spanish America.

'he gold unce

The hammered coinage of Spanish America frequently presents a relatively tidy appearance, being very nearly round and containing all the lettering and required symbols; but the press or mill type coinage is frequently of very poor appearance. These coins of rude mintage are called macuquinas (cob). In the 18th century, by ordinances of Philip V, the setting up of machinery for the minting of a perfectly round coinage, with milled and corded (ropelike) edge, became mandatory.

The type of the Hispanic-American coin was very characteristic: its most constant elements were the Pillars of Hercules and the motto *Plus Ultra,* plus the monarchy's coat of arms. In edge-milled coinage the same elements were employed in silver pieces, with the addition between the Pillars of an image of the two crowned hemispheres; this was called the moneda columnaria ("columnar coinage") and was minted until 1772. From that date, by ordinances of Charles III, silver coinage carried on the face a bust of the reigning monarch and on the reverse the coat of arms, a system already utilized in the gold pieces.

**Hispanic-American colonial mints.** At the beginning of the colonial period, stamped metal foundry pieces frequently substituted for scarce currency. In time, several mints were established, of which the Mexican (1535–1821) and the one at Potosí (1574–1825) were particularly important. Other minor ones, and their dates of operation, were those of Santo Domingo (1542 to the end of the 16th century and 1818 to 1821), Lima (1568 to 1570, 1575 to 1588, 1659 to 1660, and 1684 to 1824), Santa Fe de Bogotá (1626 to 1821), Guatemala (1731 to 1822), Santiago de Chile (1749 to 1817), Popayán (1732 and 1749 to 1822), and Cuzco (1698 and 1824). Coinage of any of these mints had uniform currency throughout the entire Spanish Empire, and the pieces had uniformity of type. They were distinguished by the symbol of the mint, carried on every coin. The following are some of the symbols used: Mexico, *M̊;* Potosí, *P* and, in the edge-milled coins, *PTSI* and *PTS* in monogram fashion; Lima, *P, L,* and, in the edge-milled coins, *LIMA* and *LIMAE* in monogram fashion; Santiago de Chile, *S̊;* Guatemala, *G* and *NG* (for Nueva Guatemala); Santa Fe de Bogotá, *NR* (for Nuevo Reino); Popayán, *P, PN,* and *P^N;* Santo Domingo, *SD;* Cuzco, *C°* and *CUZ.*

Spanish colonial ypes

**Dissemination of Hispanic-American coinage.** The larger silver and gold pieces, the eight-reales, or pesos fuertes, and the ounces, became in modern times the international currency par excellence. Their dissemination throughout the world was brought about by the uniformity of their standard and milling characteristics. In many countries they were counterstamped to adapt them to a local monetary system or to authorize their currency.

**Emergency coinages in the era of independence.** During the wars of independence, between 1810 and 1826, emergency mints were established in different parts of the continent, by the royalists as well as by the patriots. The coinages were almost always crudely designed, being in some instances merely foundry coinage. On other occasions the coins had merely fiduciary value—that is, no intrinsic value at all—as was the case with the numerous coinages of the Mexican patriot José María Morelos (1765–1815), who produced eight-real pieces in copper. Only in Mexico were there mints of any importance, situated in 10 different localities. The coinage situation became further complicated when the authorities of the opposing forces started counterstamping each other's coins in order to use them within their own camps.

**The independent countries.** The independent states that arose in Latin America after the revolutions of 1810 proceeded to mint new coins, retaining the bimetallic system established by Spain, with units in reales and escudos, except for copper fractionary units. After 1850, within a period of about 15 years, all the states adopted the decimal system, and the peso became the unit, though in several cases it took a special name. Within the second half of the 19th century, bimetallism was generally replaced by the gold standard, which in the 20th century was replaced in turn by fiduciary currency or paper money, coinages being limited to fractionary pieces or to "merchandise coins" (trade tokens with little inherent metal value).

**Brazil.** Coins minted in Spanish America circulated abundantly in Brazil from the 17th to the 19th century. They were given their official value in terms of the Portuguese reis, the corresponding amount being indicated by counterstamping. Hispanic-American eight-real pieces carried an overstamp that was at first of 480 reis, increasing until in edge-milled coins it amounted to 960 reis. By the 18th century, mints were established in Rio de Janeiro, Bahia, and Pernambuco, but joint circulation of both Hispanic-American and Portuguese coinages continued. Counterstamping ceased during the first decades of the 19th century, although Hispanic-American eight-real pieces and the equivalent coins of the independent Latin-American countries continued to be reminted with the value of 960 reis for some time. The Brazilian monetary unit that eventually became the milreis later became the cruzeiro, divided into 100 cents. (A. de A.-M.)

## COINS OF THE UNITED STATES

The first coins struck in the North American Colonies were silver shillings, sixpences, and threepences, made by silversmiths John Hull and Robert Sanderson at a mint in Boston from 1652 to 1682, by order of the general court of Massachusetts Bay Colony. This mint dated its coins 1652 over the entire 30-year period to conceal the continuous mintage from British authorities in London.

Pre-Revolutionary coinage

With very few exceptions, the coins circulating in the Colonies until the Revolution were unauthorized private issues or old worn coppers no longer acceptable in England or Ireland. Silver was rare (consisting mainly of Spanish and Mexican dollars) and gold almost nonexistent. Copper was then a semiprecious metal, and in theory (though seldom in practice) 24 copper halfpence contained a shilling's worth of copper. Some of the Colonies, notably those in New England, repeatedly experimented with paper money, with disastrous results. Ostensibly to satisfy the colonists'

needs for metallic currency, but in reality for the benefit of owners of mines in Cornwall, the Royal Mint in 1688 issued tin farthings bearing the image of James II on horseback and the curious denomination of 1/24 of a Spanish real. The Rosa Americana pieces, struck by William Wood of Wolverhampton under royal patent dated July 12, 1722, received a disappointingly small circulation in New York and New England. Another coinage by Wood in 1722–24, intended for Ireland but rejected there because of scandalous circumstances surrounding his purchase of the royal patent, was shipped to the North American Colonies. Later these coins were supplemented by quantities of lightweight imitation halfpence, made principally in Birmingham, Eng. Alone among the Colonies, Virginia (because of a provision of its 1606 charter) had an official copper coinage executed at the Royal Mint in 1773. New Hampshire authorized William Moulton to make coppers in 1776, but the number was extremely small. The Continental Congress, colonial delegates of the incipient United States, uttered pewter dollars in the same year to provide moral support for its inflated paper currency. These bore a sundial, the word *Fugio* ("I flee"), the motto "Mind Your Business," and 13 links for the united colonies.

British copper tokens

The end of the Revolution in 1783 occasioned the manufacture and circulation of immense quantities of British copper tokens designed for the American trade. Between 1785 and 1789 the Republic of Vermont and the states of Connecticut, New Jersey, and Massachusetts awarded contracts to various individuals to strike copper coins, and Congress similarly licensed James Jarvis in 1787 to make cents of the same design as the 1776 dollars. All these ventures were failures, the authorized coins being driven out of circulation by British tokens, Birmingham halfpence, and the lightweight issues of "Machin's Mill" (a clandestine mint near Newburgh, N.Y.). The copper panic of 1789–90 followed, coppers of all kinds dropping to 72 to the shilling from their former 14 or 15.

Congressional efforts to establish a national mint had resulted in the issue of the historic 1783 Nova Constellatio silver patterns of 1,000, 500, and 100 units, from dies by the Englishman Benjamin Dudley, exemplifying the extraordinary Morris Plan, drawn up by Robert Morris, superintendent of finance, which reconciled the diverse colonial moneys of account. In 1786, however, Congress adopted instead the proposals of Thomas Jefferson for a decimal monetary system based on the dollar, and in 1792 the mint was finally built in Philadelphia, with David Rittenhouse as director. Jefferson tried vainly to hire as die engraver a Swiss, Jean Pierre Droz, who nevertheless furnished dies, hubs, and presses. Before the mint was quite ready, the first official American silver coin, the half dime, was struck in October 1792 in John Harper's cellar a short distance away, from dies by Robert Birch and Joseph Wright, who were also responsible for the regular cents and half cents of 1793. Silver followed in 1794 and gold in 1795, the engraver being Robert Scot.

Later American designers

Later designers of American coins included Gilbert Stuart (1796 silver), Titian Peale and Thomas Sully (the 1836 dollars engraved by Christian Gobrecht), Augustus Saint-Gaudens (1907–33 10- and 20-dollar gold pieces, called eagles and double eagles), Bela Lyon Pratt (1908–29 half eagles and quarter eagles), Victor Brenner (the Lincoln cent), James Earle Fraser (the buffalo nickel), A.A. Weinman and Hermon MacNeil (1916 silver), John Flannagan (1932 quarter dollar), Laura G. Fraser, and Chester Beach and Gutzon Borglum (various commemorative coins).



Figure of Liberty on a U.S. $20 gold piece designed for Pres. Theodore Roosevelt by Augustus Saint-Gaudens, 1907. The relief being too high, the coin proved unsuitable for circulation. In the American Numismatic Society, New York City. Diameter 34 mm.

By courtesy of the American Numismatic Society, New York City

The discovery of gold and silver in various regions and the difficulty of transporting large quantities of bullion through country menaced by Indians and bandits prompted the founding of both private and federal local mints. The Bechtlers of Rutherfordton, N.C., coined locally mined gold long before the government built a mint in Charlotte. The California gold rush stimulated coining by many bankers and assayers. Private coinage was legal so long as the coins contained full bullion weight and purity and imitated no official issues; Bechtler and Moffat gold (the latter coined at San Francisco) circulated at about par until the Civil War, while lightweight private gold took a discount. The California private mints mostly ceased operations when the San Francisco federal branch started, but those in the less accessible regions of Colorado continued long afterward, and as late as 1901 Joseph Lesher struck octagonal silver dollars in that state.

From 1851 to 1900 many brief experiments with odd denominations were tried, all of which proved superfluous. A law of 1873 discontinued the silver dollar until political pressure from mine owners forced through Congress an 1878 act requiring the mint to buy $2,000,000 to $4,000,000 worth of silver monthly and coin the entire amount into silver dollars. Coinage was discontinued in 1935. Millions of the silver dollars long remained stored in banks and treasury vaults, but eventually they became scarce; and in 1964 a new minting was authorized. Gold was recalled in 1934, but gold coins of numismatic interest may be retained in any quantity by "collectors of rare and unusual coin" (Presidential Order 6260). By an act of 1853 all silver coins except the dollar are fiduciary. The passing of 19th-century artistic canons has been reflected in American coin designs, which since 1909 have portrayed statesmen rather than personifications of liberty. All the above influences have combined to make the 20th-century American coinage system the simplest in use in any major nation. (W.H.Br.)

Recall of gold currency

## COINS OF ASIA

**Ancient Persia.** *Achaemenids.* The ancient kingdoms of the Middle East—Egyptian, Sumerian, Babylonian, Assyrian, and Hittite—had no coined money. The use of coins reached Persia from the Lydian kingdom of Croesus and the Persian satrapies of Asia Minor. The first ruler of the Achaemenid dynasty to strike coins was probably Darius I (522–486 BC), as the Greek historian Herodotus suggests. The coins of the dynasty were the daric struck from gold of very pure quality and the siglos in silver; 20 sigloi (shekels) made a daric, which weighed 8.4 grams. The types of both coins were the same: obverse, the Persian king in a kneeling position holding a bow in his left hand and a spear in his right; reverse, only a rough irregular incuse caused in the striking. These roughly oval pieces were uninscribed and remained in issue unaltered in type until the fall of the empire. The issue of gold was the royal prerogative, but the conquered Greek and other cities and states were allowed to issue silver and copper, while a number of Persian satraps struck silver in their own names, producing some of the earliest and finest coin portraits. At the fall of the empire, various satraps struck silver coins of their own.

Daric and siglos types

*Parthians.* Alexander's coinage and that of the Seleucids were purely Greek in character. In the mid-3rd century BC the Parthians became a great power in Persia. They had an extensive but monotonous coinage in silver (tetradrachms and drachmas) and copper. The coins do not bear the name of the issuer but that of Arsaces, which was used as a dynastic title. Some of the coins are dated in the Seleucid era; on the later coins the Greek becomes corrupt and is often joined by an inscription in Persian. Some local dynasties (*e.g.*, of Persis and Characene), vassals of the Parthian kings, also struck coins.

*Sāsānians.* The Sāsānian coinage was very extensive in silver, and the early emperors also coined gold and copper, although rarely. The coin types throughout the dynasty are the same: on the obverse is a bust of the king with his name and titles, and on the reverse a fire altar, usually with two attendant priests. From about the 4th century AD, with a few earlier examples, the reverse legend gives the mint and the regnal year of issue. The standard of the gold coins is derived from that of the Roman solidi; the silver coins are drachmas following the Parthian standard and are remarkable for their broad, thin form, which was copied by the Arabs for their silver coins. The execution of the portraits in the early coins was remarkably fine but was later debased.

**Islāmic coins of the West and of western and Central Asia.** The conquering Muslims at first imitated the coinage of their predecessors. In the western provinces they issued gold and copper pieces imitated from contemporary Byzantine coins, modifying the cross on the reverse of the latter somewhat to suit Muslim sensibilities. In the eastern provinces the Arab governors issued silver dirhams that were copies of late Sāsānian coins (mostly of those of Khosrow II) with the addition of short Arabic inscriptions on the margin and often the name of the Arab governor in Pahlavi; even the crude representation of the fire altar was retained. Toward the end of the 7th century, the fifth Umayyad caliph, ʿAbd al-Malik, instituted a coinage more in keeping with the principles of Islām. This "reformed coinage" was of gold (first issued in AD 698–699), silver (first issued in 696–697), and copper. The old coin, called dinar (from the Aramaic derivation of the Roman denarius aureus), derived its standard (4.25 grams) from the Byzantine solidus; the standard of the silver coin (dirham, from the name of the Sāsānian coin, which in its turn was derived from Greek drachma) was reduced to 2.92 grams, but it retained in its thin material and style some features of its Sāsānian predecessor; the name of the copper change, *fals,* comes from the Latin word *follis* ("money bag," by derivation a copper coin of low value). The reformed gold and silver coinage has no pictorial type, only skillfully arranged inscriptions, which are nonetheless of high historical value.

Reformed dinar and dirham types

The reformed dinar and dirham bear on the obverse the Muslim profession of faith—"There is no god but God: he has no associate"—and around it the marginal legend "In the name of God; this dinar (or dirham) was struck at . . . in the year . . . ." The reverse area has a quotation from Qur'ān CXII, "God is one; God is eternal; He begets not and is not begotten nor is there any one like unto Him." Around is Qur'ān IX, 33: "Muḥammad is the apostle of God sent with guidance and the religion of truth to make it prevail over all other religions averse though the idolators may be." In the mid-18th century the ʿAbbāsids overthrew the Umayyad caliphate but at first made little change in the coinage. In time the caliph's name was added and, at the provincial mints, that of the local governor, and in the 9th century a second marginal inscription was added: "To God belongs the order before and after and that day believers shall rejoice in the help of God" (Qur'ān XXXI, 3–4).

The ʿAbbāsid caliphate broke up in the 9th and 10th centuries, and the succeeding independent rulers regularly put their own names on the coins, although they retained that of the caliph of Baghdad, whose nominal authority was still recognized. Thus, in northern Africa and Egypt the Idrīsids, Aghlabids, Ṭūlūnids, and Ikhshīdids had their own coinage. From the eastern provinces there are the coins of the Ṭāhirids, Ṣaffārids (both in the 9th century), and the Būyids (10th–11th century). In Central Asia there was the extensive coinage of the Ṣāmānids, mainly in silver. In northern Africa and Egypt the extensive Fāṭimid currency in gold introduced a new type of dinar with legends arranged in three concentric circles. In the west the Umayyads of Spain issued a copious coinage from the mid-8th to the beginning of the 11th century, first in silver and later also in gold; their tradition was continued during the 11th century by the small local rulers of Spain who succeeded them and by the Almoravids, who united Morocco and Spain in one empire.

Islāmic gold coinage became one of the great currencies of the medieval world, and the dinar enjoyed great popularity on the western shores of the Mediterranean. It was referred to in Europe in earlier times under the name of *mancusus,* while the Almoravid dinar was known as *morabiti* (whence Spanish *maravedí*). The quarter dinars (known as taris) of the Fāṭimids, who ruled also in Sicily, were imitated in southern Italy and Sicily and by their Norman successors. Huge quantities of silver dirhams also reached eastern and northern Europe and especially (as a result of the fur trade) Scandinavia.

The Almohads, who succeeded the Almoravids in the 12th century, introduced a coinage that was new in both standard and form. Their fine gold dinars (2.3 grams) are among the most beautiful coins of the Muslim world; the dirham (1.5 grams) is square. The coinage of the Almohads survived also among their successors, well into the late Middle Ages, and was also widely current, and imitated, on the European shores of the Mediterranean.

In the east the successors of the Seljuqs (Artukids, Zangids, etc.), who, because of the scarcity of silver, issued large copper coins, introduced a striking innovation: they adopted types borrowed from ancient Greek and Roman, Sāsānian, and Byzantine sources. The Seljuqs of Asia Minor (12th–13th century) had silver coins showing a horseman with a mace over his shoulders, or a lion and sun. Farther east the Ghaznavids (10th–12th century), on their conquest of India, struck coins with Sanskrit inscriptions.

Mongol coinage

In the 13th century the Mongols swept through all Asia except India. The khans of the Golden Horde issued an extensive series of small silver coins (which influenced early Russian coinage). The Il-Khans of Persia struck large and handsome coins in all three metals. In the 14th century, Timur (Tamerlane) revived the power of the Mongols and struck silver and copper coins. His son Shahrukh introduced a new type of dirham, with, obverse, profession of the faith with the name of the first four caliphs on the margin and, on the reverse, his title.

Meanwhile, the new gold Venetian ducat spread in the East. It was used until the 18th century, and its standard (3.56 grams) was adopted for Islāmic coins.

*Ottoman Empire.* The original coinage of the Ottomans consisted of small silver coins (*akche,* called asper by Europeans). Gold coins were not struck before the end of the 15th century; before and after that century, foreign gold, mainly the Venetian ducat, was used. A notable Ottoman innovation was the *tughra,* an elaborate monogram formed of the sultan's name and titles, which occupies one side of the coin. Various European silver dollars also circulated extensively.

*Later Persia, Afghanistan, and Turkestan.* The earlier coins of the shahs of Persia were large, thin silver pieces of Central Asian style, but in the 18th century the coins became smaller and thicker, as in India. Legends were usually in rhyming couplets; gold was scarce until the 18th century. Cities issued copper with local emblems.

The emirs of Afghanistan, who became independent of Persia in the 18th century, struck gold and silver on the standard of the Mughal emperors, whose poetic inscriptions they also copied. Of the various smaller modern dynasties that ruled Central Asia until the Russian conquest, the emirs of Bukhara and of Khokand were notable for their extensive issues in gold. From the 19th century gradual westernization resulted in the adoption of European types.

**India.** *Ancient and early medieval.* India derived the idea of coinage from the Greek world via Iran. The earliest coins were weighed from pieces of stamped silver and were decorated with stylized depictions of animals and plants. These coins were soon augmented by copper ones, some made in the same way, others by casting. These pieces circulated over most of northern India during the 4th to 1st centuries BC. From the 1st century BC onward there were also copper coinages of numerous small states, tribes, and dynasties, which show increasing Greek influence. Their few silver coins were directly influenced by the hemidrachms of the Greek rulers of northwestern India of the 1st century BC.

Early in the 2nd century BC the Greeks of Bactria began

to invade India, and their coinage is remarkable for its fine series of portraits and for the number of names it records of rulers otherwise unknown. Prākrit legends began to appear alongside Greek and, as the Greek rulers were replaced by Central Asian invaders who copied their types, the Greek deities gradually gave place to local ones on the coins.

Kushān coinage

In the mid-1st century AD another group of Central Asian invaders, the Kushāns, founded a great empire in northwestern India; they left a wealth of gold and copper coins with legends in the Bactrian language, written in cursive Greek letters. Their coin types—of king on obverse and deity on reverse—became the general style of northern Indian coinage for the next 1,000 years. The type was continued by the kings of Kashmir to the 10th century and adopted, with modifications, by the great Gupta emperors in the 4th century. The Guptas struck an extensive gold coinage; among the more notable Gupta coins are those that commemorate Candra Gupta I's horse sacrifice or depict him as a lyrist.

In western India a dynasty of satraps of Persian origin had been ruling since the 1st century AD. Their extensive silver coinage is dated and therefore of unusual historical value. This kingdom was overthrown by the Guptas at the end of the 4th century, and they at once began to imitate this silver coinage locally. The Huns (Hephthalites), who destroyed the Gupta and other smaller states in northern India in the 6th century, left numerous coins, imitated from Sāsānian, Gupta, or Kushān prototypes. Copies of these continued to circulate in parts of northern India until the revival of various Hindu dynasties from the 10th century onward. A notable adaptation of a Hun design was

Gandharan silver

the neat silver coinage of the Shahis of Gandhara of the "bull and horseman" type in the 9th and 10th centuries, extensively imitated by the Muslim conquerors of India and the contemporary minor Hindu dynasties. The other type favoured by the medieval Hindu dynasties for their gold coinage was that of a seated goddess—going back to a Gupta reverse—and an inscription with the king's name on the other side.

The coinages of southern India form a class by themselves. In the later centuries BC and early AD, the Andhras ruled a great kingdom in central southern India; they issued coins mainly of lead but also of copper and silver with types based on Greek or local northern Indian designs.

The later medieval dynasties of southern India struck coinages mainly of gold, the type of which is usually the badge of the dynasty; the Cheras of Malabar, for example, had an elephant, the Chalukyas of the Deccan a boar, the Pandyas a fish, and the cup-shaped pieces of the Kadambas a lotus. The Chola dynasty introduced under northern influence the type of a king standing, on obverse, and, on the reverse, the king seated, which spread through southern India and was taken to Ceylon by the Chola conquest and adopted locally. The great Hindu kingdom of Vijayanagar (Mysore) left a large series of small gold and copper coins with types of various deities.

*Islāmic.* The earliest Arab invaders had reached India in the 8th century and founded a dynasty in Sind, which left numerous very small silver coins of the Umayyad type. The coinage of the Ghūrid dynasty of northwest Afghanistan and its successors from the 12th century onward is varied and extensive, mainly gold and silver tangas (or rupees) of 10.76 grams. Gold was hardly issued at all in the 15th and 16th centuries, and for a time the coinage was mainly billon. Shēr Shāh of Sūr (1540–45), of northern India, issued a large silver currency of a type carrying the profession of the faith and names of the four caliphs, that was imitated by the Mughal successor of the Sūrs.

The coinages of Bābur and Humāyūn, the first two of the Mughal conquerors of India, are not extensive and are of Central Asian character. With the next two emperors, Akbar and Jahāngīr, is found a series unrivaled for variety and, within limitations, beauty—the gold coins of Jahāngīr are noble examples of Muslim calligraphy. In

Mughal types

the 16th century the type that goes back to Shēr Shāh prevailed: the profession of the faith with the names of the first four caliphs and the emperor's titles on the other side; Aurangzeb replaced the confession of faith by the



Jaipur Nazarana silver rupee struck for presentation purposes by Isvari Singh (1743–60) in the name of the Mughal emperor 'Ālamgīr II. In the British Museum. Diameter 30 mm.

Reproduced with permission of the trustees of the British Museum; photograph, Ray Gardner for The Hamlyn Publishing Group Limited

mint and date, and this remained the usual type until the end of the dynasty. The emperor's name is usually enshrined in a Persian couplet to the effect that the metal of the coins acquires added lustre from bearing the emperor's name. Nearly 50 such verses are found on Jahāngīr's coins. His reign is also remarkable for the series of coins bearing signs of the zodiac and for the set of portrait mohurs, one of which represents him holding a wine cup. From the beginning of the 18th century the coins become stereotyped, and the epigraphy loses its beauty. The European East India companies copied the native types from the local coinages and did not strike on European lines until the 19th century. A uniform coinage for territories under British administration was introduced in 1835. The right of native states to mint their own coinage was gradually curtailed by the British government. Since 1948, India, Pakistan, and Sri Lanka have had their own coinages. Bangladesh commenced independent coinage on Jan. 1, 1972.

*Miscellaneous.* Mention should also be made of the extensive Nepalese coinage in gold and silver with Sanskrit legends; the coinage of Tibet, related to that of Nepal; and the long series of octagonal gold and silver coins of Assam, struck until *c.* 1821.

**China.** Before coins were invented, cowrie shells were used as money in China. The earliest Chinese coins are small bronze hoes and knives, copies of the tools that previously had been used for barter. The knife coins (*tao*) were about six inches (15 centimetres) long and some bore inscriptions naming the issuer and giving the value. Hoe coins bore similar inscriptions. Both types circulated during the 4th and 3rd centuries BC. Round money with a hole in the centre was issued about the mid-3rd century, but it was not until 221 BC that the reforming emperor Shih huang-ti (221–210/209 BC) superseded all other currencies by the issue of round coins (*pan-liang*) of half an ounce. (There were 24 grains in the Chinese ounce, and in the Han period the ounce weighed 16 grams.) These *pan-liang* coins were continued by the Han dynasty. The official weight of this coin was gradually reduced until it was replaced in 118 BC by the emperor Wu-ti's five-grain (*chu*) piece, which remained the standard coin of China for the next three centuries; a break in the monotony of the regular coinage occurred in the archaistic innovations of the usurper Wang Mang (AD 9–23), who issued a series of token coins based on the *tao* and on square Japanese *pu* coins and various new round coins.

After the Han period (206 BC–AD 220), the standard coin underwent many modifications. The coin was issued in iron and lead, in six-grain and four-grain weights, and in token versions. Yet the ideal of the five-grain coin of Han survived until the rise of the T'ang dynasty, when the

Kai-yuan coinage

emperor Kao-tsu in 621 issued the Kai-yuan coin, which gave the coinage of all the Far East its form until the end of the 19th century—a round coin with a square hole and a four-character legend stating the function (*tong-bao,* which means "circulating treasure") and date of the coin. The Southern Sung dynasty (1127–1279) dated their coins on the reverse with regnal years, and the T'ang and Ming dynasties (618–907 and 1368–1644, respectively) put the mint name on the reverse, as did the Ch'ing dynasty (1644–1911/12), this last giving it in Manchu characters.



Bronze token coin designed by Emperor Hui-tsung, Northern Sung dynasty, 1107. In the British Museum. Diameter 41 mm.

Reproduced with permission of the trustees of the British Museum; photograph, Ray Gardner for The Hamlyn Publishing Group Limited

Paper money has been in use in China since the 9th century and was current almost to the exclusion of regular coins under certain Mongol emperors, such as Kublai Khan, whose paper money is described by Marco Polo. For more than 2,000 years the copper cash was the only official coinage of China; gold and silver were current by weight only, the latter in the form of ingots. As a result of the popularity of imported Spanish colonial and Mexican dollars, several attempts were made to institute a silver coinage based on the dollar in the 19th century; not until the very end of the 19th century were mints established to strike silver and copper coins of European style. Under the republic, coins were at once struck with the portraits of Sun Yat-sen and Pres. Yüan Shih-k'ai, and the various generals who fought for control of China issued their own coins. The currency of both the People's Republic of China and Taiwan is the yuan (dollar). The very extensive series of talismans, coinlike in shape but usually larger and in their legends and types reflecting popular Chinese religious thought, is noteworthy.

**Japan.** The art of coinage was borrowed from China by Japan, whose first bronze coins were issued in AD 708. To the mid-10th century, 12 different issues were made, each of a different reign. For the next 600 years, however, no government coins were issued, and grain and cloth were used as money. From the Middle Ages imported Chinese coins began to circulate along with locally minted imitations. In 1624 the copper *kwan-ei* was first issued and remained in vast variety the usual issue for more than two centuries. The *ei-raku* and *bun-kyū sen* of the 19th century were the only other regular copper coins. Unlike China, Japan has had a gold and silver coinage since the 16th century. The gold coins are large flat pieces in the shape of rectangles with rounded corners, the largest size being *ōban* and the smaller *koban*. Other gold pieces are the small rectangular pieces of one and two *bu* issued from time to time; round gold is rare and usually of provincial mints. Silver was originally in the form of stamped bars called long silver; these were supplemented by small lumps, also stamped, called bean silver. They were later augmented by issues of silver pieces in the same shape as the small rectangular gold coins.

The Tokyo mint

In 1869 a mint on European lines was established in Tokyo, and gold, silver (yen or dollars), and copper were regularly issued from it until World War II, when nickel and various alloys superseded the precious metals. After World War II the yen was retained as the unit of currency. The *e sen* of Japan are not coins but amulets.

**Korea.** The earliest coins found in Korea were Chinese knife coins of the 3rd century BC. The local production of coins did not begin until the 9th to 10th century AD, when copies of contemporary Chinese Kai-yuan coins were made. Coins with local inscriptions, still based on the Chinese model, were issued from the 12th century. Chinese-style coins continued to be used until Japanese and Russian influence led to the introduction of Western-style coinage in the late 19th century.

**Vietnam, Kampuchea (Cambodia), Laos.** Nam Viet (present-day Vietnam) began by imitating Chinese coins and had a regular bronze coinage of its own on the Chinese model from the 10th to the 19th century. Silver became common in the 19th century in the form of narrow oblong bars. Presentation pieces in gold, silver, and copper were created in a variety of designs bearing, for example, auspicious inscriptions and quotations from the Chinese Classics, in addition to the king's name. The native coinage continued until World War II but had largely been replaced by French colonial issues. After independence from France, Vietnam substantially retained the Western alphabet on its often very attractive coinage. Cambodia (Kampuchea) had its own coinage from the 15th century—curious uniface round pieces decorated with simple religious pictorial designs. Western-style coinage began to replace these from the mid-19th century. Separate coinages subsequently were in circulation in Cambodia and Laos, as they were in North and South Vietnam during the 1945–76 partition.

**Burma and Thailand.** The earliest coinages of Southeast Asia were issued in Burma and Thailand during the late 1st millennium BC. They were derived from Indian prototypes (examples of them have also been found in Cambodia and Vietnam). From as early as the 17th century Thailand struck gold and silver coins in the form of balls made by doubling in the ends of a short, thick bar of silver and bearing the stamp of the reigning monarch ("bullet" coins). After *c.* 1860 it had a coinage on European lines with issues in gold, silver, tin and copper, and later nickel. (J.Al./S.M.Sn./Ed.)

## COINS OF AFRICA

The Aksumite kings, powerful rulers of a kingdom in northern Ethiopia from the 2nd to the 9th century AD, and Christian from the 4th century, issued small gold coins, with a little bronze and very rare silver, from the 3rd century onward; the initially Greek inscriptions were replaced ultimately by Amharic. Indigenous coinage lapsed in the 10th century, the country becoming dependent on imported currencies, of which the silver Maria Theresa thalers of Austria were conspicuous from the 18th century onward. National coinage was resumed by King Menelik II, emperor of Ethiopia (1889–1913) with silver coins called talaris and their fractions and subsidiary copper, showing the Lion of Judah reverse—an allusion to the tradition that Menelik I had been the son of King Solomon and the Queen of Sheba. Some gold came later, to be continued by Emperor Haile Selassie (1930–36), who coined also in nickel and bronze until the Italian occupation and after his restoration in 1941. A national coinage continued after he was deposed in 1974.

Ethiopia

**North Africa.** Elsewhere the 19th-century partition of Africa by colonial powers led to a great miscellany of currencies before decolonization and independence were achieved from the mid-20th century. Egypt, gaining independence from the Ottoman Empire in 1914, based its currency on the piastre, with Arabic inscriptions; some gold and silver multiples were produced. Under Fu'ād I (1922–36) and Farouk I (1936–52) the royal portrait was used. The subsequent republic, with its piastres of aluminum–bronze alloy accompanied by rare silver and even rarer gold, has often chosen types referring to national history (*e.g.*, the Great Sphinx, Ramses II, the Aswān High Dam).

The piastre became the unit of Libya, which, after a period as an Italian colony, briefly became a kingdom under Idris I (1951–69), with a fine portrait coinage, before the regime of Col. Muammar al-Qaddafi. The piastre was also the unit of the French protectorate of Tunisia until 1891, when a coinage of francs and centimes was introduced. Independence from France in 1956 brought Arabic inscriptions. The piastre was also adopted in 1956 as the unit of the new republic of The Sudan. In Morocco, however, which was an early 20th-century protectorate of France, the unit was the Arabic silver dirham, replaced in 1902 by the silver rial until the introduction of the franc in 1921.

**Sub-Saharan Africa.** Further south the various regional currencies grew out of the 19th-century European colo-

nization. Thus Ghana, before independence in 1957, had been the British colony of the Gold Coast, in which the British denominations of shilling and penny were traditionally used; special gold was coined to mark the declaration of a republic in 1960. Similar developments took place in the British colonies of East Africa and in the colonial territories of Northern and Southern Rhodesia (later independent as Zambia and Zimbabwe), Nyasaland (later Malawi), and Nigeria. The currency of Liberia (founded by former American black slaves) from the mid-19th century consisted mainly of copper or bronze, with an elephant displacing the head of Liberty, of U.S. type.

In South Africa, before the Union was established in 1910, the only coinage of note was that of the South African Republic. During South Africa's membership in the Commonwealth its currency was assimilated to that of Great Britain, but when South Africa left the Commonwealth in 1961 it established a new system based on the gold rand.

Coinage for the French colonies such as the Cameroons, French West Africa and Equatorial Africa, Madagascar, and French Togoland, showing the French cockerel or the head of "Marianne" (emblem of the spirit of the French Revolution), was in general more standardized than in the British colonies. The principal Portuguese colonies were Angola and Mozambique; the former used macutas (equal to 50 reis) of copper, followed by centavos and silver escudos, while copper reis were current in the latter, followed by escudos or centavos. In the Belgian Congo (Zaire since 1971), originally established as a free state by King Leopold II in 1885, currency was based on silver francs and copper centimes. (C.H.V.S.)

## MEDALS AND MEDALLIC ART

**Italy.** The modern commemorative medal, in both form and content, was invented by the Italian painter Antonio Pisano (*c.* 1395–1455), called Pisanello. His first medal portrayed the Byzantine emperor John VIII Palaeologus and was made in 1438–39. His medals provided a portable portrait relief of the sitters, reproducible by casting in lead or bronze and small enough to be held in the hand. He placed a profile portrait on the obverse and an allegorical or pictorial scene on the reverse. This formula for the medal has lasted to the present day. Pisanello made medals of 16 sitters for the courts of Ferrara, Mantua, Milan, Naples, and Rimini. Major schools of medal making developed, particularly in Mantua, Florence, the Veneto, and Rome. The papal court had no local school but attracted medalists from all over Italy. Toward the end of the century the portrait effigy became bolder and more sculptural in the work of Niccolò Fiorentino and Sperandio of Mantua.

Pisanello

During the 16th century in Italy the cast medal continued in favour, and Leone Leoni (1509–90) of Milan and Pier Paolo Galeotti were its principal masters. Leoni was engraver at the papal mint in Rome from 1537 to 1540, Master of the Habsburg mint at Milan (1542–45, 1550–59), and court sculptor to Charles V. His most masterly cast medal is of Michelangelo (1561). He also produced struck portrait medals, like those of the Genoese statesman and admiral Andrea Doria. For the first time the struck medal became a common instrument of court propaganda, especially for the popes and for the ruling Medici family in Florence. Galeotti made more than 80 cast portrait medals, which rival the work of Leoni. Pastorino da Siena produced a long series of portraits of sitters of lesser rank, cast in lead without reverse type. The finest struck portraits were the work of the medalists Domenico di Polo and Domenico Poggini in Florence and Giovanni Bernardi, Alessandro Cesati, and Benvenuto Cellini at the papal court. Antonio Abondio drew his style from Leoni and from the charming Mannerist portrait medalists of Reggio nell'Emilia, particularly Alfonso Ruspagiari.

**France.** The earliest French medals were heraldic pieces struck in gold and silver, *c.* 1455, to commemorate the expulsion of the English. The first portrait medal was a struck gold presentation piece of Charles VIII and Anne of Britanny, made by local goldsmiths for a visit to Lyon in 1494. Italian medalists had worked in France and directly inspired the work of Jacques Gauvain and Jérôme Henry at Lyon. In 1550 mint officials were sent by Henry II to seek out and obtain German minting machinery, and in consequence numerous propaganda medals were produced, ascribed to the Huguenot goldsmith Étienne Delaune and to Claude de Héry. With the appointment in 1572 of the great Mannerist sculptor Germain Pilon (1535–90) by Charles IX to the new office of "contrôleur général des effigies," a new form of medal appeared. Pilon produced a superb series of large cast portrait plaques for members of the Valois dynasty and a series of struck medals for Henry III. For Henry IV the Danfrie family produced a series of struck medals. Jean Warin (1604–72) also made elegant cast pieces, and between 1636 and 1670 he held almost a monopoly of the production of struck pieces for the court. Guillaume Dupré (1574–1647) followed Pilon, charmed Henry IV with his portrait medals, and was appointed in 1604 "conducteur et contrôleur général" of the Paris Mint. Nicolas Briot (1579–1646), rival of Dupré, was a lesser master who was a skilled mechanic and engraver general at the Paris Mint from 1600. In 1625 he went to London, where he revived the English court's interest in the medal.

**Germany and Austria.** The free imperial cities under the Holy Roman Empire were important centres of patronage, and the sitters were proud burghers depicted in a realistic idiom. A few fine medals are ascribed to Albrecht Dürer, but the first professional medalist was Hans Schwarz of Augsburg, active in Germany and elsewhere between 1512 and 1532. Christoph Weiditz produced numerous Augsburg medals and with Schwarz showed the greatest sensitivity in capturing individual character in his portraits. Friedrich Hagenauer, active in Munich and in Augsburg (1527–32), produced more than 230 medals. In Nürnberg, Matthes Gebel (active 1525–54) and his follower Joachim Deschler (active 1540–69) were the principal medalists. Ludwig Neufahrer worked mainly in Nürnberg and the Austrian Habsburg domains, employed by Ferdinand I from 1545. The Italian expatriate medalist Abondio was called to Vienna and also appointed court medalist by Emperor Maximilian II in Prague in 1566.

**The Netherlands.** The famous medal of Erasmus of 1519, by Quentin Massys, made in Antwerp, is the grandest northern Renaissance medal, but it had no progeny. Of the regular professional medalists some, like Steven van Herwyck (*c.* 1530–67) and Jacob Jonghelinck (1530–1606), who worked in Italy for Leoni, adopted the Italian style, somewhat more idealized than the German. The war with Spain (1568–1648) stimulated the production of propaganda medals, which became a popular vehicle of nationalist sentiment. The Netherlands' tradition of silversmithing was also adapted to the medal. Highly skilled engraved portraits on thin plates of gold and silver were made. The masterpiece of this genre is a portrait of Elizabeth I of England's favourite, Robert Dudley, earl of Leicester, in 1586, engraved on gold by the Dutch Mannerist engraver and painter Hendrik Goltzius (1558–1617). Simon van de Passe produced similar work and went to London, where he created a series of Tudor and Stuart portraits.

Massys' Erasmus medal

**The Baroque period.** The large struck propaganda medal was issued widely in northern Europe in the 17th century. The Thirty Years' War and later the Dutch wars with France and England stimulated such issues. Sebastian Dadler (1586–1657) was employed by the courts of Saxony, Sweden, Poland, and the Holy Roman Empire to produce large struck medals on the political events of the time. The Swiss Johann Carl Hedlinger (1691–1771) was trained in Paris, became court medalist in Stockholm, and produced numerous historical medals on commission. His portraits are the most elegant and individualistic effigies of the 18th century. The European medal was dominated by the court style of Versailles. The grand propaganda series of the *Histoire métallique,* a series of medals struck to commemorate Louis XIV's reign, was envied and imitated throughout Europe, though the Dutch copied it in a manner calculated to ridicule the French. The technical excellence of the Paris Mint was also imitated. The first fully Baroque medal in England, on the Restoration of

The *Histoire métallique*

King Charles II in 1660, was made by the Paris-trained medalist John Roettier, in the full French court style.

Caspar Gottlieb Lauffer of Nürnberg from 1679 issued a large number of medals engraved by numerous artists and commemorating contemporary events. He eventually published a catalog, in 1742, entitled *Das Laufferische Medaillen-Cabinet.*

The cast medal continued to be made. In Italy, the Tuscan sculptor, scholar, courtier, and mint master Massimiliano Soldani-Benzi (1656–1740) revived the cast portrait medal in 1677 and founded a school with his pupils Antonio Selvi (1679–1753) and Lorenzo Maria Weber (1697–1774). The school lasted until the 1740s. In Rome, the few cast medals included works by Charles-Jean-François Chéron (1635–98) and by Gioacchino Francesco Travani (active 1634–75), after designs by the great Italian sculptor of the Baroque, Gian Lorenzo Bernini. The Dutch silversmiths invented chased medals made of shells of silver hammered into relief from the reverse (repoussé) and soldered to a rim, the work by Pieter van Abeele (1608–84) being particularly charming.

In England, Thomas and Abraham Simon produced cast portrait medals of great refinement in a northern European realistic tradition. The cast portrait plaque was revived by the Romantic sculptor Pierre-Jean David d'Angers (1789–1856) in his series of portraits forming a *Galérie des contemporaines,* begun in 1827. The Paris school of the late 18th century, especially the work of Benjamin Duvivier (1728–1819) for King Louis XVI, combined Rococo elegance with realism. Duvivier's work included commissions from the U.S. Congress. The Napoleonic regime ordered an elaborate *Histoire métallique.* The Duvivier era saw the introduction of steam-powered presses for coin and medal making, perfected by the English engineer Matthew Boulton at Birmingham in 1786, and the use of the reducing machine, which permitted the translation of a sculptor's large-scale relief model into a working die (see below *Techniques of production*). This invention was crucial to the development of a new Parisian school of the Art Nouveau, founded by Jules-Clément Chaplain (1839–1909) and Louis Oscar Roty (1846–1911).

The reducing machine

A rival and similar school developed in Vienna and spread in Hungary and Bohemia. Britain was touched by the missionary zeal for the Art Nouveau style shown by Alphonse Legros (1837–1911), and a few sculptors, most importantly Alfred Gilbert (1854–1934), took up medal making. Frank Bowcher (1864–1938) studied under Legros in Paris, where he produced both struck and cast medals. He became engraver at the Royal Mint, London. In the United States, Augustus Saint-Gaudens (1848–1907) produced admirable medals and portrait plaques in the same Art Nouveau style.

The German reaction to the power of the Parisian school produced a school of expressionist medalists, while in France the 1920s saw the beginning of the Art Deco medal. After World War II the artistic medal continued to show remarkable possibilities as a medium for portraiture. Wit, imagination, and experiments with form brought the medal to resemble the plaquette or ornamental tablet. The variety and originality can be seen through the biennial exhibitions of the Fédération Internationale de la Médaille. (J.G.P.)

## TECHNIQUES OF PRODUCTION

The essential advantage of using metals for currency, apart from durability, is that they can be shaped by melting and casting. Casting, therefore, has always been an integral part of the coin manufacturing process. Indeed, in some instances, it has been the only part. In early China bronze was cast into the form of the hoes and knives originally used for payment, and up to the 19th century the objects called "cash," with their square central holes, were also cast. Similarly, the first Roman issues, aes grave (heavy bronze), were ponderous cast pieces, the heaviest actually corresponding in weight to the libra, the Roman pound. However, as soon as the state realized that it could make a profit from issuing coins by decreeing that their value in the market should be greater than the intrinsic value of their metal content, casting—so simple an operation—at once led to counterfeiting. Provided the mold was made from an official coin, there was no straightforward visual way of distinguishing true from false. For this reason, casting alone seems not to have been employed for precious metal currency.

Casting and counterfeiting

The prototype coinage of Greek Ionia on the west central coast of Anatolia in the 7th century BC consisted of pellets of electrum (a gold–silver alloy) made by pouring the molten metal onto the striated surface of an anvil, where, under the action of surface tension, they assumed a characteristic lenslike shape before solidification. The weight of the pellets was checked and confirmed for use by stamping them with a punch—a naillike piece of metal, probably of bronze or iron. The punch sometimes had a crudely fractured end surface (which, of course, would be unique), sometimes an engraved design (the latter produced on the punch by drilling with abrasive corundum dust, which ate away at the surface, as in the lapping process perfected over millennia for sealstones). A counterfeiter would have had great difficulty in simulating the exact form of the punch surface. Within a short time, the issuing authorities began putting their own validating punch marks on the pellets, thus producing coins as opposed to bullion. In India before the Mauryan Empire (*c.* 321–185 BC), currency consisted of small silver bars carrying as many as six marks; the boat-shaped silver emanating from Southeast Asia in the 19th century was also officially confirmed in this fashion.

**Ancient minting.** Most of the ancient dies that have survived are of bronze, although iron dies are thought to have been widely used also. Lower dies seem generally to have been disk-shaped so that they could sit in a recess on an anvil. In some instances the design may have been cut directly on the anvil. Engraving of the details was carried out using small steel tools (scorpers), or designs were drilled out using corundum dust. It is possible that major elements of the design were inserted by a "hub," or master punch stamped into the die, but not all scholars accept that this method was employed in antiquity.

Making dies

Blanks or planchets (*i.e.,* the small metal disks from which coins are made) seem first to have been cast by pouring the molten alloy from a crucible onto a flat surface, where they cooled into the characteristic lens shape. Later the metal was poured into molds, which sometimes consisted of two parts so that the metal was completely enclosed; traces of the "flash," or joining line, can still be seen on surviving coins. At Alexandria in the Ptolemaic period (323–30 BC), open molds were common; in these a sequence of disk-shaped impressions in the mold were connected by channels, and a number of blanks were thus obtained at one pouring. The upper surface of the blank, where slag and oxide accumulated, had to be "turned" off, or drilled out, presumably by a tool like a carpenter's bit, and the centre punch mark to accommodate the tool point is characteristic of Ptolemaic, Seleucid, and Greek imperial coins. Contemporary issues in India were often square in outline and were cut by chisel from metal sheets. Many Greek and Roman silver coins were plated; an envelope of silver sheet was soldered on a copper core, and it is by no means certain that all such specimens were the work of counterfeiters, since solid silver and plated coins sometimes appear to have been struck from the same dies. In the later Roman Empire (3rd century AD) silver issues were heavily debased with copper; prior to striking, the blanks were immersed in an acid bath that leached out the surface copper to expose more silver, giving a much more acceptable appearance to the coins when they were first issued.

Striking—the impression of the die designs on the blanks—was startlingly simple. The lower die, set in the anvil, was covered by the blank; the upper die, which was positioned above, was then given one or more hammer blows (see Figure 1). A two-pound hammer, wielded by one hand, could easily give a force at the die face of seven tons. To get the high relief typical of Greek issues, two or three blows were necessary, and often there is evidence of double-striking on the coins. However, by preheating the blank, as practiced in Athens in the 5th century BC, less force was required and die life was extended. Analysis

Striking

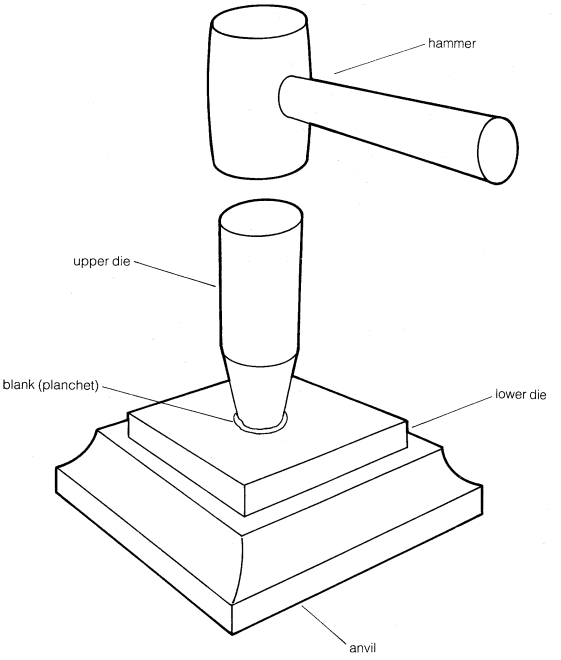



Figure 1: Striking, an ancient form of minting.

of the documentary evidence implies that one obverse (lower) die could produce upward of 20,000 coins, while 10,000 coins have been struck from a simulated bronze die without significant deterioration of the working surface. Receiving the hammer blows more directly, the reverse (upper) dies enjoyed about half the life of the obverse. Production rates varied. In small mints, operated by one man, a rate of 100 coins per hour has been shown to have been feasible. At important centres such as Rome or Antioch teams of four probably operated. An eyewitness account of a Persian mint in the 1870s describes how, with a hammerer, a die holder, a blank placer, and a coin remover, one piece could be struck about every two seconds.

Rate of production

**Medieval minting.** Alterations in the flan (the coin disk, a term deriving from the French *flatir,* "to beat flat") led to corresponding changes in the manufacture of dies. In about AD 220 the Sāsānian dynasty of Iran introduced the concept of thin flan coins, issues that were struck in relief on both sides. In order not to produce intolerable stresses in the dies, since the thinner the material the more force necessary to make it flow into the recesses of the die's design, the depth of relief on such coins was of necessity much shallower than with earlier currency. Such techniques spread by way of Byzantium to northern Europe, where the emperor Charlemagne (*c.* 742–814) struck thin flan deniers (small silver coins), or pennies, which became characteristic of both his own and neighbouring kingdoms.

The Franks and Saxons inherited an art that was formalistic rather than realistic, and this permitted their coin designs to be made up from a small number of standard elements that were reproducible using punches. It has been shown, for example, that the complete dies for all of the coin types of Edward the Confessor of England could be obtained from seven punches, giving individual wedges, crescents, pellets, and bars, each of which was independently struck to make up a legend and design. Consequently, a single workshop could supply the 70 or so contemporary English mints in a relatively short time. An experimental pair of dies took less than an hour to fabricate. Of course, many European dies were produced by a combination of punching and engraving, while engraving alone was typical of early Islāmic and contemporary Oriental dies. With the advent of larger denominations, such as the gros tournois (based on the weight standard of Tours, in France) in the 13th century, more florid designs came to be preferred, but the elements of the royal crown, for example, or the letters of the legend were still punched into the dies. To judge from surviving specimens, both upper and lower dies (trussells and piles) were by then produced from wrought iron. While the upper die retained the cylindrical shape of antiquity, the lower die was tanged (having protrusions added) so that it could be wedged into a wooden block.

Design elements

The thin silver sheet required for the new coins needed to be beaten out from its cast state, and this in turn necessitated annealing (strengthening by slow cooling) to prevent cracking. By the 10th century, squares of sheet, somewhat larger than the eventual penny, were being struck between square dies and then separated by a circular cutter. A few imitative coins on square flans are known from Scandinavia, while die identical coins have exactly matching edge irregularities, proving use of the same cutter. With the introduction of the gros tournois, the blanks were cut roughly circular with shears, then gripped by special tongs in rouleaux (columnar rolls) of a dozen or more, and finally hammered into circularity on a flat anvil. Alternatively, the silver was cast into thin rods of rectangular section; pieces of the correct length (and hence weight) were next cut from the rod by chisel and then, with several annealings, were beaten to the appropriate thickness, before being rounded and struck by a die. Blanching (cleaning) of the blanks by an acid dip was necessary before striking to produce an acceptable surface if oxidation had occurred during annealing.

Blanching

Striking was carried out in much the same way as during antiquity, although contemporary illustrations indicate that only one operator, not a team, was employed. Twelfth-century Byzantine coins were often cup shaped. A full impression of the curved dies could not easily be obtained by one blow; hence there evolved a method of striking one half of the coin with a slightly inclined upper die, which was then rocked over to the other side for a second blow. Bracteates, issues of foil thickness, were common in 12th-century Germany. To make these, a single die was used to strike a column of several blanks resting on a piece of leather, so that the reverse of each was the incuse (hollowed impression) of the obverse. Die life in general was higher than during antiquity, and documentary evidence for 13th-century English pence indicates perhaps 30,000 coins per obverse. There is no evidence for production rates.

**Early modern minting.** The increase of mining activity in central Europe during the 15th century gave a great impetus to the development of modern minting processes. The dies themselves were still made by punches, but these, in turn, had become much more sophisticated, often embodying a complete portrait of the monarch. Their general shape depended on the striking process employed, but the material used was a steel that could be hardened by carburizing (putting iron in a bed of carbon in a sealed air-tight box, and thence into a furnace, where the carbon diffused into the outer layers) after the designs had been punched in, or sunk.

The metal for the coins was cast as ingots, a typical size being $\frac{1}{2} \times 1 \times 20$ inches. These were then passed between steel rollers, powered by a water mill or horse gin (a mechanism that translated horsepower into rotational energy), to reduce the thickness. Several passes and annealings were necessary to obtain the correct thickness. The blanks, particularly for the larger crown-sized coins being introduced, continued to be roughly cut with shears from the rolled fillet (metal strip), so that, as previously, they could easily be adjusted for weight before being rounded in rouleaux. In some cases, however, they were punched by a machine from the fillet to a fixed diameter, so that the thickness was critical for controlling weight. To protect against clipping, during the next century a security edge was sometimes rolled onto the blank; this might consist of an inscription or a serrated or milled edge.

Security edging

Hand-operated screw presses were developed for stamping the designs on the blanks; although the blanks originally were centred on the lower die by eye, it soon became clear that a locating collar would prevent off-centre striking. Such a method was used by Benvenuto Cellini, who struck coins for Italian princes in the first half of the 16th century, and it was then introduced first to Paris and then to London in the 1550s.

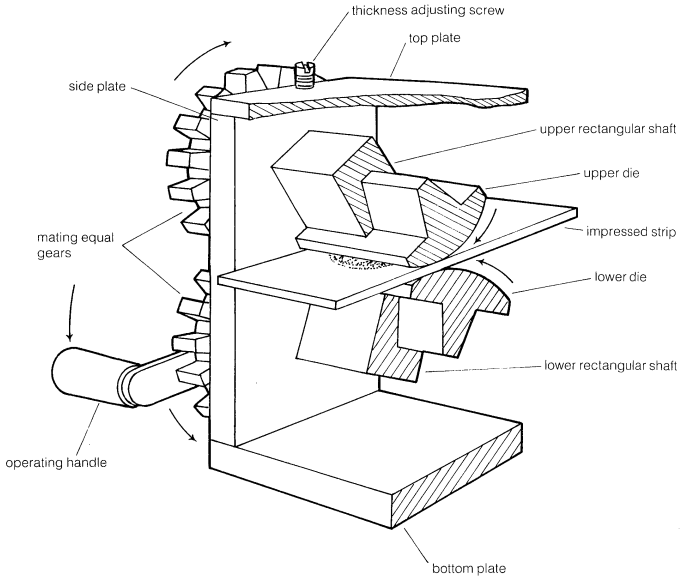



Figure 2: The *Taschenwerke* minting machine.

The roller press

At the same time the roller press was under development in Germany. Initially, the die designs were engraved or punched into the curved surfaces of two rollers that were geared together so that the whole fillet (rather than single blanks) could be fed between them and emerge impressed. This method was advantageous in requiring less power: only part of the blank was being deformed at any one time, and so, as compared with the screw press, the stresses on the machine were reduced. Because of imperfections, the fillet and the finished coins as punched out were markedly curved, and the coins required flattening (planishing) by light tapping with a smooth-faced hammer.

The difficulty of taking out the complete rollers from such a press led to an ingenious variation—the *Taschenwerke* (see Figure 2). In this machine the rollers were replaced by rectangular shafts pierced in the middle to take a pair of dies with tapered extensions (tangs). The axis of the upper shaft could be raised or lowered a short distance to accommodate variations in the dies or differing coin thicknesses. Such machines continued to be used in Germany into the 18th century.

The rocker press represents another variation. The bottom roller (actually a quadrant insert, as in the *Taschenwerke*) remained stationary; the axis of the upper roller rotated about this lower axis as a small circle around a larger, so that the upper die face rolled over a stationary fillet that had been positioned over the lower die. One such mechanism, now in the British Museum, produced minor copper coins in Spain soon after 1600.

The fly press

The prolific Jean Warin, one of the great engravers, finally established the use of the fly press, a variation on the screw press in which the helix angle of the screw was much increased. The rotational arms ended in heavy weights that were swung with great velocity by two operators (working for only 20 minutes in each hour), and the elasticity of the system caused a rebound of the arm to its original position after the coin had been struck. Again, with a team of moneyers, a rate of production of a coin every second or so could be achieved. In some Russian mints of about 1800 a guided dropping weight functioned in much the same way, regaining its original position partly by rebounding and partly by operators pulling on return ropes running over pulleys.

**Contemporary mints.** In the 1770s the steam engines of Matthew Boulton and the Scottish inventor James Watt made available new sources of power that were soon adapted to the coining process. Initially used to strike commercial tokens, these methods were eventually taken up by the Royal Mint in London. Experiments produced new steels that could cope with the much higher stresses involved, while a French invention, the pantograph, or reducing machine, permitted the manufacture of a standardized design for every denomination, all being reproduced identically but to differing scales.

Modern die manufacture

In modern minting, the sequence of die manufacture is as follows. A plaster model of the proposed design, about one foot (0.3 metre) in diameter, is received from the artist and a mold is made; from this is obtained an electrotype copy in nickel and copper. Mounted in the reducing machine, the copy permits the cutting of the design to the appropriate coin size in a block of steel, the master punch, which has features in relief, as on the eventual coin. The master is then used to punch-in, or sink, a matrix; this raises a working punch, which is used to sink a working die. Imperfections at any stage are removed by hand tooling and, for best results, the surface of the working die is highly polished before it is sunk.

The U.S. Mint

The production of blanks (called planchets in the United States) is highly automated. At the U.S. Mint in Philadelphia, for example, the incoming metal is assayed to ensure that it is of the correct specification. After being sheared into small pieces, the metal, together with the waste clippings (scissel) from previous blanking operations, is conveyed to a computer-controlled weighing section, where a charging bucket is filled with the correct proportions of each constituent of the required alloy; the metal then goes into a 15,000-pound-capacity electric induction melting furnace. During the melting, deoxidation additives are introduced, and the furnace is then tilted to pour the metal into a water-cooled, semicontinuous casting machine mold with a movable bottom. The resulting vertical ingot, with dimensions of 16 inches by 6 inches by 18 feet and a weight of 6,600 pounds, is set on a roller conveyer in a horizontal position preparatory to being cut by a rotary saw into two equal lengths. The bars are next raised in temperature in a high-frequency induction coil so that they can be hot-rolled. After nine passes the thickness of the bars is reduced to less than one-half inch, and the length is extended to approximately 115 feet. Quenching with water is followed by skim milling in order to remove the oxide layer on the top and the bottom surfaces. The coils of strip metal are next cold-rolled, reducing the thickness to about one-tenth of an inch. The ends of the individual coils are then trimmed and welded together, giving a large coil weighing some 4½ tons. Finally, the coil is rolled under tension in a finishing mill, where the thickness is controlled by sensors.

The blanking presses are typically high speed, punching out from the coils as many as 21 planchets per stroke at 100 or more strokes per minute, the scissel being returned to the melting pot. For coins of small denomination the planchets are then fed into an annealing furnace, quenched with water, cleaned with acid, washed, and dried. The subsequent operation is edge-up setting, the partial formation of a protective rim by forcing the blank into too small a hole. The planchets then proceed to the coining presses, many of which are adapted to cope with four coins at one blow. The spread of the metal under the force of the die is confined by a collar, and the radial recovery of the metal as the load is removed prevents its adherence to the collar, as the latter is retracted below die level. The struck coins are taken to a checking point, after which they are counted and bagged, ready for distribution. Die life is upwards of 200,000 coins.

Edge-up setting

The Royal Mint

The procedures at the United Kingdom's Royal Mint, at Llantrisant, Wales, are analogous. There, however, the ingots are cast continuously, not in discrete lengths, and they are subsequently sawed for the rolling operations. The edge-up setting is sometimes combined with the impression of a channeled security edge of the type found on Indian issues. To assist metal flow during striking, the washing and lubricating of the blanks are combined. The production of seven-sided coins (20- and 50-pence denominations) from circular blanks indicates the extent of the flow of the blank metal within the collar; the striking presses are capable of 600 strokes per minute. The counting and bagging operation is performed by robots. An experiment in obtaining the correct weight of gold issues, where the "remedy," or tolerance (permitted range of variation in the standard), is very limited, showed that when the blanks were punched from fillets that were one-third

inch thick and then pressed out to the normal thickness of one-tenth of an inch, the same error in initial thickness had less eventual effect on the weight of a one-third-inch blank than on a one-tenth-inch blank. In this experiment the same solution to obtaining correct weight was applied as in the medieval period, when the use of easily measurable lengths of thin silver rods gave the correct weight per penny. (D.G.J.S.)

**BIBLIOGRAPHY.** RICHARD G. DOTY, *The Macmillan Encyclopedic Dictionary of Numismatics* (1982); and EWALD JUNGE, *World Coin Encyclopedia* (1984), are informative and well-illustrated reference sources. PERCY GARDNER, *A History of Ancient Coinage, 700–300 B.C.* (1919, reprinted 1974), is an exhaustive study of the development of early measures of value, the origin of coin standards, and the mutual relationship of precious metals. The general history of coins and coinage is given in GEORGE MACDONALD, *The Evolution of Coinage* (1916, reprinted 1980); THOMAS W. BECKER, *The Coin Makers,* rev. ed. (1970); LIONEL CASSON and MARTIN PRICE (eds.), *Coins, Culture, and History in the Ancient World* (1981); SIR JOHN CRAIG, *The Mint: A History of the London Mint from A.D. 287 to 1948* (1953); R.A.G. CARSON, *Coins Ancient, Mediaeval & Modern,* 2nd ed. (1970; U.S. title, *Coins of the World*); MARTIN PRICE and BLUMA L. TRELL, *Coins and Their Cities: Architecture on the Ancient Coins of Greece, Rome, and Palestine* (1977); MARTIN PRICE (general ed.), *Coins: An Illustrated Survey, 650 BC to the Present Day* (1980); and GERALD HOBERMAN, *The Art of Coins and Their Photography* (1982).

Topical treatments can be found in J.G. MILNE, *Greek Coinage* (1931); COLIN M. KRAAY, *Archaic and Classical Greek Coins* (1976); BARCLAY V. HEAD, *Historia Numorum: A Manual of Greek Numismatics,* new ed. (1911, reprinted 1983); P.D. WHITTING, *Byzantine Coins* (1973); PHILIP GRIERSON, *Byzantine Coins* (1982); C.H.V. SUTHERLAND, *Roman Coins* (1974); HAROLD MATTINGLY, *Roman Coins from the Earliest Times to the Fall of the Western Empire,* 2nd ed. (1960, reprinted 1977); ARTHUR ENGEL and RAYMOND SERRURE, *Traité de numismatique du moyen âge,* 3 vol. (1891–1905, reprinted 1977); and GEORGE C. BROOKE, *English Coins from the Seventh Century to the Present Day,* 3rd rev. ed. (1950, reprinted 1976).

Most of the above mentioned works include bibliographies. For a separate bibliographic treatment, see PHILIP GRIERSON, *Coins and Medals: A Select Bibliography* (1954); E.E. CLAINSTEFANELLI, *A Numismatic Bibliography* (1985); and the comprehensive annual bibliography provided in *Numismatic Literature* (semiannual), published by the American Numismatic Society of New York.

Catalogs and guides include WILLIAM D. CRAIG, *Coins of the World, 1750–1850* (1960); RICHARD S. YEOMAN, *A Catalog of Modern World Coins,* 10th ed. (1972); CHESTER L. KRAUSE and CLIFFORD MISHLER, *Standard Catalog of World Coins,* 7th ed. (1980); L.V. WRIGHT, *Colonial and Commonwealth Coins* (1959); and RICHARD S. YEOMAN, *A Guide Book of United States Coins, 1986: Fully Illustrated Catalog and Valuation List 1616 to Date,* 39th rev. ed., edited by KENNETH BRESSETT (1985), which is updated annually.

General works on medals and medallic art include JEAN BABELON, *La Médaille et les médailleurs* (1927); MAX BERNHART, *Medaillen und Plaketten,* 3rd ed., edited by TYLL KROHA (1966); SIR GEORGE HILL, *Medals of the Renaissance,* rev. ed., edited by GRAHAM POLLARD (1978); MARK JONES, *The Art of the Medal* (1979); GRAHAM POLLARD and GIUSEPPE MAURI MORI, *Medaglie e monete* (1981); and L. FORRER (comp.), *Biographical Dictionary of Medallists: Coin, Gem, and Seal-Engravers, Mint-Masters, &c., Ancient and Modern,* 8 vol. (1902–30, reprinted 1970). For current sources of information, see *Monete e Medaglie* (bimonthly, Italy); and *Muenzen und Medaillen-Monnaies et Medailles* (monthly, Switzerland).

Technology of minting is explored in GEORGE F. HILL, *Ancient Methods of Coining* (1977), a work that was originally a pioneering article in a 1924 issue of the *Numismatic Chronicle* of the Royal Numismatic Society; CORNELIUS C. VERMEULE, *Some Notes on Ancient Dies and Coining Methods* (1954), a description of surviving ancient dies and other evidence relating to minting; and D. SELLWOOD, "Minting," in *Roman Crafts,* ed. by DONALD STRONG and DAVID BROWN (1976).

Illustrations of copies of early machinery can be found in LUDWIG VEIT, *Das Liebe Geld: Zwei Jahrtausende Geld- und Münzgeschichte* (1969). F. MAZEROLLE, *Les Medailleurs Français du XV^e siècle au milieu du XVII^e,* 3 vol. (1902–04), discusses documents relating to the introduction of machinery to French mints. A general overview of minting with emphasis on modern technology is presented in DENIS R. COOPER, *Coins and Minting* (1983).

(C.H.V.S.)

# Collective Behaviour

Collective behaviour is a sociological term that refers to the ways in which people behave together in crowds, panics, fads, fashions, crazes, publics, cults, and followings as well as more organized phenomena, such as reform and revolutionary social movements. Because it emphasizes groups, the study of collective behaviour is different from the study of individual behaviour, although inquiries into the motivations and attitudes of the individuals in these groupings are often carried out. Collective behaviour resembles organized group behaviour in that it consists of people acting together; but it is more spontaneous—and consequently more volatile and less predictable—than is behaviour in groups that have well-established rules and traditions specifying their purposes, membership, leadership, and method of operation.

The U.S. sociologist Robert E. Park, who coined the term collective behaviour, defined it as "the behavior of individuals under the influence of an impulse that is common and collective, an impulse, in other words, that is the result of social interaction." He emphasized that participants in crowds, fads, or other forms of collective behaviour share an attitude or behave alike, not because of an established rule or the force of authority, and not because as individuals they have the same attitudes, but because of a distinctive group process.

The absence of formal rules by which to distinguish between members and outsiders, to identify leaders, to establish the aims of the collectivity, to set acceptable limits of behaviour for members, and to specify how collective decisions are to be made accounts for the volatility of collective behaviour. The leader of a mob can become the object of the mob's hatred in a matter of minutes; a fashion leader can suddenly become passé.

Although agreeing that collective behaviour does not generally adhere to everyday rules, some investigators emphasize the emergence of rules and patterns within the collectivity that are related to the surrounding social structure. The U.S. psychologist Ralph H. Turner and the U.S. sociologist Lewis M. Killian define collective behaviour on the basis of "the spontaneous development of norms and organization which contradict or reinterpret the norms and organization of society." Somewhat similar is the U.S. sociologist Neil J. Smelser's definition: "mobilization on the basis of a belief which redefines social action." The distinctive belief—which is a generalized conception of events and of the members' relationships to them—supplies the basis for the development of a distinctive and stable organization within the collectivity. But Smelser's definition points attention, in a way that other definitions do not, toward the unique manner in which members perceive reality; without such a view a group of people would not be engaged in collective behaviour.

The U.S. sociologist Herbert Blumer determined a desire for social change in collective behaviour, as expressed in his definition: "a collective enterprise to establish a new order of life." This definition, however, excludes many of the temporary escapes from conventional life through revelry and orgies, punitive actions such as lynchings, and panics, which are not oriented to any kind of reconstruction of social life or society. Most students of collective behaviour, however, would not restrict the field so severely.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* Part Five, Division II, especially Section 521.

This article is divided into the following sections:

Elementary forms of collective behaviour 556
  Milling 556
  Rumour 557
    Rumour-creating situations
    The transmission of rumour
    Stages of rumour transmission
  Social unrest 557
Major forms of collective behaviour 557
  Responses to disaster 557
    Common misconceptions
    The disaster cycle
  Collective obsessions 558
    Fads
    Hysterical contagion
    Deviant epidemics
    Fashion
    Crazes
  Crowds 559
    Active crowds
    Expressive crowds
  Panic 560
  Publics and masses 561
  Social movements 561
    Characteristics of social movements
    Types of social movements
    The dynamics of social movements
    Relations between structural elements
    The causes of social movements
    The consequences of social movements
Theories of collective behaviour 564
  Individual motivation theories 564
  Interaction theories 565
  Social change 565
The results of collective behaviour 565
  The variety of effects 565
    Short-term effects
    Contingencies
    Long-term effects
  Attempts at control 566
Bibliography 566

## Elementary forms of collective behaviour

Regardless of where or how collective behaviour develops, it requires some kind of preparation. In organized groups there are rituals, such as personal introductions, the toastmaster's humour, and group singing, to facilitate the transition from individual action to group interaction. People may act together efficiently if they have been prepared for a pattern of behaviour such as a fire drill, but the result is organized rather than collective behaviour. Lacking organization, people must first become sensitized to and begin to communicate with one another. These processes of sensitization and communication have been called elementary collective behaviour. Three important elementary forms are milling, rumour, and social unrest.

### MILLING

Prior to most instances of collective behaviour there is a period during which people move about in a somewhat agitated but aimless way. Early students of crowd behaviour, struck by the resemblance to the milling of cattle before a stampede, gave this form of human activity its name. Its characteristic physical restlessness can be seen in an audience waiting for a late-starting program to begin or among citizens who have just received word of an assassination attempt. In the former case people scuffle their feet, leave their seats and walk about, and sometimes join spontaneously in rhythmic behaviour, such as foot stamping. In the latter case people discontinue routine activities and talk with neighbours, friends, and strangers. In most situations milling also means looking for clues to others'

feelings, such as sweating, nervousness, and changes in tone of voice.

Effects of milling

Human milling has at least four important effects. First, it sensitizes people to one another. In this sense milling focuses people's attention on the collectivity and on a subject or problem. Second, milling tends to produce a common mood among the interacting individuals. Where some might react with sorrow, others with anger, and still others with partisan delight or indifference, milling helps to diffuse a single mood within a group. Third, milling develops a common image or interpretation of the situation. The milling throng decides whether the Western tourist taking pictures of a marketplace in the native quarter of an Asian city is harmless or an affront to native dignity; whether the police in an American city are simply arresting a drunken driver or harassing an oppressed minority. Finally, milling sets in motion the process of redefining the rules that govern behaviour. The milling of an audience is usually the signal that customary rules of courtesy toward performers and speakers are no longer applicable and that different forms of behaviour may be expected.

RUMOUR

**Rumour-creating situations.** Rumour abounds under certain circumstances. The U.S. psychologists Gordon W. Allport and Leo Postman offered the generalization that rumour intensity is high when both the interest in an event and its ambiguity are great. The U.S. sociologist Tamotsu Shibutani agreed, contending that rumour abounds when the demand for news is greater than is the supply provided through institutional channels.

At least two conditions must be added to interest and ambiguity as prerequisites for rumour. First, rumour abounds when a group of people share the need to act but are reluctant to do so until the situation can be better defined. Second, rumour abounds only when the situation requires that in some essential respect the members of the group act in concert rather than individually.

Conditions encouraging rumour

There are three major kinds of situations in which these four conditions are commonly met and rumour is rampant. First, in a social order in which information is, or is believed to be, strictly controlled by authorities, rumour is intense. When control over news is a continuing (rather than temporary) condition, rumour becomes regularized as an essential aspect of daily life. The so-called grapevines created by these conditions are regularly utilized by totalitarian regimes, military organizations, and subordinated ethnic groups, races, and social classes.

Second, rumour spreads when events threaten the understandings upon which normal life is based. A major disaster or scandal presents such a challenge. Any change in the regular accommodations between potentially conflicting or competing groups in society similarly calls into question routine patterns of conduct. The suggestion that management may enforce factory rules more strictly, for example, or the suggestion that a college faculty may stiffen or relax degree requirements, immediately provokes a siege of rumour.

Third, rumour springs up when a strong, shared incentive to act is blocked in some way, even by merely the lack of an occasion for action. During states of boredom, rumour capitalizes on minor events, magnifying them into occasions for exciting collective action.

**The transmission of rumour.** Rumour spreads most rapidly along preexisting social networks: among friends, associates, and peers rather than among persons of unequal standing. The messenger who first relates a rumour earns prestige by doing so. Moreover, any specific rumour tends to spread most rapidly when it first enters a group, and to reach persons faster who have responsibilities and interests connected with the event.

The credibility factor

It is frequently assumed—incorrectly—that people transmit rumours only when they believe them and that discrediting a rumour will stop its spread. Other evidence suggests that people pass on rumours whether they believe them or not and that the likelihood of belief increases with their repeated hearing. This latter pattern is understandable if rumour is seen as a seeking, rather than a believing, process, in which every idea, no matter how invalid, provides a way of comprehending a strange or troublesome event. But since the group finds it urgent to reach a common understanding, pressure toward acceptance of a favoured version grows as the rumour process expands. Eventually, there is a sorting out of accounts and an insistence that everyone agree to a consensual account, which then serves as a basis for collective action.

**Stages of rumour transmission.** There is evidence that rumour follows a typical course. Evidence suggests that the rumour process eliminates the most improbable and unreliable accounts and achieves a high degree of veracity when (1) there is considerable recirculation of rumour and (2) there is a fairly well-organized grapevine. When rumour is recirculated the opportunity to compare versions with different groups of people acts as a brake on exaggeration and rubs off the idiosyncratic aspects of the story. With an established grapevine, the source of rumours can often be checked, and individuals who are known to have inside information are regularly consulted for verification.

In both early and late stages, rumour content changes with successive retelling in the direction of the understandable and familiar and in the direction of supporting the actions that the group is starting to take. The former is called assimilation by Allport and Postman and is illustrated by the tendency to make rumour details consistent with prejudice. The latter trend indicates that a group is inclined to support those beliefs that supply justification for some course of action toward which they are already predisposed.

SOCIAL UNREST

The general condition of the community in which milling is both frequent and widespread and in which rumour is recurrent is the crucible in which the more highly organized forms of collective behaviour develop. This condition, known as social unrest, can lead to outbursts of violence. The American urban black uprisings of the 1960s were preceded and accompanied by social unrest in the form of a rise in tensions in black communities throughout the country; the Russian Revolution was preceded by several years of constant unrest and turmoil, involving random assassinations, strikes, and riots.

Characteristics of social unrest

There are several distinguishing characteristics to social unrest. First, there is a general impairment of collective life routines. People find it difficult to concentrate on their work or even to adhere to rules in playing games. Any occasion to abandon routines is welcomed. Second, people are hyperreactive. The magnitude of the response is out of proportion to the usual meaning of any stimulating incident. A small police provocation elicits a major outcry of police brutality; a trivial success is the occasion for large-scale celebration. Milling and rumour abound because incidents that would normally pass with little notice become occasions for both. Third, social unrest is marked by contagiousness. When restlessness is strictly individual, one person's restlessness merely annoys another. But when restlessness becomes a shared experience, people are highly suggestible to one another. Questioning and exploring alternative courses of action are reduced to a minimum. Fourth, social unrest is not specific with respect to grievances or activities. When there is social unrest in a school, students complain of both restrictions on their behaviour and the lack of clearly defined rules; they find fault both with school administrators and with their fellow students. Finally, social unrest is perhaps the most volatile of collective states. Unlike rumour or milling, it does not remain focused on an issue or problem. Unlike crowd behaviour or fads, it has not yet been channeled into one main direction. Although social unrest may eventually die down without any serious aftermath, it is a condition in which people can be easily aroused.

## Major forms of collective behaviour

RESPONSES TO DISASTER

A disaster-stricken community affords a prototypical situation for collective behaviour. The lives of persons are disrupted indiscriminately by a tornado, flood, or earthquake, and coping with the resulting destruction and dis-

order is beyond the capacity of conventional institutions. Of perhaps greatest importance, the assumption of a reasonably stable and predictable reality is undermined.

**Common misconceptions.** A number of common assumptions about behaviour under stress have been dispelled by research on responses to disaster. First, panic is rare. The quite specific conditions under which panic occurs is described below, but stoic, unbelieving, or even resigned reactions are more common than panic. Second, scapegoating is not the rule. Some investigations have suggested an almost unnatural avoidance of singling out villains and placing blame. Within the disaster community the establishment of solidarity is a concern that dampens scapegoating, at least until the immediate emergency is past. Third, there is much less looting and vandalism than is popularly supposed. Even among persons who converge from outside the community there is more petty pilfering for souvenirs than serious crime. Fourth, initially an altruistic selflessness is more prevalent than self-pity and self-serving activity. Frequently noted are dramatic instances of persons who have suffered injury or property damage themselves devoting their time to helping others in no greater need. Fifth, the disruption of established organizations and customary behaviour does not lead primarily to innovation and the exercise of freedom from old restraints. Instead, people more frequently cling to the familiar and seek reinstatement of the old.

**The disaster cycle.** Collective behaviour in disaster follows a characteristic cycle, from first warning to community rehabilitation.

*Warning period.* Although individuals read widely different meanings into disaster warnings, the striking feature of this initial stage is the slowness to believe and the reluctance to act upon warnings. People often remain in their houses in spite of imminent flooding and remain on familiar low ground in the face of tidal wave warnings. The surface calm that each person seeks to maintain in the presence of others can lead to collective self-deception and the inhibition of tendencies toward flight.

Slow reaction time

*Impact and stocktaking period.* In disasters such as floods and some hurricanes there is a distinctly long period of impact, which can be separated from a subsequent period of stocktaking or immobility. In earthquakes and explosions, on the other hand, the impact is so brief that the periods can hardly be separated. The combined period of impact and stocktaking is marked initially by a fragmentation of human relations, as each individual is separated from others and from his customary moorings; it is then marked by a resurgence of interpersonal warmth that transcends customary social barriers within the disaster community.

*Rescue period.* Just as initial fragmentation is followed by unnatural solidarity, stunned immobility gives way to a frenzy of activity in the rescue stage. Although activity is often inefficient, the task of rescuing persons who are trapped and of getting the injured to first-aid facilities is usually accomplished fairly expeditiously, often before outside help arrives. This is the period in which altruism becomes the norm, and old rivalries and conflicts are suspended. Many business concerns adopt an uneconomic generosity, and some individuals disregard their personal welfare. The imperious demand to "do something" at once creates an urgent demand for leadership. People turn first to established community leaders, and, when they are equal to the demands, such figures as police and fire officials, school principals, and mass-media personages are quickly accepted as leaders. Frequently these public figures are as bewildered and distracted as everyone else in the community and are soon abandoned in the restless search for leadership. The leaders then are found among persons who have the specific skills and tools required for the rescue efforts of the moment. Often these are people who do not normally exercise community leadership.

Prevalence of selfless aid

*Rebuilding or "brickbat" period.* The buoyed-up state of the disaster community can last only a short time. Tasks that call for intense effort within a brief time span are completed, and the slow and discouraging work of rebuilding confronts the community. Because the old community cleavages begin to reappear, and because tensions created and repressed during the rescue phase are now released, this period has been called the brickbat stage. The most notable characteristic of this period is the tendency to reinstitute the old community—to rebuild homes on old foundations, to reinstate old forms of organization. In spite of criticism against the ineptitude of established authorities, and in spite of evidence that building locations and methods are vulnerable to the elements, it requires strong leadership to guide the community toward innovation that makes use of what can be learned from the disaster experience.

## COLLECTIVE OBSESSIONS

Characteristics of collective obsessions

The various kinds of collective obsession—fads, hysterias, and the like—have three main features in common. (1) The most conspicuous sign is a remarkable increase in the frequency and intensity with which people engage in a specific kind of behaviour or assert a belief. There was an "epidemic" of flying-saucer sightings; children in every residential neighbourhood in the United States played on skateboards; there was a sudden rush to buy Florida land. (2) The behaviour—or the abandon with which it is indulged—is ridiculous, irrational, or evil in the eyes of persons who are not themselves caught up in the obsession. In the case of recreational fads, such as skateboarding, nonfaddists are amazed at the tendency to drop all other activities in order to concentrate on the fad; the hundreds of incidents in which swastikas were daubed on synagogues during a few weeks in 1959 and 1960 in the United States, West Germany, and other countries shocked the sensibilities of a world that remembered the Nazi persecution of the Jews. (3) After it has reached a peak, the behaviour drops off abruptly and is followed by a counterobsession. To engage in the fad behaviour after the fad is over is to be subjected to ridicule; after the speculative land boom declines, there is a mad rush to sell property at whatever price it will bring. The following discussion covers five types of collective obsession: fads, hysterical contagion, deviant epidemic, fashion, and crazes.

**Fads.** It is tempting to explain fads on the basis of a single motive such as prestige. Prestige is gained by being among the first and most adept at a skill that everyone else covets. That the skill fails as a source of prestige when it is no longer scarce is an important explanation for the abrupt end of a fad. But motives are complex and varied. The exhilaration of joining a band of devotees in an intense preoccupation and the joy of mastering the novel are not to be discounted.

An examination of fads in such enterprises as scientific research and recreation sheds light on the fundamental dynamics of all kinds of fads. First, the scientific fad begins with a new idea or a rediscovered idea—though not just any new idea will set off a fad. The new idea must be a "key invention," one that opens up the possibility for a wide range of minor innovations. Discovery of a potent new drug, for example, is followed by a rush to test the drug in all kinds of situations. Similarly, recreation and style faddists do not merely copy a pattern; they try out a variety of novel uses and variations on the basic pattern. The Hula-Hoop was an ideal fad because each child could develop his own particular variation in spinning the hoop.

Second, the termination of fads is largely explained by the exhaustion of innovative possibilities. The drug has been tested in all of the apparently relevant settings; children have run out of new ways to twirl the Hula-Hoop.

Third, the faddish preoccupation means holding in abeyance many routine activities as well as awareness of drawbacks to the fads. So long as the fad is in full force, a sharp ingroup-outgroup sense insulates faddists against these concerns. But once the faddists run out of new variations they begin to be aware of the extent of their neglect of other activities and to consider possible dangers in the fad.

**Hysterical contagion.** Occasionally waves of fear find expression in a rash of false perceptions and symptoms of physical illness. Girls in an English school fainted in great numbers; women in Mattoon, Ill., reported being anesthetized and assaulted by a mysterious prowler. The best documented case is that of a clothing factory that had

Spread of physical symptoms through fear

to be closed down and fumigated because of reports of toxic insect bites—reports that could not subsequently be substantiated. The U.S. sociologist Alan C. Kerckhoff and the U.S. psychologist Kurt W. Back found that the crisis came after a period during which the women employees had performed unusual amounts of overtime work. The women who became ill from the mysterious insect bites had generally worked more overtime than others and had serious family responsibilities that they could not fulfill because of job demands. Afraid to refuse overtime work lest their job prospects be damaged, yet increasingly upset over neglect of family responsibilities, they found themselves in a conflict from which they could not extricate themselves. Illness from an insect bite provided an excuse to leave work for a day or two. The epidemic continued for about 11 days. It began immediately after a large shipment of foreign cloth had arrived, rendering plausible the assumption that some strange new insect had been introduced to the plant. The first women "bitten" were social isolates, lacking normal social defenses and controls. A rapid spread then took place among women who belonged to intimate cliques, in accord with the theory that social diffusion occurs most readily along well-established lines of social interaction. In the final stage the illness spread to others, irrespective of friendship ties or isolation.

**Deviant epidemics.** Obsessive behaviour also is observed within deviant groups in society. After Edward G. Robinson starred in the motion picture *Little Caesar* (1932), a rash of undersized juvenile delinquents aped his manner. In 1959 and 1960 there was a rash of incidents in which synagogues were desecrated, usually by painting Nazi swastikas on them, and anti-Semitic slogans were painted in public places. In the United States the epidemic began the day after Christmas and continued for nine weeks, encompassing 600 reported incidents. Incidents reached a peak in the third week, with the cycle in small communities lagging a little behind the large cities. In the early and late weeks Jewish synagogues, houses, and other specifically Jewish properties were the main targets. During the middle three weeks anti-Semitic symbols were often placed elsewhere, leading investigators to infer that during the peak of the epidemic many participants were drawn in who were less preoccupied with anti-Semitism than were those who initiated the incidents. Only a minority of the perpetrators were identified and arrested, but these were principally adolescent boys who worked together in small unorganized and heterogeneous groups. Some were strongly anti-Semitic in their attitudes, while others were no more hostile toward Jews than they were toward many other groups or aspects of society.

In this kind of episode socially disapproved feelings are given vent following an initial incident. Beginning with persons who have been holding back a specific feeling for some time, the epidemic builds up until persons with other types of suppressed feelings join in. As the epidemic recedes, these secondary participants drop out first.

**Fashion.** Fashion is much like fads and other collective obsessions, except that it is institutionalized and regularized, becoming continuous rather than sporadic, and partially predictable. Whereas fads often emerge from the lower echelons of society, and thus constitute a potential challenge to the class structure of society, fashion generally flows from the higher levels to the lower levels, providing a continuous verification of class differences. Continuous change is essential if the higher classes are to maintain their distinctiveness after copies of their clothing styles appear at lower levels. Fashions tend to change cyclically within limits set by the stabler culture.

Fashion as a show of class differences

**Crazes.** Another term frequently used to characterize collective obsessions is craze. The term is not analytically separate from "fad" and "fashion," but it does carry somewhat different connotations. Frequently it refers to a collective focus on important figures in the entertainment or sports world—Rudolph Valentino, Frank Sinatra, James Dean, the Beatles, Michael Jackson, and Pelé to name a few. Fans idolize these personalities, relish and mimic real or imaginary details of their lives, and often form clubs or societies to share their fascination. In many instances crazes suffer the same fate as fads—they die abruptly. In some cases, however, figures such as Sinatra and the Beatles outlast the craze and endure as public figures.

Personality idolization

The term craze also has a special connotation in the financial world. There crazes develop when the value of land, stock, or other merchandise is driven well above its intrinsic value by speculation, creating a boom. Such crazes are mainly modern phenomenon, since they require that there be surplus wealth and a flexible and storable medium of exchange. They represent the escalation of a buoyant confidence in the economic future that goes far beyond realistic limits. Financial crazes normally occur after a period of economic expansion and are associated with what seems to be the sudden emergence of a new area of opportunity. The postwar opening up of Spanish New World colonies to British trade was the occasion for the famous 18th-century South Sea Bubble. Combined with craze optimism is the fear of lost opportunity—that is, that the supply of land (or whatever) is not inexhaustible and that only those who buy early will benefit from the initial low prices. The famous crazes have generally received the stamp of authenticity from respected figures who themselves invested and endorsed the enterprise. No less a person than the king of England lost money when the South Sea Bubble burst. Speculative crazes in modern times evolve particularly out of stock exchange activities, although government controls have somewhat curtailed their volatility.

## CROWDS

A thin line separates crowd activities from collective obsessions. The crowd is, first, more concentrated in time and space. Thus a race riot, a lynching, or an orgy is limited to a few days or hours and occurs chiefly within an area ranging from a city square or a stadium to a section of a metropolitan area. Second, a concern of the majority of the crowd (many participants do not always share the concern) is a collaborative goal rather than parallel individual goals. The "june bug obsession" cited earlier, in which dozens of women went home from work because of imaginary insect bites, could have turned into a crowd action if the women had banded together to demand a change in working conditions or to conduct a ceremony to exorcise the evil. Third, because the goal is collaborative, there is more division of labour and cooperative activity in a crowd than in collective obsessions. Finally, a major concern of a crowd is with some improvement or social change expected as a result of its activity. Labour rioters expect management to be more compliant after the riot; participants in a massive religious revival expect life in the community to be somehow better as a result.

The crucial step in developing crowd behaviour is the formation of a common mood directed toward a recognized object of attention. In a typical riot situation a routine police arrest or a fistfight between individuals from opposing groups focuses attention. Milling and rumour then establish a mood of indignation and hostility toward an identified enemy or enemies. In a collective religious experience there is usually an amazing event that rivets attention. Through elementary collective behaviour the mood is defined as religious awe and gratitude toward the supernatural and its agents.

As the mood and object become established, either an "active" crowd or an "expressive" crowd is formed. The active crowd is usually aggressive, such as a violent mob, though occasionally it acts to propel members into heroic accomplishments. The expressive crowd has also been called the dancing crowd because its manifestations are dancing, singing, and other forms of emotional expression.

**Active crowds.** The active crowd identifies an object or group of objects outside itself and proceeds to act directly upon it or them. It will brook no delay or interference, no discussion of the desirability of acting, and no dissent from its course of action. Because of the high pitch of crowd interaction, subtle and indirect courses of action cannot win crowd support, though members are highly suggestible to all proposals and examples for action in keeping with the mood and the object. The stage of transformation from shared mood to shared action constitutes the beginning of the true crowd or mob.

The crucial feature of this stage is overcoming the barriers to such behaviour as the destruction of property or violence toward persons—actions against which most people have strongly ingrained inhibitions. At least four aspects of the way crowd members feel about the situation make this possible. First, there is a sense of an exceptional situation in which a special moral code applies. The crowd merely carries further the justification for a special code of ethics incorporated in the slogan "You have to fight fire with fire!" Second, there is a sense of power in the crowd, with its apparent determination and uniform will, that overcomes the individual's doubts concerning his own ability to carry out a momentous task successfully. Third, there is a sense of impunity, of safety from personal injury and punishment so long as the individual is on the side of the crowd. And finally, there is a sense of inevitability—that the crowd aim will be accomplished regardless of the doubts and opposition of individuals.

Collapse of conventional restraints

Once the crowd breaks through the barrier of conventional restraints there is typically a "Roman holiday" period during which all restraint appears to be dropped. To the outsider, people seem to have gone mad. Rage is entirely uninhibited. But at the same time an atmosphere of intense enjoyment and release is evident. There is laughing and cheering as the violence and destruction become part of a tremendous carnival.

Under cover of the Roman holiday, people pursue many different interests. Looting for personal gain is infrequent in the early stages of rioting. The leading agents in bringing the mob into being are too preoccupied with their indignation for this. But once the general attack is under way, looting for gain, vandalism for fun, and attacks on specific objects to pay off old grudges become prevalent. In Russian and Polish pogroms of the 19th and early 20th centuries, peasants came with their carts to loot Jewish property after they heard that the pogrom was under way. Lynchings in the southern United States in the early part of the 20th century were frequently followed by general forays on black neighbourhoods.

The active crowd normally ends with a tapering-off period, which is sometimes preceded by a stage of siege. In riots of limited scale in which no massive police or military forces are used, the peak day is followed by a few more days of successively smaller numbers of widely scattered encounters. Often the last incidents are in areas not previously hit by rioting. There seems to be some internal mechanism limiting the duration of crowd behaviour, though whether it is fatigue, catharsis, or reassertion of ingrained standards of behaviour is uncertain. In serious riots, however, the police and other armed forces are brought into action long before the riot declines on its own. When police power is applied with only enough force to ensure a standoff between rioters and authorities, there is a period—usually ranging from one to three or four days—of siege. The mood of buoyancy gives way to a mood of dogged persistence. Rioters are more cautious and deliberate in what they do. The desire to have the riot over grows among the participants and in the community, but there is reluctance to give up the fight until concessions have been won.

Development of active crowds

A crowd develops only when a necessary sequence of events occurs and when conditions conducive to crowd development are present. There are at least six such conditions of importance. The first is a deep frustration that is shared by an important segment of the population and that has been festering for a considerable period of time. The frustration is especially poignant when widening intergroup contacts make the frustrated segment more vitally aware of its disadvantages, when its members have been encouraged by education or a public policy statement to aspire to relatively unattainable objectives, and when a period of steadily improving conditions is suddenly interrupted. Second is the presence of deep intergroup cleavages in society. A crowd must have not only a grievance but also an oppressor whom it can blame for its condition. Third is some contradiction in the value system of society, so that there is support both for the social arrangements that the group finds frustrating and for its demands for change. Fourth is a failure of communication, so that grievances can no longer be presented to the appropriate authorities with confidence that they will be given some consideration. Fifth is some failure in the system of control. Mobs often catch police unprepared. In many instances the police, by virtue of their class or ethnic identity, are in sympathy with mobs and unwilling to enforce order. Sixth are experiences leading people to hope that conditions will be improved as a result of violent or disruptive action. Many riots have the support of a well-developed ideology, or they follow occasions when demonstrations and other less extreme tactics have won gains. Among the reasons that mob actions do not soon recur in a given location are that the forces of order are usually strengthened, the hope of great gain is dampened, and channels of communication are often improved after a mob action.

**Expressive crowds.** Not all crowds act. In some crowds the participants are largely preoccupied with themselves or with one another, and with participation in a common experience. Beginning as early as the 7th century in Europe, and continuing throughout the Middle Ages, there were reported epidemics in which groups of people were caught up in a frenzy of dancing that continued until they dropped. Later a collective frenzy of dancing, singing, and shouting became a regular feature of frontier revivals in 19th-century America. Crowds that exceeded conventional limits of revelry have been common in many historical eras. In San Francisco in 1945, license for public violation of sexual mores characterized the day of celebration at the end of the war with Japan.

Expressive crowds may be secular or religious. What distinguishes them is that the production of a shared subjective experience is the crowd's measure of its accomplishment, rather than any action upon objects outside the crowd. One interpretation is that the same determinants of social unrest and frustration may give rise to both the expressive crowd and the active crowd, but the expressive crowd fails to identify an object toward which to act; hence members must release accumulated tension through motions and gestures expressing emotion. According to this view an expressive crowd can fairly quickly metamorphose into an active crowd if an object becomes apparent to them. Another interpretation sees the expressive crowd as equally equipped with an object, but with an object that must be acted upon symbolically rather than directly. Thus, one crowd engages in a wild dance to exorcise evil spirits, whereas another seeks to destroy buildings associated with the "establishment" that it blames for many ills.

The expressive crowd may serve best those types of frustrations requiring revitalization of the individual and group rather than direct modification of external circumstances. Expressive crowds may be especially frequent in periods of frustration and boredom over the predictability and routinization of life, from lack of a sense of meaning and importance in the daily round of life, and from a sense of interpersonal isolation in spite of the physical closeness of others.

## PANIC

The word panic is often applied to a strictly individual, maladaptive reaction of flight, immobility, or disorganization stemming from intense fear. For example, a student "panics" during an examination and is unable to call upon his knowledge in answering questions, or a disaster victim in a situation of mild danger panics and flees into much greater danger. Individual panic frequently occurs as a unique individual response without triggering a similar reaction in others.

Panic as collective behaviour, however, is shared behaviour. When an entire military unit breaks into disorderly flight, a group pattern of orderly behaviour is replaced by a group pattern of panic.

Characteristics of panic

There are a number of distinguishing features to collective panic, four of which are noted here. First, several persons in social contact with one another simultaneously exhibit intense fear and either flee (or demonstrate disorganization leading toward flight) or remain immobile. Second, each individual's fear and his evaluation of the danger are augmented by the signals he receives from others. Third,

flight is indicated as the only conceivable course of action by the signals each is receiving from others. Fourth, the usual rules according to which individuals adjust their behaviour so as not to work at cross-purposes are nullified. In the more dramatic instances of collective panic, people trample one another in vain efforts to reach safety.

Four types of causes for collective panic are generally recognized. First, collective panic usually occurs in the kind of situation that arouses fear in any individual. Hence the psychological causes for individual panic are also the fundamental causes for collective panic.

A second cause of panic is the special character of the situation in which people find themselves. Students of responses to disaster observe that collective panic occurs only when people perceive a danger that is both immediate and severe, when they know of only a very limited number of escape routes from the danger, and when they believe those routes are being closed off so that the time for escape is extremely limited. The requirement that all three conditions be present underlines the observation that intense fear in situations from which there is apparently no escape elicits no collective panic and little individual panic.

Psychologists have suggested that collective panic be viewed as part of a broad class of individualistic crowds. Individualistic crowds include such phenomena as the crush and breakdown of order that sometimes occur at a bargain sale, or the transformation of an orderly ticket-window queue into a shoving and pushing crowd. All the usual mechanisms of crowd behaviour are in operation, but, in contrast to the lynch mob or race riot, the situation encourages the intensified pursuit of individual rather than collective goals.

Behaviour of passengers during the *Titanic* disaster

The situational explanation is not complete by itself, however, as indicated by such occasions as the sinking of the ocean liner *Titanic* with great loss of life but without panic. The ship was visibly sinking, and it was known that there were too few lifeboats for all the passengers, and yet men were frequently reluctant to board the lifeboats until all women and children on board had first been rescued. Hence the third set of causes is the interstimulation of elementary crowd behaviour, the milling, rumour, and social unrest, through which the group forms a collective view of the situation and of the appropriate behaviour. It is difficult to find any logical explanation for the difference in behaviour between the passengers on the *Titanic* and passengers who have panicked in other maritime disasters, except that a norm of gentility and heroism came to dominate the collective definition through these elementary processes.

Since the most dramatic feature of panic behaviour is every individual's disregard for his fellows' lives, many students believe that the fourth set of causes lies in the quality of every individual's relations with his fellows. The U.S. sociologists Kurt Lang and Gladys E. Lang view panic as the end point in a process of demoralization in which behaviour becomes privatized and there is a general retreat from the pursuit of group goals.

### PUBLICS AND MASSES

Crowd behaviour and such related forms as fads and panics are often contrasted with "publics," in which more of an attitude of deliberation prevails. The most important distinction between crowds and publics is that people in the public recognize that there is a division of opinion about an issue and are prepared to interact with a recognition and tolerance of difference. Blumer defines the public as "a group of people who (a) are confronted by an issue, (b) are divided in their ideas as to how to meet the issue, and (c) engage in discussion over the issue." Another important difference is that the product of interaction in the public is public opinion, rather than the collective action or experience of collective ecstasy that eventuates from active and expressive crowds.

Publics are common in societies where public officials and institutional leaders are thought to be responsive to indications of public opinion. When this condition does not prevail, collective behaviour does not usually crystallize beyond the elementary forms, stopping with the establishment of a rumour grapevine. When disillusionment over official response to public opinion reaches a high pitch, publics either do not form or turn quickly into crowds that take direct action.

Comparison of crowds, publics, and masses

The public and crowd should be distinguished from the "mass." Members of a mass exhibit similar behaviour, simultaneously, but with a minimum of interaction. Masses include a wide range of groups. They include, for instance, people simultaneously reading the newspaper advertisement for a department store sale and simultaneously converging on the store with similar objects in mind; but masses also involve people converging in a disaster or a gold rush or a mass migration. In the public and the crowd, social interaction plays a large part in accounting for common definitions of an issue and similar views about how to deal with a problem. But in a mass a great many people react similarly to a common stimulus just because they have common attitudes and motivations. Election behaviour is often closer to the mass than to the public, when taboos on discussing controversial topics lead each person to make up his mind privately on the basis of what he gleans from the mass media of communication.

### SOCIAL MOVEMENTS

Collective behaviour in crowds, panics, and elementary forms (milling, etc.) are of brief duration or episodic and are guided largely by impulse. When short-lived impulses give way to long-term aims, and when sustained association takes the place of situational groupings of people, the result is a social movement.

**Characteristics of social movements.** A movement is not merely a perpetuated crowd, since a crowd does not possess organizational and motivational mechanisms capable of sustaining membership through periods of inaction and waiting. Furthermore, crowd mechanisms cannot be used to achieve communication and coordination of activity over a wide area, such as a nation or continent. A movement is a mixture of organization and spontaneity. There is usually one or more organizations that give identity, leadership, and coordination to the movement, but the boundaries of the movement are never coterminous with the organizations. For example, although organizations such as California's Sierra Club are influential in the movement to preserve the natural environment, anyone who works for the cause and interacts with other workers for this purpose is a member of the conservationist movement. The famous John Brown was not a member of any major abolitionist organization, but his martyrdom made him a leader and symbol for the movement, even though organizational leaders were reluctant to recognize him.

(R.H.T./N.J.S.)

*Social movements and social change.* All definitions of social movement reflect the notion that social movements are intrinsically related to social change. They do not encompass the activities of people as members of stable social groups with established, unquestioned structures, norms, and values. The behaviour of members of social movements does not reflect the assumption that the social order will continue essentially as it is. It reflects, instead, the faith that people collectively can bring about or prevent social change if they will dedicate themselves to the pursuit of a goal. Uncommitted observers may regard these goals as illusions, but to the members they are hopes that are quite capable of realization. Asked about his activities, the member of a social movement would not reply, "I do this because it has always been done" or "It's just the custom." He is aware that his behaviour is influenced by the goal of the movement: to bring about a change in the way things have "always" been done or sometimes to prevent such a change from coming about.

The individual in the social movement

*The membership.* The quixotic efforts of bold, imaginative individuals do not constitute social movements. A social movement is a collectivity or a collective enterprise. The individual member experiences a sense of membership in an alliance of people who share his dissatisfaction with the present state of affairs and his vision of a better order. Like a group, a social movement is a collectivity with a common goal and shared values.

The sense of membership suggests that the individual is

subject to some discipline. In addition to shared values, a social movement possesses norms. These norms prescribe behaviour that will symbolize the member's loyalty to the social movement, strengthen his commitment to it, and set him apart from nonmembers. The norms prohibit behaviour that may cause embarrassment to the movement or provide excuses for attacks by opponents. Commitment is strengthened by participation in group activities with other members and by engaging in actions, individual or collective, that publicly define the individual as a committed member.

A social movement also provides guidelines as to how members should think. Norms of this kind constitute something resembling a "party line"—a definition of the "correct" position for members to take with regard to specific issues. There is subtle pressure on the individual to espouse this position even in the absence of personal knowledge of the arguments for it. Not every member can be expected to study and think through the philosophy that justified the movement and its values. Ideology provides him with a ready-made, presumably authoritative set of arguments.

One of the defining characteristics of a social movement is that it is relatively long lasting; the activity of the membership is sustained over a period of weeks, months, or even years rather than flaring up for a few hours or a few days and then disappearing. A social movement is usually large, but, like duration, largeness is only relative. Some social movements, lasting many decades, may enlist hundreds of thousands of members. Some movements take place within the boundaries of a specific secondary group, such as a religious association or a local community, and may include only a few score or a few hundred members.

The exact size of a social movement is impossible to determine exactly, for membership is not formally defined.

The lack of formal structure

Indeed, one of the salient characteristics of a social movement is the semiformal character of its structure. It lacks the fully developed, formal structure of a stable association, such as a club, a corporation, or a political party. The leaders do not possess authority in the sense of legitimatized power, and members are not formally inducted. The informal, noncontractual quality of membership and the absence of formal decision-making procedures place a premium on faith and loyalty on the part of members. While not all members display these traits, the ideal member gives his total, unselfish loyalty to the movement. Since no legal obligation is assumed on becoming a member, either to conform to the movement's norms or to remain a member, commitment to the movement and its values becomes one of the most important sources of control. The deeply committed member, accepting without question the decisions and orders conveyed by the leaders, sacrificing self, family, and friends if required to do so, is likely to be regarded by outsiders as a fanatic. Some students of social movements, particularly those whose analysis has a psychoanalytic orientation, have suggested that the fanaticism of dedicated members results from individual psychopathological states. An alternative explanation is that the social movement becomes a reference group that provides the member with a new and deviant view of social reality. His basic assumptions about the nature of the social order become so divergent from those of the "normal" members of society that his logic and conclusions are incomprehensible to them.

**Types of social movements.** There is no single, standard typology of social movements. As various scholars focus on different aspects of movements, different schemes of classification emerge. Hence any social movement may be described in terms of several dimensions.

Dimensions of social movements

Many attempts at categorization direct attention to the objective of the movement. The social institution in or through which social change is to be brought about provides one basis for categorizing social movements as political, religious, economic, educational, and the like. It may be argued that all movements tend to be either political or religious in character, depending upon whether their strategy aims at changing political structures or the moral values of individuals.

A commonly used but highly subjective distinction is that between "reform" and "revolutionary" movements. Such a distinction implies that a reform movement advocates a change that will preserve the existing values but will provide improved means of implementing them. The revolutionary movement, on the other hand, is regarded as advocating replacement of existing values. Almost invariably, however, the members of a so-called revolutionary movement insist that it is they who cherish the true values of the society and that it is the opponents who define the movement as revolutionary and subversive of basic, traditional values.

Some attempts to characterize movements involve the direction and the rate of change advocated. Adjectives such as radical, reactionary, moderate, liberal, and conservative are often used for such purposes. In this context the designations "revolutionary" and "reform" are often employed in a somewhat different sense than that described above, with the implication that a revolutionary movement advocates rapid, precipitous change while a reform movement works for slow, evolutionary change.

Killian advances still another typology based on the direction of the change advocated or opposed. A reactionary movement advocates the restoration of a previous state of social affairs, while a progressive movement argues for a new social arrangement. A conservative movement opposes the changes proposed by other movements, or those seeming to develop through cultural drift, and advocates preservation of existing values and norms.

Turner and Killian argue that it is useful at times to categorize social movements on the basis of their public definition, the character of the opposition evoked, and the means of action available to the movement. This scheme is designed to eliminate the subjective evaluation of goals inherent in such categories as reform and revolutionary. A movement that does not appear to threaten the values or interests of any significant segment of society is publicly defined as respectable. If there is no competing movement advocating the same objective, it is also nonfactional. The respectable nonfactional movement must contend primarily with the problems of disinterest and token support, but it has access to legitimate means of promoting its values. A respectable factional movement must contend with competing movements advocating the same general objective but also has access to legitimate means of extending its influence. A movement that appears to threaten the values of powerful and significant interest groups within the society is publicly defined as revolutionary and encounters violent suppression. As a result, it is denied access to legitimate means of promoting its program. Another type of movement is defined as neither respectable nor dangerous but as peculiar; this type, seen as odd but harmless, encounters ridicule and has limited access to legitimate means.

Social movements may also be categorized on the basis of the general character of their strategy and tactics; for instance, whether they are legitimate or underground. The popular distinction between radical and moderate movements reflects this sort of categorization. An obvious difference between types of movements depends upon their reliance on violent or nonviolent tactics. But a nonviolent movement may also be defined as revolutionary or radical because it accepts civil disobedience, rather than legal or parliamentary maneuvering, as a major feature of its strategy. It should be added that the distinction between violent and nonviolent movements is a relative one because a movement may shift rapidly from one to the other as it develops.

**The dynamics of social movements.** As an enduring, sustained collectivity a social movement undergoes significant changes during its existence. This characteristic has led some scholars to formulate a theory of a "life cycle" or "natural history" common to all social movements. Other scholars question the value of the life-cycle approach to social movements, arguing that empirical studies of numerous movements fail to support the notion of invariant stages of development. Smelser suggests as an alternative a value-added theory, which postulates that while a number of determinants are necessary for the occurrence of a social movement, they need not occur in any particular

The pattern of social movements

order. Some may be present for some time without effect only to be activated later by the addition of another determinant. At most it can be said that the idea of the life cycle permits the discovery of conditions that must be present if any movement is to proceed from one stage to another. It may also help identify the conditions that cause a movement to change direction. Still, it can be said that a social movement has a career; for as it endures it always undergoes changes in many of its characteristics, though the sequence of these changes may vary from movement to movement.

*Progressive changes in leadership and membership.* One of the most apparent changes is a shift in leadership. In its earliest stages the strongest influence on a movement is likely to be the charismatic leader who personally symbolizes its values. At some point intellectuals play a leadership role by contributing to the developing ideology of the movement. And if a movement endures and grows for any length of time, administrative leaders arise who are concerned with the practical matters of organization and strategy. Influence in the movement may shift between these types.

Usually the membership of a movement grows during its career, which introduces an element of greater heterogeneity. In the early stages the followers typically are deeply committed with an almost fanatical dedication to the movement's values. If the movement gains a measure of respectability in some segment of society, members may be acquired who are not deeply committed. They are likely to have significant reservations about the movement, and their participation is sporadic. This heterogeneity also can be the basis for internal conflict in a movement. On the other hand, if a movement is publicly defined as revolutionary and subjected to harsh oppression, the membership is likely to be reduced mainly to deeply committed converts or to fanatics who derive some satisfaction from the feeling of being persecuted.

*Progressive changes in goals and strategies.* The goals rarely remain unchanged. As the movement endures and grows, they are likely to become broader and vaguer than they were at the beginning. Proposals for limited, specific reforms become embedded within programs of general social reform. As the leaders and members begin to acquire a sense of power through early victories, the power orientations of the movement may increase. Acquisition of greater power by the population segment that the movement purportedly represents, rather than the implementation of the values of the movement, then becomes a goal. At the same time, the statement of the movement's aim in acquiring power becomes vaguer and more utopian.

Diminishment of original purpose

Changes also occur in the strategy, which may tend in either of two general directions. It may emphasize personal transformation, bringing about social change by converting a majority of society to implement the values by their actions. Or it may emphasize a strategy of societal manipulation, changing social institutions so that the program may be implemented without regard to the number of people favouring the new order. Failure of a movement to gain a large number of converts, combined with indications that it has at its disposal effective means of coercion, leads to a shift to this type of strategy.

Strategy and changes in strategy are strongly influenced by the relationship of the social movement to the larger society and to other social movements. The social structure and the prevailing belief system may suggest either that change can be brought about by changing the hearts and minds of the individual members or that individuals have little effect on the social order. A public definition of the movement as dangerous and subversive may force it to rely increasingly on a strategy of societal manipulation, including violent tactics. The opposition posed by a countermovement may have the same effect, making attempts at persuasion difficult and dangerous and causing a nonviolent, noncoercive movement to use force.

**Relations between structural elements.** As a collectivity, a social movement is characterized by an emergent social structure and a culture. The social structure is reflected in the relationship between leaders and followers, the culture in the values and norms.

Unlike an association, a social movement does not possess legitimate leaders in the sense of being endowed with authority through some formal process. Leaders must constantly substantiate their claims to leadership by demonstrating the effectiveness of their influence on the followers. There is a relationship of reciprocal influence. The followers, for their part, lack institutionalized means of making their influence felt, such as referendums, legislatures, or periodic elections of leaders. It falls to the leaders, therefore, to formulate policies and decisions that will strike a responsive note in their following. Having advanced such proposals, they must rely on either persuasion or coercion to create the illusion that these are collective decisions made by the entire movement. Propaganda thus becomes an important tool of leadership.

Interior structures

Propaganda is also important for maintaining morale and unity. A social movement lacks both the intimacy of a primary group and the formal boundaries of an association. The speeches and writings of leaders serve, in part, to assure the followers of the size, the strength, and the potential for success of the movement—matters difficult for the followers to observe directly. Movements do utilize interpersonal relations to enhance their unity, encouraging small groups of members to meet frequently in circumstances in which they can form personal ties. Mass meetings and parades, with the accompanying ritual, reduce the feelings of isolation that scattered members may experience. Of extraordinary value to a movement is the example of martyrs whose fate arouses indignation in the members, symbolizes unreserved commitment, and lightens the burden of sacrifices.

The culture of a movement encompasses norms and values. Norms are standardized expectations of behaviour developed by members. Values include the program and the ideology. The program is the scheme of change, the new social order that the movement proposes to bring about. The ideology is a body of ideas justifying the program and the strategy of the movement. It usually includes a reinterpretation of history, a projection of the utopia that the success of the movement will introduce, a projection of the disastrous consequences of failure, and a reevaluation of the relationship between population segments and the movement.

**The causes of social movements.** Both individual psychological states and the characteristics of a society at a particular time may be considered as causes of social movements.

*Psychological factors.* Individual factors are psychological states that either convince a person to join a movement or so weaken his commitment to conventional groups that he is willing to risk their disapproval because of his belief in an unpopular cause. Failure to achieve a satisfying status and identity within normal membership groups may be such a factor. The prestige and sense of belonging, which such a person may gain as a member of a social movement, may be even more important to him than the values of the movement. Alienation, feelings of powerlessness, hopelessness, and estrangement from society may predispose an individual to participation. Some scholars argue, however, that there are different kinds of alienation. One type leads merely to apathy and resignation. Political alienation, however, reflects a loss of faith in the political community and predisposes the individual to join a movement that challenges it.

Deprivation, discontent, and frustration are frequently assumed to be sufficient causes for initiating or joining a social movement. The relationship is not a simple one, however. There is little evidence that the most deprived segments of a population are the most likely to participate in social movements. The concept of relative deprivation has been used to explain the fact that persons who could be much worse off than they are but still feel deprived in comparison with even more fortunate groups often play a prominent part in social movements.

*Social factors.* An important task of the student of social movements is to identify those conditions under which social movements are most likely to arise. While the existence of widespread poverty and suffering might seem sufficient to give rise to efforts at reform, it must

be emphasized again that some basis for hope must also exist to stir people to make the effort. Paradoxically, partial alleviation of conditions of deprivation may provide such a basis, serving as the impetus for the formation of a social movement just as things seem to be getting better. The success of other people similarly situated, such as victorious revolutionaries in a neighbouring nation, may be another source of hope.

Social movements and discontent

More general theories of the origin of social movements, such as those of Smelser, Turner, and Killian, suggest that social change may result in strains or conflicts in one or more crucial aspects of the social order. Normative strain arises when changing conditions create a situation in which the established norms no longer lead to the attainment of important, accepted values. Strain in values arises when the values themselves seem to interfere with the satisfaction of important needs of a segment of the society. This sort of strain often arises when different groups, such as immigrants, minorities, or the younger generation, develop values that conflict with those of more established groups. Even with little change in norms and values, changes in social structure reflected in the failure of important functionaries to play their roles adequately may lead to discontent.

The general nature of the belief system existing in the culture of a society affects the likelihood that social movements will arise and defines the type that will occur. For example, a system that is essentially fatalistic is less conducive to social movements, particularly those with a strategy of societal manipulation, than one that emphasizes the perfectibility of man and his control over his own fate.

**The consequences of social movements.** It has been suggested that the committed participant in a social movement undergoes a psychological reorganization. It is clear that his new sense of security and importance is acquired at the sacrifice of autonomy. As a loyal member he tends to let the leaders do his thinking for him, suppressing doubts as to the validity of the ideology and the wisdom of the leader's decision. He repeats their arguments in a dogmatic fashion; persons who are not in the movement find it difficult to debate with him since they start from different premises. His perception is selective in a different way from theirs. The ideology, for example, may lead him to view all governmental authorities as villains, while the ordinary citizen views them as legitimate leaders, some good, some bad. The end product of this surrender of autonomy may be an altered worldview. Some things taken for granted before becoming part of the movement will never seem the same again, even after leaving the discipline of the movement.

The end products of social movements as collectivities attempting to change the social order cannot be analyzed simply in terms of success or failure. Failure may come as a result of ruthless suppression of the movement or through widespread apathy. A movement may wither away because too few take it seriously and it does not develop enough power to force its program on society. Sometimes the remnants may linger for a long time as a cult, oriented inward toward the gratifications that the members obtain from participation but making no serious effort to change the social order.

Consequences of success

Success is most apparent when a movement manages to have its power legitimized as authority. In a successful revolution the social movement becomes the new source of authority and respectability, and opposition to its values is defined as counterrevolutionary. In other instances, the movement achieves power through secession. Failing to compel acceptance of its values in the larger group or society, the members withdraw into a new social system in which they can attempt to implement the values separately from a hostile or indifferent society.

A less obvious form of success is the institutionalization of the values or some part of them. Accepting the legitimacy of the movement's values, the traditional associations in the society incorporate them into their own values and implement them without a transfer of authority to the movement. Thus the U.S. Socialist Party has seen many of its proposals adopted by the two major political parties and the government without winning a major election or overthrowing the government. Sometimes the social movement itself is institutionalized by being accorded authority as the legitimate custodian of the new values. The movement is then transformed into a bureaucratic association, as happened with the American labour movement of the early 20th century and the Congress Party of India after British rule ended. (L.M.K./Ed.)

## Theories of collective behaviour

Because much collective behaviour is dramatic, unpredictable, and frightening, the early theories and many contemporary popular views are more evaluative than analytic. The French social psychologist Gustave Le Bon identified the crowd and revolutionary movements with the excesses of the French Revolution; the U.S. psychologist Boris Sidis was impressed with the resemblance of crowd behaviour to mental disorder. Many of these early theories depicted collective behaviour as an atavism, in which the evolutionary accomplishments of civilization were stripped away and human behaviour returned to an earlier stage of development. Freud retained this emphasis in viewing crowd behaviour and many other forms of collective behaviour as regressions to an earlier stage of childhood development; he explained, for example, the slavish identification that followers have for leaders on the basis of such regression.

Social and revolutionary theories

More sophisticated recent efforts to treat collective behaviour as a pathological manifestation employ social disorganization as an explanatory approach. From this point of view collective behaviour erupts as an unpleasant symptom of frustration and malaise stemming from cultural conflict, organizational failure, and other social malfunctions. The distinctive feature of this approach is a reluctance to take seriously the manifest content of collective behaviour. Neither the search for enjoyment in a recreational fad, the search for spiritual meaning in a religious sect, nor the demand for equal opportunity in an interest-group movement is accepted at face value.

An opposite evaluation of many forms of collective behaviour has become part of the analytic perspective in revolutionary approaches to society. From the revolutionist's point of view, much collective behaviour is a release of creative impulses from the repressive effects of established social orders. Revolutionary theorists such as Frantz Fanon depict traditional social arrangements as destructive of human spontaneity, and various forms of crowd and revolutionary movements as man's creative self-assertion bursting its social shackles.

### INDIVIDUAL MOTIVATION THEORIES

Among the analytic theories that seek to eschew evaluation, the most popular ones stress individual motivation in accounting for collective behaviour. Frustration and lack of firm social anchorage are the two most widely used explanations for individual participation in collective behaviour of all kinds. In the psychiatric tradition, frustration heightens suggestibility, generates fantasy, brings about regressions and fixations, and intensifies drives toward wish fulfillment so that normal inhibitions are overcome. Since most forms of collective behaviour promote thoughts that are otherwise difficult to account for and that breech behavioral inhibitions, this is often a fruitful source of explanation.

Psychiatric and sociological traditions

In the sociological tradition of Émile Durkheim, absence of firm integration into social groups leaves the individual open to deviant ideas and susceptible to the vital sense of solidarity that comes from participation in spontaneous groupings. Drawing upon both the psychiatric and the sociological traditions, Erich Fromm attributed the appeal of mass movements and crowds to the gratifying escape they offer from the sense of personal isolation and powerlessness that people experience in the vast bureaucracies of modern life. Extending Karl Marx's theory of modern man's alienation from his work, many contemporary students attribute faddism, crowds, movements of the spirit, and interest-group and revolutionary movements to a wide-ranging alienation from family, community, and country, as well as from work.

According to the approach suggested by the U.S. political scientist Hadley Cantril, participation in vital collectivities supplies a sense of meaning through group affirmation and action and raises the member's estimate of his social status, both of which are important needs often frustrated in modern society. Eric Hoffer, a U.S. philosopher, attributed a leading role in collective behaviour to "true believers," who overcome their own personal doubts and conflicts by the creation of intolerant and unanimous groups about them.

#### INTERACTION THEORIES

Sociologists and social psychologists, without denying the place of individual motivation in any complete explanation for collective behaviour, have more often stressed a distinctive quality or intensity of social interaction. The U.S. sociologist Ernest Burgess, along with Park, associates collective behaviour with "circular reaction," a type of interaction in which each person reacts by repeating the action or mirroring the sentiment of another person, thereby intensifying the action or sentiment in the originator. Blumer adds a subtlety to this theory by sharply distinguishing circular reaction from "interpretative interaction," in which the individual first interprets another's action and then makes a response usually different from the stimulus action. Another stream of thought has stressed difference of intensity rather than kind of interaction. Following the lead of the French social scientist Gabriel Tarde and the French psychologist Alfred Binet, many investigators have looked for clues that normal imitative tendencies and suggestibility may be intensified in collective behaviour. An important approach is based on the U.S. psychologist Floyd H. Allport's criticism of Le Bon and William McDougall, a British-born U.S. psychologist, for their concept of "group mind," and for their apparent assumption that collective behaviour makes people do things to which they are not predisposed. Allport insisted instead that collective behaviour involves merely a group of people doing what they previously wanted to do but for which they lacked the occasion and the support of like-minded associates.

Contagion, convergence, and emergent norm

These interaction theories have been labeled contagion and convergence theories, respectively—the former stressing the contagious spread of mood and behaviour; the latter stressing the convergence of a large number of people with similar predispositions. Both have sought to explain why a group of people feel and act (1) unanimously, (2) intensely, and (3) differently from the manner in which they customarily act. Other interaction theorists have challenged the assumption of unanimity, proposing that in most kinds of collective behaviour a single mood and course of action is established with such force and intolerance that the many who privately dissent are silenced, creating an illusion of unanimity. Rather than contagion, it is an emergent norm or rule that governs external appearances and, to a lesser extent, internal convictions in collective behaviour.

Freud, too, stressed a distinctive pattern of interaction in collective behaviour. The key to these groupings is the desire to possess a beloved leader. Because the leader is unattainable, and because his attentions must be shared among many followers, a relation of identification is expressed in the demand for uniformity that the followers insistently impose on each other, according to the example of the leader.

#### SOCIAL CHANGE

A final set of theories stresses characteristics of social organization that generate collective behaviour. Collective behaviour is commonly seen by sociologists as a normal accompaniment and medium for social change, relatively absent in periods of social stability. With the more or less continuous shifts of values in any society, emerging values are first given group expression in collective behaviour; efforts to revitalize declining values also bring forth collective behaviour. Again, the constant readjustments in the power of different population segments are implemented and resisted through collective behaviour. Because it is a means of communication, and because it is always characterized by novel or intensified control over individuals, collective behaviour also arises to bypass blockages in communication and to install an emergent order when formal or informal regulation of behaviour is inadequate.

The most comprehensive theory specifying necessary conditions for the development of most major forms of collective behaviour was advanced by Smelser. He noted six conditions that must be present: (1) the social structure must be peculiarly conducive to the collective behaviour in question; (2) a group of people must experience strain; (3) a distinctive type of belief must be present to interpret the situation; (4) there must be a precipitating event; (5) the group of people must be mobilized for action on the basis of the belief; and (6) there must be an appropriate interaction between the mobilized group and agencies of social control. The detail for each condition varies with the type of collective behaviour.

Since the early 1970s two new strands of theory and empirical research have arisen, one in the United States and one in western Europe. The first, called resource mobilization theory, takes as its starting point a critique of those theories that explain social movements as arising from conditions of social disorganization and strain and as finding their recruits among the isolated and alienated in society. By contrast, research mobilization theorists argue that the success of social movements rests mainly on the resources that are available to it; this means forming coalitions with already-existing organizations, securing financial support, and mounting effective and organized campaigns of political pressure. As a result of this emphasis, resource mobilization theorists downplay the factor of ideology—and irrational factors generally—in the study of social movements.

The second theory is the new social movement theory. It derives from an intellectual dissatisfaction with the predominantly Marxist view that treats social movements as reflecting a fundamental struggle among classes organized around economic production. That theory, it is argued, has become less relevant as these classes have been drawn into collective bargaining, the welfare system, and other social advancements within the state. The "new social movements" that have arisen in their place are interpreted as struggles against the social inequalities, the dominance of the mass media, and other features of postindustrial capitalism and the welfare state. These include youth, feminist, peace, and ecological movements, as well as the rise of group conflicts based on ethnicity and race. Jürgen Habermas, a German sociologist, interpreted such movements as protests against the excessive size and rationality of the state and its bureaucracies and their intrusion into the private worlds of individuals.

## The results of collective behaviour

#### THE VARIETY OF EFFECTS

**Short-term effects.** The most notable immediate effect of all kinds of collective behaviour is to alter the salience of various problems, issues, and groups in public awareness. Popular concern about disarmament grew large as "Ban-the-Bomb" demonstrations proliferated during the late 1950s and early 1960s; then public interest waned as demonstrations became infrequent or ceased. A fad calls attention to recreational needs; the circumstances surrounding a panic monopolize public attention. Second, all forms of collective behaviour contribute to polarizations, forcing people to take sides on issues and eliminating the middle ground. Often a three-sided conflict develops among the two polarized groups and mediators who wish to de-emphasize divisive issues altogether. Third, every instance of collective behaviour either alters or strengthens the makeup of group and community leadership. The swings of fashion discredit some clothes designers and boost others to prominence. A riot or a wildcat strike usually reveals the inability of established leaders to control their members and produces emergent leaders from among the spokesmen acceptable to members.

**Contingencies.** How the immediate effects of collective behaviour are translated into long-term consequences depends upon several contingencies, of which four merit

attention. First, the nature of the response by authorities affects the immediate course of the collective behaviour. Some evidence suggests that alarmed and repressive reactions strengthen polarization, that moderate reactions strengthen the mediation viewpoint, and that inaction or ineffectual action facilitates efforts toward usurpation of authority.

Second, the response of authorities affects public definitions of the meaning of the collective behaviour. Publics have variously defined particular fads as harmless diversions, threats to authority and order, threats to health and well-being, visitations of the Holy Spirit, and possession by the devil, treating them quite differently in consequence. Lynchings are vigilante actions, or they are criminal subversions of justice. Riots can be viewed as mass criminality or as social protest. Social movements are defined as respectable, or as peculiar but harmless, or as dangerous and revolutionary, evoking polite support, embarrassed avoidance, or active repression, respectively.

A third contingency affecting the aftermath of collective behaviour concerns the nature and strategy of the countermovements or counterfads that arise. When the countermovement arises, acquires a bitter and reactionary tone, and becomes a backlash, polarization and heightened disorder often lead to demands for order at any cost, at the expense of any amelioration that might otherwise have occurred. But backlash is often self-discrediting as "extremism," and over the long run it sometimes pushes many people onto the side of amelioration. Countermovements that avoid the backlash pattern typically try to undermine the group they oppose by taking some of the latter's aims as their own, thereby helping to effect reforms sought in the initial protest.

Conventionalization of collective behaviour

Finally, the effect of collective behaviour depends upon the ubiquitous process of conventionalization. In a spontaneous fad or mob action, participants usually copy the pattern of earlier incidents with which they are familiar, so that separate incidents in a wave of collective behaviour exhibit a similarity indicating the development of customary ways of rioting, or playing at a fad, and possibly even of panicking. When incidents are repeated, a gradual accommodation between participants in collective behaviour and the authorities becomes routinized. Once the behaviour is conventionalized in this fashion, there are increasing efforts to create and use the conventionalized form of collective behaviour for private and public aims. Much advertising seeks to create fads in conventionalized ways. Political rallies, sports rallies, and some of the ceremonies of established religious organizations seek to conventionalize the enthusiasm and sense of solidarity of expressive crowds. Social movements rapidly acquire stable organizations, sects become denominations, political movements become political parties or are absorbed into parties, and humanitarian movements become stabilized as associations to promote some form of human betterment. Conventionalization extends the influence of orienting ideas, but it also ensures compromise and abandonment of the most disruptive and controversial features of the initial behaviour.

**Long-term effects.** In the long run it is difficult to be sure whether a particular type of collective behaviour actually makes a difference or whether it is merely a shadow cast by passing events. Scattered collective behaviour is endemic in every society. But when there is widespread discontent, collective behaviour soon becomes a prominent feature of group life. When there are no exciting new ideas—such as the liberal humanitarian vision of the 18th and 19th centuries, the Socialist idea of the 19th and 20th centuries, and the nationalist mystique of the 20th century—collective behaviour consists principally of expressive behaviour, panics, and unfocused disruption or intergroup vengeance such as pogroms. This kind of collective behaviour probably contributes little to change. But when there is a new perspective to give meaning to discontent, many forms of collective behaviour appear to become agents of change. Even a recreational fad becomes a form of self-assertion for a rising class or age group. Le Bon suggested that in a period of widespread discontent crowd action serves to destroy an old order in preparation for a new one. Social movements help to build the new order.

One view holds that collective behaviour supplies a testing ground on which new ideas are tried out for general acceptability and on which groups test their strength against forces of resistance. The outcome of this testing is sometimes change and sometimes public demonstration that the old order is still viable. This view suggests that collective behaviour has as great a function to play in maintaining social stability as in implementing social change.

## ATTEMPTS AT CONTROL

Attempts to control collective behaviour vary according to whether change or stability is sought. Advocates of change seek to control countermovements and backlash crowds, as well as those expressive crowds and fads that anesthetize people to their grievances, whereas advocates of stability seek to control crowds and movements that undermine public order or threaten revolution. Advocates of both change and stability likewise make use of collective behaviour in achieving their aims. The volatile and unpredictable nature of all collective behaviour renders manipulation and control highly problematic, however, and masters of control, such as the French revolutionary Robespierre, have often been victims of the followers they once manipulated.

The most sensitive and difficult control problem occurs at the moment of the first precipitating incident and during the stage of transformation in an active crowd. A show of weakness—or maybe even unnecessary repression—will escalate the crowd into the Roman-holiday stage. It is essential to identify spokesmen who command a hearing with the crowd—often not the established group leaders—and open serious negotiations with them. Poorly arranged negotiating sessions before television cameras are easily turned into occasions for incitement of the crowd. If the provocations of excessive policing are avoided and one or two dramatic concessions of great symbolic importance made, a cooling-off period may be secured in which more comprehensive measures to relieve tensions in the situation can be undertaken.

Once collective behaviour is fully escalated there is seldom any control technique available except massive suppression, and some experts believe that crowd behaviour will spring up again if crushed before it has substantially run its course. Interference with an expressive crowd, and even with many fads and instances of hysterical contagion, often turns it into a hostile, active one. As the intensity of feeling begins to decline, the time is then ripe to quicken the end of crowd behaviour by intensifying negotiations with spokesmen respected by the crowd. (R.H.T./N.J.S.)

## BIBLIOGRAPHY

*Theoretical and general studies:* Theories of collective behaviour are introduced in HADLEY CANTRIL, *The Psychology of Social Movements* (1941, reprinted 1973); SIGMUND FREUD, *Group Psychology and the Analysis of the Ego* (1922, reissued 1975; originally published in German, 1921); WILLIAM A. GAMSON, *Power and Discontent* (1968); ERIC HOFFER, *The True Believer: Thoughts on the Nature of Mass Movements* (1951, reissued 1980); RICHARD T. LAPIERE, *Collective Behavior* (1938); and DAVID L. MILLER, *Introduction to Collective Behavior* (1985). See also JOHN C. BRIGHAM, *Social Psychology* (1986). The major general treatments of the subject include HERBERT BLUMER, "Collective Behavior," in ALFRED M. LEE (ed.), *Principles of Sociology,* 3rd ed. (1969), a classic sociological statement of a widely used approach; ROGER BROWN, "Mass Phenomena," in GARDNER LINDZEY (ed.), *Handbook of Social Psychology,* vol. 2, pp. 833–876 (1954); and STANLEY MILGRAM and HANS TOCH, "Collective Behavior: Crowds and Social Movements," in GARDNER LINDZEY and ELLIOT ARONSON (eds.), *Handbook of Social Psychology,* 2nd ed., vol. 4, pp. 507–610 (1968), comprehensive reviews presented by psychologists; ROBERT R. EVANS (ed.), *Readings in Collective Behavior,* 2nd ed. (1975), a collection of classic journal articles; KURT LANG and GLADYS ENGEL LANG, *Collective Dynamics* (1961), a standard textbook; NEIL J. SMELSER, *Theory of Collective Behavior* (1963, reissued 1971), a classic theoretical treatise and text; RALPH H. TURNER, "Collective Behavior," in ROBERT E.L. FARIS (ed.), *Handbook of Modern Sociology,* pp. 382–425 (1964), an analytic statement of the field for the advanced student in sociology; RALPH H. TURNER and LEWIS M. KILLIAN, *Collective Behavior,* 3rd ed.

(1987), a standard textbook; and JOHN LOFLAND, *Protest: Studies of Collective Behavior and Social Movements* (1985).

*Specialized studies:* Elementary collective behaviour is studied in GORDON W. ALLPORT and LEO POSTMAN, *The Psychology of Rumor* (1947, reprinted 1975); TAMOTSU SHIBUTANI, *Improvised News: A Sociological Study of Rumor* (1966); and FREDRICK KOENIG, *Rumor in the Marketplace: The Social Psychology of Commercial Hearsay* (1985). Responses to disaster are the subject of GEORGE W. BAKER and DWIGHT W. CHAPMAN (eds.), *Man and Society in Disaster* (1962); WALTER LORD, *A Night to Remember* (1955, reissued 1984); HARRY E. MOORE, *Tornadoes over Texas: A Study of Waco and San Angelo in Disaster* (1958); and UNITED STATES. FEDERAL EMERGENCY MANAGEMENT AGENCY, *Behavior and Attitudes Under Crisis Conditions: Selected Issues and Findings* (1984). For discussion of collective obsessions, see DAVID CAPLOVITZ and CANDACE ROGERS, *Swastika 1960: The Epidemic of Anti-Semitic Vandalism in America* (1961); JOHN CARSWELL, *The South Sea Bubble* (1960); ALAN C. KERCKHOFF and KURT W. BACK, *The June Bug: A Study of Hysterical Contagion* (1968); and CHARLES MACKAY, *Memoirs of Extraordinary Popular Delusions,* 3 vol. (1841, reissued in 1 vol. as *Extraordinary Popular Delusions and the Madness of Crowds,* 1981).

*Crowds:* E. LOUIS BACKMAN, *Religious Dances in the Christian Church and in Popular Medicine* (1952, reprinted 1977; originally published in Swedish, 1945); HUGH DAVIS GRAHAM and TED ROBERT GURR (eds.), *Violence in America: Historical and Comparative Perspectives* (1969; published also as *The History of Violence in America*); GUSTAVE LE BON, *The Crowd: A Study of the Popular Mind,* 2nd ed. (1968, reprinted 1984; originally published in French, 1895); *Report of the National Advisory Commission on Civil Disorders* (1968); ARTHUR F. RAPER, *The Tragedy of Lynching* (1933, reprinted 1970); GEORGE RUDÉ, *The Crowd in History: A Study of Popular Disturbances in France and England, 1730–1848,* rev. ed. (1981); CARL F. GRAUMANN and SERGE MOSCOVICI (eds.), *Changing Conceptions of Crowd Mind and Behavior* (1986); and FRANK STAGG, E. GLENN HINSON, and WAYNE E. OATES, *Glossolalia: Tongue Speaking in Biblical, Historical and Psychological Perspective* (1967).

(R.H.T.)

*Social movements:* RUDOLF HEBERLE, *Social Movements: An Introduction to Political Sociology* (1951), develops general theories but focuses on the relationship between social movements and political parties. CLARENCE W. KING, *Social Movements in the United States* (1956), develops general principles from the analysis of selected social movements. HANS TOCH, *The Social Psychology of Social Movements* (1965); and MUZAFER SHERIF and CAROLYN W. SHERIF, *Social Psychology* (1969), represent theoretical approaches, placing greater emphasis on individual motivational and perceptual processes. LYFORD P. EDWARDS, *The Natural History of Revolution* (1927, reprinted 1970); CRANE BRINTON, *The Anatomy of Revolution,* rev. ed. (1965); and GEORGE S. PETTEE, *The Process of Revolution* (1938, reissued 1971), represent attempts to develop general theories of revolution as a type of social movement through the analyses of American and European revolutions. A similar effort based on studies of revolutions in Latin America and the Middle East is found in CARL LEIDEN and KARL M. SCHMITT, *The Politics of Violence* (1968, reprinted 1980). There are numerous studies of particular social movements. Representative of those that include theoretical propositions, as well as historical descriptions, are E.J. HOBSBAWM, *Primitive Rebels: Studies in Archaic Forms of Social Movement in the 19th and 20th Centuries,* 3rd ed. (1971); CORA DU BOIS, *The 1870 Ghost Dance* (1939, reprinted 1976); KENELM BURRIDGE, *Mambu: A Melanesian Millennium* (1960, reissued as *Mambu: A Study of Melanesian Cargo Movements and Their Ideological Background,* 1970), studies of nativistic movements; and FREDERICK KRANTZ (ed.), *History from Below: Studies in Popular Protest and Popular Ideology in Honour of George Rudé* (1985). Analyses of Communism and Nazism are found in PHILIP SELZNICK, *The Organizational Weapon: A Study of Bolshevik Strategy and Tactics* (1952, reissued 1979); and THEODORE ABEL, *Why Hitler Came into Power* (1938, reprinted 1986). Essays on the Algerian revolution are found in FRANTZ FANON, *Studies in a Dying Colonialism* (1965; originally published in French, 1959). Analyses of social movements in the United States are found in THOMAS H. GREER, *American Social Reform Movements: Their Pattern Since 1865* (1949, reprinted 1980); LEWIS M. KILLIAN, *The Impossible Revolution? Black Power and the American Dream,* 2nd ed. (1975); JEROME H. SKOLNICK, *The Politics of Protest* (1969); and ALEC BARBROOK and CHRISTINE BOLT, *Power and Protest in American Life* (1980). A source on the new social movements is JÜRGEN HABERMAS, *Die Neue Unübersichtlichkeit: Kleine politische Schriften V* (1985).

(L.M.K./N.J.S.)

# Cologne

One of the key inland ports of Europe, Cologne (German: Köln) is the fourth largest city in Germany and the largest city in the *Land* ("state") of Nordrhein-Westfalen (North Rhine-Westphalia). It is the historic, cultural, and economic capital of the Rhineland. Cologne's commercial importance grew out of its position at the point where the huge traffic artery of the Rhine (German: Rhein) River intersected one of the major land routes for trade between western and eastern Europe. In the Middle Ages it also became an ecclesiastical centre of significance and an important centre of art and learning. This rich and varied heritage is still much in evidence in present-day Cologne, despite the almost complete destruction of the Inner City during World War II. Cologne is the seat of a university and the see of a Roman Catholic archbishop. Its cathedral, the largest Gothic church in northern Europe, is the city's major landmark and unofficial symbol.

This article is divided into the following sections:

Physical and human geography 568
  The landscape 568
    The city site
    The city layout
    Architecture
  The people 569
  The economy 569
    Finance and industry
    Transportation
  Administration and social conditions 569
    Government
    Education
  Cultural life 569
History 569
  Early settlement and medieval growth 569
  The 19th and 20th centuries 570
Bibliography 570

## Physical and human geography

### THE LANDSCAPE

**The city site.** Cologne is situated about 21 miles (34 kilometres) northwest of Bonn and 25 miles southeast of Düsseldorf. It lies 210 feet (65 metres) above sea level, just below where the Rhine enters the fertile North German Plain. The river at this point is navigable to seagoing vessels. The immediate surroundings of Cologne are varied. The picturesque hills of the Bergisches Land lie to the east, while on the west is another group of hills forming a chain called the Ville. The North German Plain stretches away to the north and northwest, and the Rhine Valley winds to the southeast toward Bonn.

The area of the modern city is about 156 square miles (405 square kilometres), the greatest distance from west to east being about 17 miles and from north to south about the same. There are 85 districts, divided into nine city areas (*Bezirke*). Most of the city lies on the left (west) bank of the river, but it also incorporates a cluster of suburbs on the right bank, some of which were annexed in 1975. Average temperatures in the Cologne area are 36° F (2° C) in January and 64° F (18° C) in July.

**The city layout.** The semicircular shape of the Inner City was originally determined by a defensive wall, four miles long, that was completed in about 1200. The wall enclosed several formerly separate parishes and afforded protection for some 35,000 to 40,000 people. (At that time Cologne was bigger than Paris.) The flat side of the semicircle was formed by the Rhine. In the 1880s the medieval fortifications were demolished and replaced by a chain of ring roads, called the Ringstrassen.

Although Cologne has spread far beyond the confines of the Ringstrassen, its focal point is still within this area, the Inner City (Innenstadt). There are found the main shopping and business streets—such as the Hohe Strasse (north–south) and Schildergasse (west–east), both of which have been closed to motor vehicles—as well as the city's historic buildings. Several bridges span the river at Cologne; five of them were rebuilt after World War II, and the rest were postwar additions.

A large proportion of Cologne's area consists of parkland, woods, lakes, sports facilities, and open areas. Two major park systems follow roughly the concentric patterns of old fortifications around the Innenstadt. The first is just outside the Ringstrassen and includes (from north to south) zoological and botanical gardens, the Stadtgarten, and the Volksgarten. The second, the Outer Greenbelt, is a wooded area that stretches for miles around the western and southern edge of the city and contains extensive recreation grounds and the Müngersdorfer Stadium. On the right bank of the river is the Rheinpark, with halls for fairs and exhibitions.

Cologne cathedral

**Architecture.** Cologne cathedral (Kölner Dom) eclipses in its size and grandeur all other historic buildings in the city. Its twin towers rise 515 feet above the city centre. After an earlier cathedral on the site was destroyed by fire in 1248, it was decided that a new one would be built in the Gothic style, emulating the cathedrals of France. The choir was completed in 1320 and consecrated in 1322. Construction continued until 1560, when it came to a halt. The cathedral stood unfinished until 1842, when work was resumed. In 1880 the enterprise was finally completed. The building was badly damaged by air raids in 1944; but by 1948 the choir had been restored and was again in regular use, as was the rest of the interior by 1956.

The 14th-century stained-glass windows in the choir are considered especially beautiful, and the cathedral is also noted for its other art treasures. On the high altar is a massive gold shrine containing what are said to be relics of the Magi, sent to Cologne from Milan in 1164. This shrine, begun by Nicholas of Verdun in 1182 and completed in about 1220, is considered one of the finest examples of medieval goldwork. The altar in the Lady Chapel (on the south wall of the choir) has a triptych, "The Adoration of the Magi," painted between about 1440 and 1445 by Stefan Lochner, the outstanding painter of the Cologne school.

By the south side of the cathedral lies a reminder of Cologne's still more ancient past: the mosaic floor of a banquet hall in a great Roman villa, discovered during excavations near the cathedral in 1941. The floor is now incorporated in the Römisch-Germanisches Museum. Other Roman remains in Cologne include a well-preserved 1st-century-AD tower from the earliest city wall, the remains of the North Gate, a large portion of the Praetorium (governor's residence) visible in the basement of the restored Gothic Rathaus ("Town Hall"), and a mausoleum in Weiden on the outskirts. The Ubier-Monument, discovered in the 1960s, dates from the period of the Ubii occupation of the area (see below *History*). Remains of the medieval walls can still be seen, and three of the original 12 gates survive: the Eigelsteintor, Hahnentor, and Severinstor. The Bayenturm, a medieval tower, stands near the Rhine.

Medieval churches

Apart from the cathedral, the Inner City possesses many other noble churches, largely built in the prosperous Middle Ages. Particularly in evidence is the Romanesque style, of which the best examples are Sankt Gereon, Sankt Severin, Sankt Ursula, Sankt Maria im Kapitol, Sankt Kunibert, Sankt Pantaleon, Sankt Aposteln, and Gross Sankt Martin. After sustaining severe wartime damage, these churches underwent a program of restoration, the completion of which was celebrated in 1985. The 14th-



The Cologne cathedral (Kölner Dom).
© Nat Norman/World Wide Photography—Rapho/Photo Researchers

century Antoniterkirche, a secularized monastery church, was made over to the Protestants in 1802 and became the first public Lutheran church in Cologne.

Among Cologne's secular medieval buildings that suffered in World War II and have undergone reconstruction are the Overstolzenhaus, a 13th-century Romanesque house, and the Rathaus, with its 16th-century porch. The Gürzenich, or Festhaus ("Banquet Hall"), of the merchants of the city (1441–47), reconstructed as a concert and festival hall, and the 16th-century Zeughaus ("Arsenal"), which contains a historical museum, were both restored to their medieval form only on the outside.

These ancient buildings share the crowded city centre with modern offices, shops, a theatre and opera house (opened in 1957), and, just north of the cathedral, the great railway station. Near the perimeter of the city is the new town hall. About a mile from the cathedral is the 798-foot Fernmeldeturm ("Telecommunications Tower"; 1981).

### THE PEOPLE

Cologne is the fourth in population of Germany's cities, after Berlin, Hamburg, and Munich. About 85 percent of its population is of German nationality; of the remainder, most are southern European guest workers who have moved to the city since the 1970s, chiefly from Turkey and Italy but also from Yugoslavia and Greece. The predominant religion of the German community is Roman Catholicism, but there is also a large Protestant minority. There is also a sizable Muslim community and a small Jewish one.

### THE ECONOMY

**Finance and industry.** The city remains a banking centre, as it was in the Middle Ages, and the wine trade and textile manufacturing are still prominent, but business activity has become greatly diversified. Insurance has assumed a major position. The city has been a centre of the automotive industry—notably engine manufacture—since the late 19th century, and engineering, electrical engineering, metals, chemicals, and pharmaceuticals have also come to the fore. Other manufactures include chocolate and the city's famous eau de cologne, which was first produced commercially at the beginning of the 18th century. In addition, several important economic organizations, such as the Federation of German Industries (Bundesverband der Deutschen Industrie), have their headquarters in Cologne, and numerous major trade fairs are held annually in the large exhibition area in the Rheinpark.

**Transportation.** Cologne's geographical position and commercial importance have combined to make it a focal point for communications. The city is the busiest rail junction in the country. Autobahns radiate outward from the peripheral road that encircles the city. The Rhine harbour, important since Roman days, has become one of the larger inland ports in Germany. Small oceangoing craft use the river, and there are several ship lines for passengers. Intracity transport consists of streetcars, buses, and a subway system.

### ADMINISTRATION AND SOCIAL CONDITIONS

**Government.** Cologne is the administrative centre of one of the five major administrative districts of Nordrhein-Westfalen. The city is governed by an elected council, which is presided over by an *Oberbürgermeister* ("chief mayor"). Many governmental services, such as welfare, planning, transportation, and cultural affairs, however, are controlled by the state government. State services

**Education.** The University of Cologne, founded in 1388, was dissolved in 1798 (during the period when the French occupied the city) and refounded in 1919. Teacher-training colleges, a school of sports, and colleges for the study of music, engineering, administration, and other professions and trades are also located in the city.

### CULTURAL LIFE

Cologne is rich in museums and galleries. These include the Wallraf-Richartz and Ludwig museum complex, with an exceptionally comprehensive collection ranging from paintings of the medieval Cologne school to contemporary art; the Schnütgen Museum of medieval ecclesiastical art; the Museum für Ostasiatische Kunst, with artworks from China and Japan; and the Rautenstrauch-Joest Museum, with ethnological collections. The Römisch-Germanisches Museum houses artifacts from the period of the migrations of the Germanic peoples and that of the Roman occupation. Special exhibitions are held in the Josef-Haubrich-Kunsthalle exhibition centre near the Neumarkt.

Throughout most of the year, Cologne provides a variety of musical programs. Particularly notable are the Gürzenich concerts and those held in the concert hall of the Westdeutscher Rundfunk (WDR; "West German Radio"), the high reputation of the latter being largely due to the WDR's encouragement of contemporary music. A full repertoire is offered in theatre and opera as well, and the municipal theatre has its own ballet ensemble.

The annual Rhenisch pre-Lenten carnival is celebrated with great ceremony, culminating in the Rosenmontag ("Rose Monday") festival before Ash Wednesday. A folk festival, the Kölscher Fasteleer, is peculiar to the city. Notable citizens have included the Dominican scholar Albertus Magnus, the novelist Heinrich Böll, and the statesman Konrad Adenauer.

## History

### EARLY SETTLEMENT AND MEDIEVAL GROWTH

After Julius Caesar destroyed the Eburones in 53 BC, the Roman general Agrippa colonized the area with another tribe called the Ubii, who came from the right bank of the Rhine. A fortified settlement was established on the site in about 38 BC. This was the birthplace of Agrippina the Younger, who was the wife of the emperor Claudius, and it was at her request that the title of Roman colony was conferred upon the town in AD 50. It was named Colonia

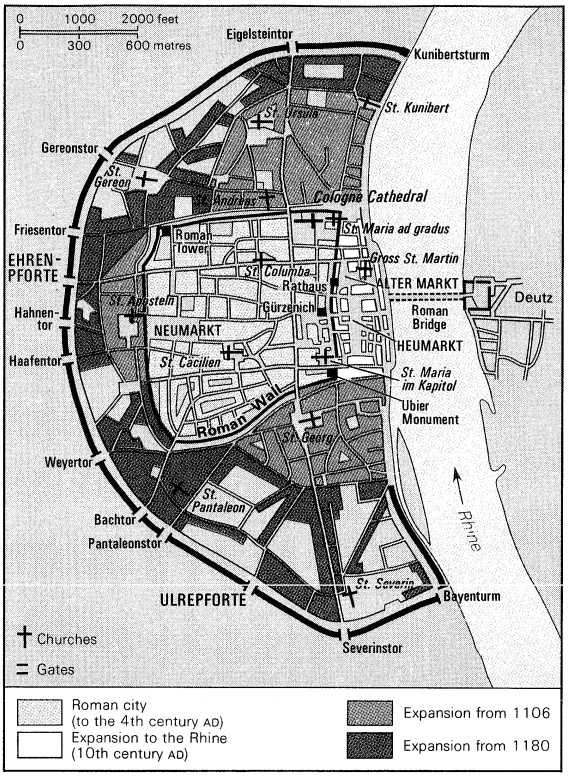
Claudia Ara Agrippinensium, shortened to Colonia; later it was made the headquarters of the governor of Lower Germany. After AD 258 it was for a time the capital of a splinter empire ruled by Postumus and comprising Gaul, Britain, and Spain. In 310 the emperor Constantine the Great built a castle and a permanent bridge to it across the Rhine. Ceramics and glass were manufactured in Cologne in Roman times. About 456 it was conquered by the Franks, and it soon became the residence of the kings of the Ripuarian part of the Frankish kingdom.

A Christian community existed in Cologne probably as early as the 2nd century, and the town is first mentioned as a bishopric in 313. Charlemagne made it an archbishopric in the late 8th century; by the 10th century the archbishop dominated the city, receiving a wide range of tolls, customs duties, and other payments. The city's industry and trade grew during the Middle Ages, especially from about the 10th century, and increasingly bitter conflicts developed between the wealthy merchants and the archbishop. The former sought commercial and political freedom, the latter the preservation of his temporal power, which was augmented from the 13th century when the archbishop became one of the electors privileged to choose the German king. It was not until the Battle of Worringen, in 1288, that the Archbishop was finally defeated, and the city of Cologne secured full self-government. From that time, Cologne was, in fact, a free imperial city, although it was only officially recognized as such in 1475.

Prosperity in the Middle Ages

Until the end of the 14th century, the government of the city was in the hands of the wealthy patricians; but in 1396, after a bloodless revolution, a new municipal constitution was established under which the 22 branches of the guilds became the basis of the government, for they elected a council that had power over all internal and external affairs.

This medieval period was a splendid one for Cologne. It was a prominent member of the mercantile Hanseatic League, and its merchants had probably the most extensive connections and the most varied trade of all the German towns. Crafts included textile manufacturing, bookmaking, leatherworking, enameling, and metalworking, the work of Cologne's goldsmiths being particularly fine. The arts and religion flourished there also. Three of the greatest Roman Catholic scholars and theologians of the medieval Scholastic movement—Albertus Magnus, Thomas Aquinas, and John Duns Scotus—all taught in Cologne's schools. After the Thirty Years' War (1618–48), however, the city declined. As late as 1794, when the French occupied Cologne, public Protestant services were still banned, and the city has remained predominantly Roman Catholic. The Jewish community, which had existed from the time of Constantine the Great, was expelled in 1424, and until 1794 no Jew was allowed to remain overnight in the city.




Cologne in 1752.

Adapted from J. von Reinhardt, *Handbuch der historischen Statten Deutschlands* (1963); Alfred Kroner Verlag

## THE 19TH AND 20TH CENTURIES

Cologne's history as a free imperial city ended when it was taken by France in 1794; and, when the Archbishop Elector died in 1801, the see was left vacant and not restored until 1821. In 1815 Cologne passed to Prussia, and from that time a new era of prosperity began. A wide range of industries flourished, and a chamber of commerce was established, the oldest in Germany. When the railways were built, its geographical position made it an ideal railway centre. The population grew from 41,685 in 1801 to 372,529 in 1900. Liberal points of view were represented in the 19th century by the *Rheinische Zeitung,* edited (1842–43) by Karl Marx and Moses Hess, while the Socialist *Neue Rheinische Zeitung* was edited (1848–49) by Marx, Friedrich Engels, and Ferdinand Freiligrath.

The city's growth was interrupted by World War I. Under Konrad Adenauer—the future chancellor of West Germany who was *Oberbürgermeister* of Cologne from 1917 until he was deposed by the Nazis in 1933—growth resumed, especially in suburban areas and in the laying out of new industrial parks. By 1939 the population had reached 768,352. In World War II, Cologne sustained 262 air raids. There were 20,000 casualties, and the city was left in ruins, with nearly all the dwellings in the old town damaged and 91 out of 150 churches destroyed. In March 1945, the war's end for Cologne, the population had sunk to 40,000. By December, however, there were 447,000 in the city, and the population continued to rise rapidly while the work of clearance and reconstruction was undertaken. Since the war the process of growth has continued, with the development of new industrial areas and satellite towns and the improvement of transportation.

World War II damage

BIBLIOGRAPHY

*General:* WILLEHAD PAUL ECKERT, *Köln: Stadt am Rhein zwischen Tradition und Fortschritt,* 3rd rev. ed. (1977); KURT KAYSER and THEODOR KRAUS (eds.), *Köln und die Rheinlande* (1961); J. KLERSCH, *Volkstum und Volksleben in Köln,* 3 vol. (1965–69), a sociological study; ERHARD SCHLEITER and RUDOLF BARTEN, III, *Köln/Cologne* (1982), a largely photographic study; and HELMUT SIGNON, *Getting to Know Cologne* (1977).

*History:* HERMANN KELLENBENZ and KLARA VAN EYLL (eds.), *Zwei Jahrtausende Kölner Wirtschaft,* 2 vol. (1975), a massive study of Cologne's economic history; HERMANN SCHMITZ, *Colonia Claudia Ara Agrippinensium* (1956); LEONARD ENNEN and GOTTFRIED ECKERTZ (eds.), *Quellen zur Geschichte der Stadt Köln,* 6 vol. (1860–79, reprinted 1970); *Köln, das Reich und Europa* (1971), a collection of essays on Cologne's involvement in politics, law, and business during the Middle Ages; ARNOLD STELZMANN, *Illustrierte Geschichte der Stadt Köln,* 8th rev. ed. (1978), an illustrated history; and KARL MARX and FREDERICK ENGELS, *The Cologne Communist Trial,* trans. from the German and ed. by RODNEY LIVINGSTONE (1971).

*Art and architecture:* PAUL CLEMEN, FRITZ WITTE, and HEINRICH NEU (eds.), *Der Dom zu Köln,* 2nd enlarged ed. (1938, reprinted 1980); LUDWIG ARNTZ, HEINRICH NEU, HANS VOGTS (eds.), *Die Ehemaligen Kirchen, Klöster, Hospitäler und Schulbauten der Stadt Köln* (1937, reprinted 1980); HANS VOGTS (ed.), *Die profanen Denkmäler der Stadt Köln* (1930, reprinted 1980), and other titles in the series "Die Kunstdenkmäler der Stadt Köln," ed. by PAUL CLEMEN, on the artistic aspects of various monuments in Cologne; HILTRUD KIER (ed.), *Köln: Kunstführer* (1980); WERNER SCHÄFKE, *Kölns romanische Kirchen: Architektur, Ausstattung, Geschichte* (1984); HANS VOGTS, *Köln im Spiegel seiner Kunst* (1950); ALBERT VERBEEK, *Kölner Kirchen,* 2nd rev. ed. (1969), a history of the art of church building in Cologne; W.D. ROBSON-SCOTT, *The Literary Background of the Gothic Revival in Germany* (1965), on aspects of the cathedral; and JAKOB SCHLAFKE, *The Cathedral of Cologne* (1980; originally published in German, 1980), a photographic study.

(Ma.Ko./H.Sr./C.A.McI.)

# Colombia

The Republic of Colombia (República de Colombia) occupies the northwestern corner of the South American continent. Its 1,000 miles (1,600 kilometres) of coast to the north are bathed by the waters of the Caribbean Sea, and its 800 miles of coast to the west are washed by the Pacific Ocean. The country is bordered by Panama, which divides the two bodies of water, on the northwest; Venezuela and Brazil on the east; and Peru and Ecuador on the south. Its area of 440,831 square miles (1,141,748 square kilometres) includes the San Andrés y Providencia archipelago, located off the Nicaraguan coast in the Caribbean, some 400 miles northwest of the Colombian mainland. The population is largely concentrated in the mountainous interior, where Bogotá, the national capital, is situated on a high plateau in the Andes Mountains.

The only American nation that is named for this discoverer of the New World, Colombia presents a remarkable study in contrasts, both in its geography and its society. The lofty snow-tipped peaks of the country's interior cordilleras tower high above equatorial forests and savannas where surviving Indian groups still follow the lifeways and traditions of their ancestors. In the cooler mountains, at intermediate elevations, modern cities are juxtaposed with traditional rural landscapes where mestizo farmers cultivate their small plots of coffee, corn (maize), and other crops. The more accessible Atlantic Lowlands, dominated by large livestock haciendas and a tri-ethnic population, have a distinctively different character.

Colombia strongly reflects its history as a colony of Spain. It is often referred to as the most Roman Catholic of the South American countries, and most of its people are proud of the relative purity of their Spanish language. Its population is heavily mestizo (of mixed European and Indian descent) but with substantial negroid admixture on the coasts. The Colombian economy is heavily dependent on the production of coffee, a crop that is highly sensitive to fluctuations in the world market. Agriculture still occupies a substantial part of the population, but employment in urban manufacturing and service industries has become increasingly important. About one-third of the inhabitants live in the six largest metropolitan areas. Political instability is closely tied to the unequal distribution of wealth, which, again, has strong historic antecedents. Illicit trade in drugs, a complex and growing issue, has become a major disruptive factor in the economy. The population has grown rapidly, and Colombia now ranks with Argentina as one of the two most populous nations in Spanish-speaking South America.

This article is divided into the following sections:


Physical and human geography 571
- The land 571
  - Relief
  - Drainage and soils
  - Climate
  - Plant and animal life
  - Settlement patterns
- The people 576
- The economy 576
  - Resources
  - Agriculture, forestry, and fishing
  - Industry
  - Finance and trade
  - Transportation
- Government and social conditions 578
  - Government
  - Education
  - Welfare and health
- Cultural life 579
  - Cultural origins
  - The arts
  - Cultural institutions
  - Recreation
  - Press and broadcasting

History 580
- Preconquest 580
- Conquest 580
- Colonial period 580
  - The audiencia
  - Viceroyalty of New Granada
- Revolution and independence 581
- The republic to 1930 581
  - Conservative–Liberal struggle, 1840–80
  - The return of the Conservatives, 1880–1930
- Colombia since 1930 582
  - The era of the Liberals, 1930–46
  - Civil unrest, dictatorship, and democratic restoration, 1946–70
  - The Conservative–Liberal rule

Bibliography 583

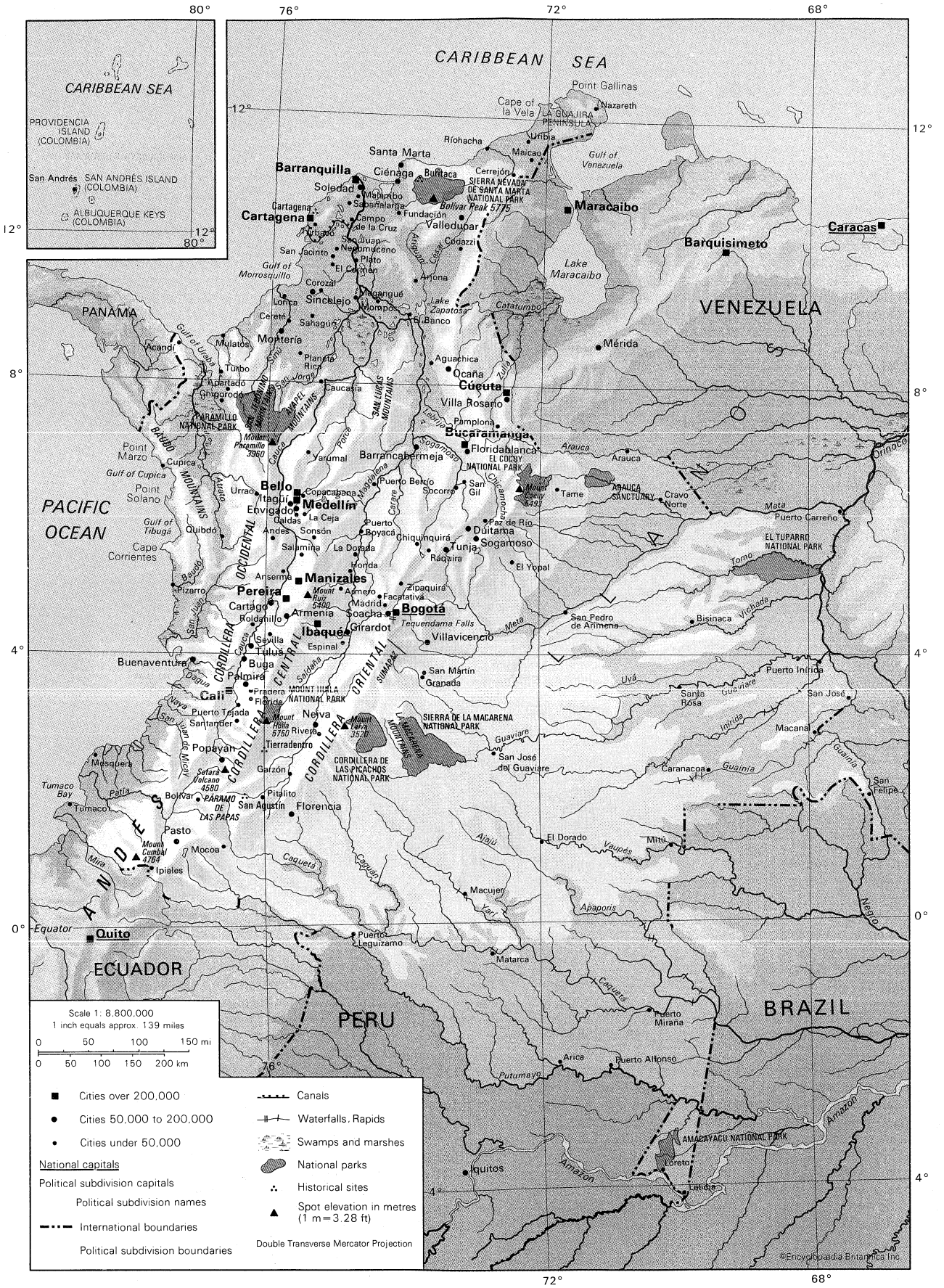

## Physical and human geography

### THE LAND

**Relief.** Few countries boast such striking physical variety as does Colombia. Its broken, rugged topography, together with its location near the Equator, creates an extraordinary diversity of climates, vegetation, soils, and crops. The Andean cordillera, one of the world's great mountain ranges, dominates the landscape of the western part of the country, where most of the people live. North of the border with Ecuador the cordillera flares out into three distinct parallel ranges. Two great river valleys, those of the Magdalena and the Cauca, separate them and provide avenues of penetration from the Atlantic coastal lowlands into the heart of the country. Volcanic activity in the geologic past blocked the middle course of the Cauca River to form a great lake that once filled the western inter-Andean trough for some 120 miles south of Cartago. The river eventually broke through the dam to leave the level floor of the Cauca Valley at some 3,000 feet (980 metres) above sea level; today it is one of the nation's most productive agricultural areas.

The Colombian cordilleras belong to the northern portion of the great Andean mountain system, which extends along the Pacific coast of South America. The Andes are among the world's most youthful mountain ranges and among the highest. The geology and history of this northern sector is less well understood than that of the central and southern parts. It is clear, however, that the entire cordillera has evolved through the subduction of the crumpled margin of the Nazca Plate and, to the north, the Caribbean Plate, under the more rigid but lighter South American Plate, represented by the ancient Guiana and Brazilian shields. These tectonic forces, similar to those found elsewhere around the Pacific rim, continue to operate, as is evidenced by the high frequency of often destructive earthquakes. At the Pasto Massif, near the Ecuador border, the mountains divide into the Cordillera Occidental (Western Range), which runs parallel to the Pacific coast, and the Cordillera Central (Central Range), which, with its numerous volcanoes, forms the backbone of the system and runs in a generally southwest to northeast direction. At the Great Colombian Massif of the Cordillera Central, near the San Agustín Archeological Park, the Cordillera Oriental (Eastern Range) branches off in a more decidedly northeasterly direction.

Geology of the north Andes


CARIBBEAN SEA
PROVIDENCIA ISLAND (COLOMBIA)
SAN ANDRÉS ISLAND (COLOMBIA)
ALBUQUERQUE KEYS (COLOMBIA)
Point Gallinas
Cape of la Vela
LA GUAJIRA PENINSULA
Nazareth
Ríohacha
Uribia
Maicao
Gulf of Venezuela
Santa Marta
Barranquilla
Ciénaga
Buritaca
Cerrejón
SIERRA NEVADA DE SANTA MARTA NATIONAL PARK
Bolívar Peak 5775
Soledad
Malambo
Sabanalarga
Cartagena
Fundación
Valledupar
Maracaibo
Caracas
Barquisimeto
Lake Maracaibo
Gulf of Morrosquillo
Corozal
Sincelejo
Lorica
Cereté
Sahagún
Montería
Mulatos
Acandí
Turbo
Apartadó
Chigorodó
PARAMILLO NATIONAL PARK
Mount Paramillo 3960
Gulf of Urabá
PANAMA
Planeta Rica
Caucasia
Mompós
El Banco
Lake Zapatosa
Catatumbo
Aguachica
Ocaña
Cúcuta
Villa Rosario
Pamplona
Bucaramanga
Floridablanca
Barrancabermeja
Yarumal
Puerto Berrío
Socorro
San Gil
EL COCUY NATIONAL PARK
Mount Cocuy 5493
Arauca
ARAUCA SANCTUARY
Cravo Norte
Tame
Mérida
VENEZUELA
Orinoco
Meta
Puerto Carreño
EL TUPARRO NATIONAL PARK
Tomo
Point Marzo
Cupica
Gulf of Cupica
Point Solano
Gulf of Tribugá
Quibdó
Cape Corrientes
PACIFIC OCEAN
Bello
Itagüí
Medellín
Envigado
Caldas
La Ceja
Andes
Sonsón
Urrao
Copacabana
Puerto Boyacá
Chiquinquirá
Paz de Río
Duitama
Sogamoso
Tunja
Ráquira
Salamina
La Dorada
Honda
El Yopal
Anserma
Manizales
Pereira
Mount Ruiz 5400
Armenia
Cartago
Roldanillo
Quimbaya
Zipaquirá
Armero
Facatativá
Madrid
Soacha
Bogotá
Tequendama Falls
Girardot
Ibagué
Espinal
Villavicencio
San Pedro de Arimena
Bisinacu
Vichada
Puerto Inírida
Sabana Larga
Sabana
Sevilla
Tuluá
Buga
Palmira
Cali
Buenaventura
Dagua
Cauca
San Juan
Patía
Mount Huila 5750
Nevado
Mount Tatacoa 3520
SIERRA DE LA MACARENA NATIONAL PARK
LA MACARENA MOUNTAINS
Guaviare
San José del Guaviare
Inírida
Macanal
Caranacoa
Guainía
San Felipe
Puerto Tejada
Santander
Neiva
Riverá
Popayán
Tierradentro
Garzón
Mosquera
Tumaco Bay
Tumaco
Bolívar
Pitalito
San Agustín
Florencia
PÁRAMO DE LAS PAPAS
Nariño
Pasto
Mocoa
Mount Cumbal 4764
Ipiales
Mira
Caquetá
Putumayo
Aljajú
El Dorado
Vaupés
Mitú
Macujer
Apaporis
Vaupés
Equator
Quito
ECUADOR
Puerto Leguízamo
Matarca
PERU
BRAZIL
Puerto Miraña
Arica
Puerto Alfonso
Iquitos
Amazon
AMACAYACU NATIONAL PARK
Loreto
Leticia
CORDILLERA OCCIDENTAL
CORDILLERA CENTRAL
CORDILLERA ORIENTAL
CORDILLERA DE LOS PICACHOS NATIONAL PARK
ANDES
Scale 1: 8,800,000
1 inch equals approx. 139 miles
0 50 100 150 mi
0 50 100 150 200 km
Cities over 200,000
Cities 50,000 to 200,000
Cities under 50,000
National capitals
Political subdivision capitals
Political subdivision names
International boundaries
Political subdivision boundaries
Canals
Waterfalls, Rapids
Swamps and marshes
National parks
Historical sites
Spot elevation in metres (1 m=3.28 ft)
Double Transverse Mercator Projection
©Encyclopaedia Britannica Inc.

**MAP INDEX**

**Political subdivisions**


Amazonas . . . . . 1 00 S 72 00 W
Antioquia . . . . . 7 00 N 75 30 W
Arauca . . . . . 6 40 N 71 00 W
Atlántico . . . . . 10 45 N 75 00 W
Bogotá . . . . . 4 15 N 74 10 W
Bolívar . . . . . 9 00 N 74 40 W
Boyacá . . . . . 5 30 N 72 30 W
Caldas . . . . . 5 15 N 75 30 W
Caquetá . . . . . 1 00 N 74 00 W
Casanare . . . . . 5 30 N 71 30 W
Cauca . . . . . 2 30 N 76 50 W
César . . . . . 9 20 N 73 30 W
Chocó . . . . . 6 00 N 77 00 W
Córdoba . . . . . 8 20 N 75 40 W
Cundinamarca . . . . 5 00 N 74 00 W
Guainía . . . . . 2 30 N 69 00 W
Guaviare . . . . . 2 00 N 72 15 W
Huila . . . . . 2 30 N 75 45 W
La Guajira . . . . . 11 30 N 72 30 W
Magdalena . . . . . 10 00 N 74 00 W
Meta . . . . . 3 30 N 73 00 W
Nariño . . . . . 1 30 N 78 00 W
Norte de Santander . . . . . 8 00 N 73 00 W
Putumayo . . . . . 0 30 N 76 00 W
Quindío . . . . . 4 30 N 75 40 W
Risaralda . . . . . 5 00 N 76 00 W
San Andrés y Providencia . . . . 12 30 N 81 45 W
Santander . . . . . 7 00 N 73 15 W
Sucre . . . . . 9 00 N 75 00 W
Tolima . . . . . 3 45 N 75 15 W
Valle del Cauca . . . 3 45 N 76 30 W
Vaupés . . . . . 0 15 N 70 45 W
Vichada . . . . . 5 00 N 69 30 W


**Cities and towns**


Acandí . . . . . 8 32 N 77 14 W
Aguachica . . . . . 8 19 N 73 38 W
Andes . . . . . 5 40 N 75 53 W
Anserma . . . . . 5 13 N 75 48 W
Apartadó . . . . . 7 54 N 76 39 W
Arauca . . . . . 7 05 N 70 45 W
Arica . . . . . 2 08 S 71 47 W
Arjona . . . . . 9 32 N 73 55 W
Armenia . . . . . 4 31 N 75 41 W
Armero . . . . . 4 58 N 74 54 W
Barrancabermeja . . 7 03 N 73 52 W
Barranquilla . . . . . 10 59 N 74 48 W
Bello . . . . . 6 20 N 75 33 W
Bisinaca . . . . . 4 30 N 69 40 W
Bogotá . . . . . 4 36 N 74 05 W
Bolívar . . . . . 1 50 N 76 58 W
Bucaramanga . . . . 7 08 N 73 09 W
Buenaventura . . . . 3 53 N 77 04 W
Buga . . . . . 3 54 N 76 17 W
Caldas . . . . . 6 05 N 75 38 W
Cali . . . . . 3 27 N 76 31 W
Campo de la Cruz . . . . . 10 23 N 74 53 W
Caranacoa . . . . . 2 25 N 68 57 W
Cartagena . . . . . 10 25 N 75 32 W
Cartago . . . . . 4 45 N 75 55 W
Caucasia . . . . . 8 00 N 75 12 W
Cereté . . . . . 8 53 N 75 48 W
Cerrejón . . . . . 11 02 N 72 39 W
Chigorodó . . . . . 7 41 N 76 42 W
Chiquinquirá . . . . . 5 37 N 73 50 W
Ciénaga . . . . . 11 01 N 74 15 W
Codazzi . . . . . 10 02 N 73 14 W
Copacabana . . . . . 6 21 N 75 30 W
Corozal . . . . . 9 19 N 75 18 W
Cravo Norte . . . . . 6 18 N 70 12 W
Cúcuta . . . . . 7 54 N 72 31 W
Cupica . . . . . 6 41 N 77 30 W
Duitama . . . . . 5 50 N 73 02 W
El Banco . . . . . 9 00 N 73 58 W
El Carmen . . . . . 9 43 N 75 08 W
El Dorado . . . . . 1 11 N 71 52 W
El Yopal . . . . . 5 21 N 72 23 W
Envigado . . . . . 6 10 N 75 35 W
Espinal . . . . . 4 09 N 74 53 W
Facatativá . . . . . 4 49 N 74 22 W
Florencia . . . . . 1 36 N 75 36 W
Florida . . . . . 3 21 N 76 15 W
Floridablanca . . . . 7 04 N 73 06 W
Fundación . . . . . 10 31 N 74 11 W
Garzón . . . . . 2 12 N 75 38 W
Girardot . . . . . 4 18 N 74 48 W
Granada . . . . . 3 34 N 73 45 W
Honda . . . . . 5 12 N 74 45 W
Ibagué . . . . . 4 27 N 75 14 W
Ipiales . . . . . 0 50 N 77 37 W
Itagüí . . . . . 6 10 N 75 36 W
La Ceja . . . . . 6 02 N 75 26 W
La Dorada . . . . . 5 27 N 74 40 W
Leticia . . . . . 4 09 S 69 57 W
Loreto . . . . . 3 48 N 70 15 W
Lorica . . . . . 9 14 N 75 49 W
Macanal . . . . . 2 45 N 67 58 W
Macujer . . . . . 0 24 N 73 07 W
Madrid . . . . . 4 44 N 74 16 W
Magangué . . . . . 9 14 N 74 45 W
Maicao . . . . . 11 23 N 72 13 W
Malambo . . . . . 10 52 N 74 47 W
Manizales . . . . . 5 05 N 75 32 W
Matarca . . . . . 0 30 S 72 38 W
Medellín . . . . . 6 15 N 75 35 W
Mitú . . . . . 1 08 N 70 03 W
Mocoa . . . . . 1 09 N 76 37 W
Mompós . . . . . 9 14 N 74 26 W
Montería . . . . . 8 46 N 75 53 W
Mosquera . . . . . 2 30 N 78 29 W
Mulatos . . . . . 8 39 N 76 44 W
Nazareth . . . . . 12 11 N 71 17 W
Neiva . . . . . 2 56 N 75 18 W
Ocaña . . . . . 8 15 N 73 20 W
Palmira . . . . . 3 32 N 76 16 W
Pamplona . . . . . 7 23 N 72 39 W
Pasto . . . . . 1 13 N 77 17 W
Paz de Río . . . . . 5 59 N 72 47 W
Pereira . . . . . 4 49 N 75 43 W
Pitalito . . . . . 1 51 N 76 02 W
Pizarro . . . . . 4 58 N 77 22 W
Planeta Rica . . . . . 8 25 N 75 36 W
Plato . . . . . 9 47 N 74 47 W
Popayán . . . . . 2 27 N 76 36 W
Pradera . . . . . 3 25 N 76 15 W
Puerto Alfonso . . . 2 11 S 70 59 W
Puerto Berrío . . . . . 6 29 N 74 24 W
Puerto Boyacá . . . 5 45 N 74 39 W
Puerto Carreño . . . 6 12 N 67 22 W
Puerto Inírida . . . . . 3 51 N 67 55 W
Puerto Leguízamo . . . . . 0 12 S 74 46 W
Puerto Miraña . . . . 1 20 S 70 19 W
Puerto Tejada . . . . 3 14 N 76 24 W
Quibdó . . . . . 5 42 N 76 40 W
Ráquira . . . . . 5 33 N 73 38 W
Ríohacha . . . . . 11 33 N 72 55 W
Rivera . . . . . 2 47 N 75 15 W
Roldanillo . . . . . 4 24 N 76 09 W
Sabanalarga . . . . 10 38 N 74 55 W
Sahagún . . . . . 8 57 N 75 27 W
Salamina . . . . . 5 25 N 75 29 W
San Andrés . . . . . 12 35 N 81 42 W
San Felipe . . . . . 1 52 N 67 06 W
San Gil . . . . . 6 33 N 73 08 W
San Jacinto . . . . . 9 50 N 75 08 W
San José . . . . . 3 15 N 67 20 W
San José de Guaviare . . . . . 2 35 N 72 38 W
San Juan Nepomuceno . . . . 9 57 N 75 05 W
San Martín . . . . . 3 42 N 73 42 W
San Pedro de Arimena . . . . . 4 37 N 71 42 W
Santa Marta . . . . 11 15 N 74 13 W
Santa Rosa . . . . . 3 32 N 69 48 W
Santander . . . . . 3 01 N 76 28 W
Sevilla . . . . . 4 16 N 75 57 W
Sincelejo . . . . . 9 18 N 75 24 W
Soacha . . . . . 4 35 N 74 13 W
Socorro . . . . . 6 29 N 73 16 W
Sogamoso . . . . . 5 43 N 72 56 W
Soledad . . . . . 10 55 N 74 46 W
Sonsón . . . . . 5 42 N 75 18 W
Tame . . . . . 6 28 N 71 44 W
Tuluá . . . . . 4 06 N 76 11 W
Tumaco . . . . . 1 49 N 78 46 W
Tunja . . . . . 5 31 N 73 22 W
Turbaco . . . . . 10 20 N 75 25 W
Turbo . . . . . 8 06 N 76 43 W
Uribia . . . . . 11 43 N 72 16 W
Urrao . . . . . 6 20 N 76 11 W
Valledupar . . . . . 10 29 N 73 15 W
Villa Rosario . . . . . 7 50 N 72 28 W
Villavicencio . . . . . 4 09 N 73 37 W
Yarumal . . . . . 6 58 N 75 24 W
Zipaquirá . . . . . 5 02 N 74 00 W


**Physical features and points of interest**


Ajajú, *river* . . . . . 0 59 N 72 20 W
Albuquerque Keys . . . . . 12 10 N 81 50 W
Amacayacu National Park . . . . 3 45 N 70 15 W
Amazon, *river* . . . . 4 00 S 70 00 W
Andes, *mountains* . . . . . 0 00 78 00 W
Apaporis, *river* . . . . 1 23 S 69 25 W
Arauca, *river* . . . . . 7 24 N 70 00 W
Arauca Sanctuary . . . . . 6 25 N 71 05 W
Ariguaní, *river* . . . . 9 35 N 73 46 W
Atrato, *river* . . . . . 8 17 N 76 58 W
Ayapel Mountains . . . . . 7 45 N 75 30 W
Baudó, *river* . . . . . 4 57 N 77 22 W
Baudó Mountains . . 6 00 N 77 05 W
Bolívar Peak . . . . 10 50 N 73 42 W
Buritaca, *historical site* . . . 11 02 N 73 58 W
Caguán, *river* . . . . 0 08 S 74 18 W
Caquetá, *river* . . . . 1 30 S 69 32 W
Carare, *river* . . . . . 6 48 N 74 06 W
Caribbean Sea . . 13 00 N 72 00 W
Cartagena, *historical site* . . . 10 25 N 75 32 W
Catatumbo, *river* . . 9 02 N 72 31 W
Cauca, *river* . . . . . 8 54 N 74 28 W
César, *river* . . . . . 9 00 N 73 58 W
Central, Cordillera, *mountains* . . . . . 5 00 N 75 00 W
Chicamocha, *river* . . . . . 6 46 N 73 12 W
Cocuy, Mount . . . . 6 25 N 72 18 W
Cordillera de las Picachos National Park . . . . 2 45 N 74 30 W
Corrientes, Cape . . 5 30 N 77 34 W
Cumbal, Mount . . . 0 57 N 77 52 W
Cupica, Gulf of . . . 6 35 N 77 25 W
Dagua, *river* . . . . . 3 52 N 77 04 W
El Cocuy National Park . . . . 6 25 N 77 05 W
El Tuparro National Park . . . . 5 20 N 68 30 W
Gallinas, Point . . . 12 28 N 71 40 W
Guainía, *river* . . . . . 2 01 N 67 07 W
Guaviare, *river* . . . . 4 03 N 67 44 W
Huila, Mount, *volcano* . . . . . 3 00 N 76 00 W
Inírida, *river* . . . . . 3 55 N 67 52 W
La Guajira Peninsula . . . . . 12 00 N 71 30 W
La Macarena Mountains . . . . . 2 45 N 73 55 W
La Vela, Cape of . . . . . 12 13 N 72 11 W
Las Papas, Páramo de, *upland* . . . . . 1 55 N 76 36 W
Lebrija, *river* . . . . . 8 08 N 73 47 W
Leiva, Mount . . . . . 2 54 N 74 48 W
Llanos, *plains* . . . . 5 00 N 70 00 W
Magdalena, *river* . . . . . 11 06 N 74 51 W
Marzo, Point . . . . . 6 50 N 77 42 W
Meta, *river* . . . . . 6 12 N 67 28 W
Mira, *river* . . . . . 1 36 N 79 01 W
Morrosquillo, Gulf of . . . . . 9 35 N 75 40 W
Mount Huila National Park . . . . 2 30 N 74 00 W
Naya, *river* . . . . . 3 14 N 77 30 W
Negro, *river* . . . . . 0 00 67 15 W
Occidental, Cordillera, *mountains* . . . . . 5 00 N 76 00 W
Oriental, Cordillera, *mountains* . . . . . 6 00 N 73 00 W
Orinoco, *river* . . . . 6 12 N 67 28 W
Párama de las Papas, *upland* . . . 1 55 N 76 36 W
Paramillo, Mount . . 7 04 N 75 55 W
Paramillo National Park . . . . 7 30 N 76 20 W
Patía, *river* . . . . . 2 13 N 78 40 W
Porce, *river* . . . . . 7 28 N 74 53 W
Providencia Island . . . . . 13 21 N 81 22 W
Putumayo, *river* . . . 3 07 S 67 58 W
Ruiz, Mount . . . . . 4 54 N 75 18 W
Saldaña, *river* . . . . 4 01 N 74 52 W
San Agustín, *historical site* . . . . 1 53 N 76 16 W
San Andrés Island . . . . . 12 32 N 81 42 W
San Jerónimo Mountains . . . . . 8 00 N 75 50 W
San Jorge, *river* . . . 9 07 N 74 44 W
San Juan, *river* . . . 4 03 N 77 27 W
San Juan de Micay, *river* . . . . . 3 05 N 77 32 W
San Lucas Mountains . . . . . 8 00 N 74 20 W
Sierra de la Macarena National Park . . . . 2 15 N 73 45 W
Sierra Nevada de Santa Marta National Park . . . 11 00 N 73 40 W
Sinú, *river* . . . . . 9 24 N 75 49 W
Sogamoso, *river* . . 7 13 N 73 56 W
Solano, Point . . . . . 6 18 N 77 29 W
Sotará Volcano . . . 2 12 N 76 31 W
Sumapaz, *upland* . . . . . 3 45 N 74 25 W
Tequendama Falls . . . . . 4 35 N 74 18 W
Tibugá, Gulf of . . . 5 45 N 77 20 W
Tierradentro, *historical site* . . . . 2 33 N 76 03 W
Tomo, *river* . . . . . 5 20 N 67 48 W
Tumaco Bay . . . . . 1 55 N 78 45 W
Urabá, Gulf of . . . . 8 25 N 76 53 W
Uvá, *river* . . . . . 3 41 N 70 03 W
Vaupés, *river* . . . . . 0 02 N 67 16 W
Venezuela, Gulf of . . . . . 11 30 N 71 00 W
Vichada, *river* . . . . 4 55 N 67 50 W
Yarí, *river* . . . . . 0 23 N 72 16 W
Zapatosa, Lake . . . 9 05 N 73 50 W
Zulia, *river* . . . . . 9 04 N 72 18 W


Of the three ranges, the nonvolcanic Cordillera Occidental, which forms the barrier between the Cauca Valley and the rain-drenched Pacific coast, is the lowest and least populated. Two passes of less than 5,000 feet between Cali and Buenaventura on the Pacific coast mark the lowest depressions in the range. Elsewhere the crest is much higher, exceeding 12,000 feet at Mount Paramillo in the department of Antioquia. From there the Cordillera Occidental fingers north into the three distinct *serranías* of Abibe, San Jerónimo, and Ayapel, forested ranges that drop gradually toward the piedmont plains of the Caribbean littoral. A lesser topographic feature on the Pacific coast, the Baudó Mountains, separated from the Cordillera Occidental by the valley of the Atrato River, which empties into the Caribbean Gulf of Urabá, represents a southward extension of the Isthmus of Panama.

The Cordillera Central

The Cordillera Central is the highest of the Andean ranges of Colombia, rising to an average height of 10,000 feet. It is a continuation of the Ecuadorian volcanic structure. Crystalline rocks are exposed at several places on its flanks and are the foci of localized gold and silver deposits. Sandstones and shales of the Tertiary Period

(from 65,000,000 to 2,500,000 years ago) are also a part of the older basement that has been capped by ash and lava derived from some 20 volcanoes of the Quaternary Period (within the past 2,500,000 years). Several of the latter reach well into the zone of permanent snow, above 15,000 feet. The highest are Mount Huila (18,865 feet), southeast of Cali, and the Ruiz-Tolima complex (17,716 feet) between Manizales and Ibagué. The fertile ash from their eruptions has produced the high, cool plateaus of Nariño department and the often steep slopes to the north that support much of Colombia's coffee production. In November 1985 Mount Ruíz erupted, melting the snow and ice that covered it and sending great mudflows downslope, destroying the city of Armero and killing more than 25,000, in one of the country's greatest catastrophes.

North of Mount Ruíz, near Sonsón in the department of Antioquia, the volcanic Cordillera Central gives way to the deeply weathered, granitic Antioquia batholith (an exposed granitic intrusion), a tableland averaging some 8,000 feet above sea level. It is divided into two parts by the deep transverse cleft of the Porce River, which occupies the U-shaped valley in which is situated the expanding metropolis of Medellín, Colombia's second city. The batholith contains gold-bearing quartz veins, which were the source of the placer gravels that gave rise to an active colonial mining economy. Beyond Antioquia the lower, remote San Lucas Mountains extend northward toward the confluence of the Magdalena and Cauca rivers.

The Cordillera Oriental

The massive Cordillera Oriental, separating the Magdalena Valley from the eastern plains (the Llanos), is composed chiefly of folded and faulted marine sediments and older schists and gneisses. Narrow to the south, it broadens out in the high, unsettled massif of Sumapaz, with elevations up to 13,000 feet. High plateaus were formed in the Quaternary Period by the deposition of sediments in depressions that had been occupied by lakes. The most important of these is the savanna area called the Sabana de Bogotá, site of the national capital. Farther northeast beyond the deep canyons cut by the Chicamocha River and its tributaries, the Cordillera Oriental culminates in the towering Sierra Nevada del Cocuy (18,028 feet). Beyond this point, near Pamplona, the cordillera splits into two much narrower ranges, one extending into Venezuela, the other, the Perijá Mountains, forming the northern boundary range between Colombia and Venezuela. The Perijás then descend northward toward the Caribbean to the arid Guajira Peninsula, the northernmost extension of the Colombian mainland.

The isolated Sierra Nevada de Santa Marta, an imposing fault-bounded granitic massif rising 18,947 feet (5,775 metres), the highest mountain in the country, ascends abruptly from the Caribbean littoral to snow- and ice-covered summits. The Atlantic Lowlands spread out southward behind it. Although it is a distinct geomorphic unit and not a part of the Andes, some geologists have suggested that it might be considered an extension of the Cordillera Central, from which it is separated by the Mompós depression in the lower Magdalena Valley.

The coastal plains

The steep and rugged Andean mountain masses and the high intermontane basins descend into plains that extend along the Caribbean and Pacific coasts and across the eastern interior toward the Orinoco and Amazon river systems. From the shores of the Caribbean Sea inland to the lower spurs of the three major cordilleras extends a slightly undulating savanna surface of varying width, generally known as the Atlantic Lowlands (also called the Caribbean Coastal Lowlands). Dotted with hills, and with extensive tracts of seasonally flooded land along the lower Magdalena and the Sinú rivers, it surrounds the inland portion of the Sierra Nevada de Santa Marta. A much narrower lowland apron extends along the Pacific shoreline from the point of Cape Corrientes southward to the Ecuadorian border.

A wide range of features characterize the country's two coastlines. Steep and articulated bays, inlets, capes, and promontories accentuate the shoreline on the Pacific side toward the Panama border and on the Caribbean side where the sea beats against the base of the Sierra Nevada de Santa Marta. These features are interspersed with sandy beaches, along with barrier islands and brackish lagoons.

The eastern lowland

The eastern two-thirds of the country lying beyond the Andes differs from cordilleran Colombia in practically all aspects of physical and human geography. The eastern lowland extends from the Venezuela boundary along the Arauca and Meta rivers in the north to the Peruvian–Ecuadorian border stream, the Putumayo, some 600 miles to the south, and from the base of the Cordillera Oriental eastward to the Orinoco–Negro river line, a distance of more than 400 miles. A region of great topographical uniformity, it is divided into two contrasting natural landscapes by a major vegetation boundary. In southern Colombia the Amazonian rain forest, or *selva,* reaches its northern limit. From the Guaviare River northward the plains between the Andes and the Orinoco River are mostly grass covered, forming the largest savanna complex in tropical America. This part of the lowland is called the Llanos Orientales or simply the Llanos.

In the central part of the plain, between the Guaviare and Caquetá rivers, the eroded rocks of the ancient Guiana shield are exposed, producing a broken topography of low, isolated mountains, tablelands, and buttes with rapids in the streams. This slightly higher ground forms the watershed between the Amazon and Orinoco systems. Some 60 miles south of Villavicencio the elongated, forested La Macarena Mountains rise 8,000 feet from the surrounding lowlands, an isolated tropical ecosystem.

**Drainage and soils.** In Colombia's rugged terrain the rivers have been historically important as routes of transportation and settlement. By far the most important river system is the Magdalena. Its drainage basin, including that of its major tributary, the Cauca, covers some 100,000 square miles, or about 22 percent of the surface of the country. Within it are found most of the nation's socioeconomic activity and more than three-fourths of its population. Originating in the Andean Páramo de Las Papas, the Magdalena flows northward in the structural depression between the Cordilleras Central and Oriental for almost 1,000 miles to empty into the Caribbean near Barranquilla. The Dique Canal, begun during the colonial period, links its lower course with the coastal city of Cartagena. The Cauca River, which contributes a substantial part of its total flow, rises in the mountains south of Popayán and, after passing through the floor of the Cauca Valley near Cali, occupies deep canyons in most of its passage through the departments of Caldas and Antioquia before emerging onto the floodplain of the lower Magdalena.

The Magdalena River system

The Magdalena, a shallow, braiding stream in its upper and middle course, served as a major transport artery for most of the country's history, but deforestation and soil erosion has led to silting and increased flow variation so that its role has become less significant. Because of its rapids, the Cauca has never been of much importance for the moving of goods. Among the major affluents of the Magdalena, besides the Cauca, are the Sogamoso, César, San Jorge, Saldaña, Lebrija, and Carare rivers. The Sinú and the Atrato are other major streams that flow directly into the Caribbean.

The great eastern watershed is subdivided into two sections, the waters flowing into the Orinoco and the Amazon rivers, which carry them to the Atlantic Ocean. The Arauca and Meta, the lower reaches of which cross into Venezuela, and the Vichada, Inírida, and Guaviare are among the main rivers that flow into the Orinoco. Among the streams that flow into the Amazon are the Vaupés, Caquetá, and Putumayo. The rivers that flow into the Pacific are relatively short, descending rapidly from the Cordillera Occidental to the sea. They carry large volumes of water, however, because they drain areas of extremely heavy rainfall. Among the rivers belonging to the Pacific watershed are the Baudó, San Juan, Dagua, Naya, San Juan de Micay, Patía, and Mira, which rises in Ecuador.

The wide variety of soils encountered in the country reflects climatic, topographic, and geologic conditions. Those best suited for modern, mechanized agriculture are the alluvial soils found in the principal river valleys, such as the Magdalena, Cauca, Sinú, César, and Ariguaní. The former lake beds of some of the inter-Andean basins, such as the Sabana de Bogotá and the Ubaté and Chiquinquirá

valleys, also fall into this category. Elsewhere, soils of volcanic origin, especially in the coffee-growing districts of the Cordillera Central, can be exceptionally productive if protected from erosion. The Quindío department, west of Bogotá, is especially renowned for its rich soils.

**Climate.** Because of the country's close proximity to the Equator, its climate is generally tropical and isothermal (without any real change of seasons). Temperatures vary little throughout the year. The only genuinely variable climatic element is the amount of annual precipitation. Climatic differences are related to altitude and the displacement of the Intertropical Convergence Zone between the two major air masses from which the Northeast Trade Winds and the Southeast Trade Winds originate. Human settlement is more oriented to vertical zoning in Colombia than anywhere else in Latin America.

The climate of the tropical rain forest in the Amazon region, the northern Pacific coast, and the central Magdalena Valley is marked by an annual rainfall of more than 100 inches (2,500 millimetres) and annual average temperatures above 74° F (23° C). A tropical monsoon climate, marked by one or more dry months but still supporting rain forest vegetation, occurs along the southern Pacific coast, on the Caribbean coast, and at places in the interior in the Quindío department and near Villavicencio.

The tropical savanna conditions of alternately wet and dry seasons comprise the predominant climate of the Atlantic Lowlands; the dry season occurs from November to April, and the wet season (broken by dry periods) from May to October. This climate is found also in the Llanos region and in part of the upper Magdalena Valley. It is characterized by an annual rainfall of 40 to 70 inches and annual average temperatures usually above 74° F (23° C). The dry season, accompanied by dust and wind, coincides with the true winter of the Northern Hemisphere.

A drier savanna climate prevails on the Caribbean littoral from the Gulf of Morrosquillo to La Guajira Peninsula in the northeast. The rains normally occur in two brief periods (in April and in October to November, respectively) but rarely exceed 30 inches annually. The average temperature is hot—more than 81° F (27° C)—with the daily range greatest where the humidity is low. This type of climate also occurs in the rain shadows of the deep gorges of such rivers as the Patía, Cauca, Chicamocha, and Zulia and in parts of the upper Magdalena Valley. The climate reaches near-desert conditions in the far northern department of La Guajira.

In the mountain regions temperature is directly related to altitude. Average temperatures decrease uniformly about 3° F per 1,000 feet of ascent (0.6° C per 100 metres). Popular terminology recognizes distinct temperature zones (*pisos térmicos*), which are sometimes referred to as *tierra caliente* (up to about 3,000 feet); *tierra templada* or *tierra del café* (3,000 to 6,500 feet); and *tierra fría* (6,500 to 10,000 feet). The majority of Colombians live in the interior cordilleras in the *tierra templada* and the *tierra fría* zones. The *tierra templada* has moderate rainfall and temperatures between 65° F (18° C) and 75° F (24° C). In the *tierra fría,* Bogotá, at 8,660 feet above sea level, has an average of 223 days of precipitation, although the average rainfall is scarcely 40 inches. Its average temperature is 57° F (14° C). The climate of the high mountain regions—the *páramos,* ranging from 10,000 feet to about 15,000 feet—is characterized by average temperatures below 50° F (10° C), fog, overcast skies, frequent winds, and light rain or drizzle. At altitudes above 15,000 feet there is perpetual snow and ice.

The *páramos*

**Plant and animal life.** The diversity of life forms and life zones in Colombia has impressed observers since the days of the German explorer Alexander von Humboldt. The complex pattern of climate, soil, and topography has produced an extraordinary range of plants and plant communities that vary through both vertical and horizontal zones. They range from the mangrove swamps of the coasts, the desert scrub of La Guajira, the savanna grasslands and gallery ecosystems of the Atlantic Lowlands and the eastern Llanos, and the rain forest of Amazonia and the Chocó region to the widely diverse and complex montane ecosystems of the Andean slopes.

Human intervention has vastly altered what must have been the original vegetation of the Atlantic Lowlands and the Andean region. Forests probably originally covered all but the highest and driest areas, where the soils were unsuited to support them. Today, in the inner Andes, they are restricted to the steepest, most inaccessible slopes and to areas of especially high rainfall. Elsewhere pasture, crops, or degraded scrub and grass have replaced the original cover of broadleaf evergreen trees. The first chroniclers often described the Andes as sparsely wooded, a condition they usually attributed to Indian agriculture and burning. In more recent times, with the increasing number of European cattle, the area in grassland has been vastly extended, both in the mountains and on the Atlantic Lowlands. Introduced grass species of African origin are particularly conspicuous.

Even in the most lush forest tracts that remain, such as those that flank the outer sides of the eastern and western ranges, there is much evidence of earlier human occupation. These wet montane forests are characterized by lianas, mosses, orchids, and bromeliads and by such economically valued plants as cinchona; the latex-bearing balata; ivory nut, or tagua; and the giant American bamboo. Lumbering has had a minor role because of the singular difficulty of access as well as the absence of forests of any one commercial species. In modern times technological advances have made possible the exploitation of forest species in accessible parts of the Atrato River basin and on the Pacific coast near Buenaventura.

The distinctive *páramo* biome of the equatorial high mountains reaches its greatest development in Colombia. This alpine-type vegetation is characterized by tussock grasses, cushion plants, and the treelike *frailejón* (*Espeletia*), a curious looking hairy-leafed genus of some 50 different species. Fire-resistant and adapted to low temperatures and high humidity, it gives special character to the *páramo* landscape. The lower *páramo,* below 12,000 feet, is a transitional belt in which scattered clumps of trees occur. Despite its bleak and forbidding climate, much of the *páramo* has been significantly altered by human activity, especially wood cutting and burning to promote better grazing. Agriculture has also impinged on its lower reaches, but extensive tracts remain relatively untouched by humans.

The *páramo* biome

The animal life of the forests of the Amazon and Pacific coastal Chocó regions is particularly rich and has supported a considerable export trade to North American and European zoos. It includes anteaters, sloths, several monkey species, tapirs, peccaries, the spectacled bear, deer, and such large tropical rodents as the agouti, paca, and capybara. Carnivores include pumas and jaguars, which were considered endangered species by the 1980s, and raccoons.

Amazonian fauna

Bird habitats are also influenced by altitude, and many species are specific to narrow altitudinal bands, ranging upward and downward only very short distances. The extremely lush birdlife encompasses more than 1,500 species, including toucans, hummingbirds, and those that migrate from North America. Among the reptiles, turtles, lizards, snakes, caimans, and crocodiles abound. Some quite unusual species inhabit the land, including earthworms that grow up to six feet in length. Freshwater fish include catfish, *bocachica,* and characin (small, brightly coloured tropical fish). Electric eels also inhabit the inland waters. The Magdalena and Cauca once supported rich river fisheries, but pollution and unrestrained commercial exploitation have taken a heavy toll.

**Settlement patterns.** Colombia can be divided into five traditional geographic regions: the Atlantic Lowlands, the Pacific coastal region, the Andean region, the eastern Llanos, and the Amazonian rain forest.

Of early colonial importance, the Atlantic Lowlands is now second to the Andean region in population and economic importance. It contains some 15 percent of the population, much of it concentrated in Barranquilla, Cartagena, and Santa Marta, the country's principal Caribbean ports. Cattle raising and mixed agriculture are the traditional economic activities, but large-scale commercial farming, especially of rice, cotton, and bananas, has been successful. Irrigation has expanded since the mid-20th

century, especially in the valleys of the Sinú and César rivers. Bananas are grown for export in the Urabá region.

The Pacific coast, including the department of Chocó, with its lush rain forest and infertile soils, is sparsely inhabited. Most of its people are descendants of liberated African slaves who settled in agricultural clearings along the rivers. It has little commercial activity, and the only port of note and the main population centre is Buenaventura.

The Andean region is the centre of the national political and economic power, with most of the country's population and large cities, including Bogotá, Medellín, and Cali,

ndean ttlement

the three most populous. The Cauca Valley, with its vast tract of alluvial soil, the Sabana de Bogotá, and the Antioquia highlands are perhaps the most dynamic centres of economic activity and growth.

Although the Llanos and the Amazonian rain forest together make up nearly two-thirds of the country, they contain only about 2 percent of the population. About half of this percentage is in the department of Meta in the Llanos, where cattle raising has long been the traditional way of life. New penetration roads extending down from the Andes have encouraged colonization along the margins of both of these areas, as have discoveries of petroleum. The remoter areas of the Amazon region are sparsely inhabited by small groups of Indians.

## THE PEOPLE

In Colombia much care has been taken to preserve the linguistic purity of the official language, Castilian Spanish, and there are close ties between the Spanish and Colombian language academies. Spanish spoken in Colombia is nevertheless marked by the presence of numerous Colombianisms, many of which have been accepted by both academies. American English terminology has also made a significant incursion. In addition to Spanish there are more than 180 indigenous languages and dialects belonging to such major linguistic groups as Arawakan, Chibchan, Cariban, Tupi-Guaraní, and Yurumanguí. Yet Indian peoples constitute less than 2 percent of the population, a much lower share than in other Andean countries.

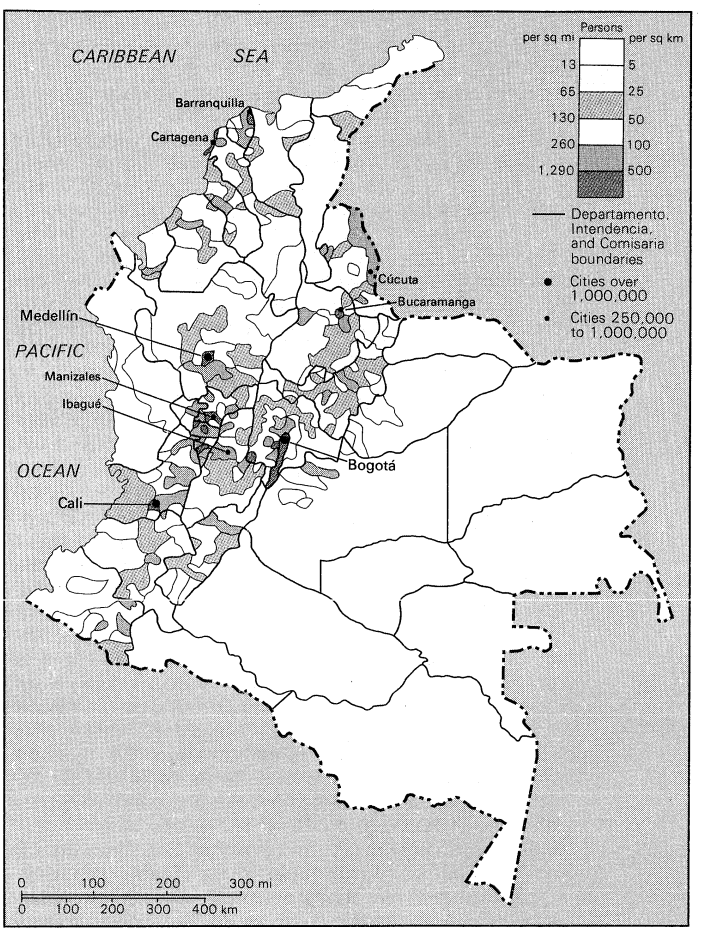
The tri-ethnic character of the population is most pronounced in the coastal departments where the admixture of Africans historically was greatest. Elsewhere mestizos clearly predominate, although official figures are unavailable. The white population has declined, probably to less than 20 percent, mostly of Spanish origin. Unlike most other South American republics, immigration has never been much encouraged in Colombia, although small numbers from the Middle East, non-Iberian Europe, and the Far East have been absorbed into the population.

High post-World War II growth rates peaked in the mid-1960s, declining subsequently to more moderate levels compared to the rest of Latin America, but still high by world standards. The decline appears to have been in part the result of a variety of government programs to reduce fertility, including family planning and educational services. In addition there has been a significant amount of emigration, especially to oil-rich Venezuela and to the United States. This emigration has been a matter of concern to Colombia, both because the loss represents a high proportion of skilled workers and because these often illegal immigrants experience human-rights problems in the

ncreasing nternal nigration

countries to which they move. There continues to be a high rate of internal migration from Colombia's rural areas to its cities, partially determined by the search for better wages and living conditions. The rates of growth in parts of the Llanos and the Caribbean coast, however, grew disproportionately high in the late 20th century, suggesting a general migration from the mountains toward the plains. The rapid growth of the cities has been accompanied by serious unemployment.

Religious freedom is guaranteed by the constitution, although more than 95 percent of Colombians are adherents of Roman Catholicism, the official religion. The church is deeply ingrained in Colombian society, usually taking a leading and authoritative role in the community and having great influence in government. The church has not generally been reform-minded, but some elements of liberalization were evident in the late 20th century. The role of Protestant communities is small, as is that of the thinly spread Jewish community. A few Indian tribes in remote areas still follow their traditional religions.




Population density of Colombia.

## THE ECONOMY

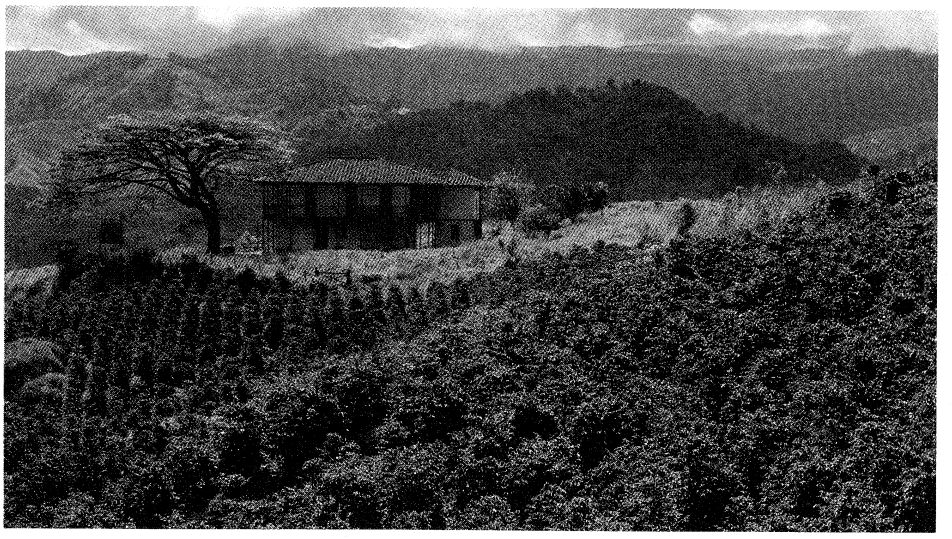
Colombia's economy is dominated by private enterprise, and direct government participation is limited to such enterprises as the ownership of the railways. The government has attempted to foster economic stability and to encourage private enterprise through indirect measures, such as a favourable system of taxation and the extension of credit to new industries. Economic growth was quite substantial through the mid-1900s, but in the later years of the the 20th century inflation and unemployment grew alarmingly as the growth rate declined.

Agriculture is the basis of the Colombian economy, although industrial development since the 1940s has been remarkable. A substantial proportion of the Colombian land is uncultivated because of the prevalence of poor soils and unfavourable climatic conditions. The eastern plains are sparsely inhabited, the Pacific coast is still in forest because of high rainfall, and large areas in the Magdalena Valley remain in open range or are unused.

**Resources.** In the colonial period the economy was based almost entirely on gold mining, including the robbing of the metal from Indian graves (*guacas*). The Gold Museum in Bogotá possesses the world's finest and largest collection of worked gold, the product of extraordinarily skilled craftsmen. Today the economy is much more broadly based with the exploitation of hydrocarbon fuels, several metals, agricultural products, and manufactured goods for export and home consumption. Regional development organizations, such as the Cauca Valley Corporation, have been established to promote more balanced industrial growth, with emphasis on hydroelectric power development and flood control.

Early importance of gold

**Agriculture, forestry, and fishing.** The mountainous



Coffee growing on the middle slopes of the Cordillera Central, near Chinchiná, Caldas department.
Victor Englebert

character of much of Colombia's territory, along with the attendant climatic variations of the different vertical zones, allows for the production of an unusually wide range of both tropical and temperate zone crops, from bananas and sugarcane to wheat, barley, and potatoes. Modern agricultural techniques are employed chiefly in those areas where they are adaptable to the topography. Chemical fertilizers are widely used, and large tracts of flatter lands have been placed under irrigation. Many small farmers, especially in the mountains, nevertheless cling to traditional methods of farming.

Importance of coffee

Coffee represents the backbone of the Colombian economy, constituting about half of all legal exports. The country is second only to Brazil in its production. Coffee, a labour-intensive crop, grows best at an elevation of between 3,300 and 6,300 feet. The farms or estates (*fincas*) on which it is produced are concentrated in the central parts of the three Andean ranges; a few are on the slopes of the Sierra Nevada de Santa Marta. Holdings tend to be small. Colombian coffee, which brings premium prices on the world market, traditionally has been grown under nitrogen-fixing leguminous shade trees, but with the introduction of the high-yielding *caturra* variety, plantings have increasingly been made in the open sunlight.

Bananas and plantains rank as important fruit crops. Most of the bananas grown are exported from plantations located in the Urabá region of the Caribbean coast. Sugarcane is a major crop in the warm and temperate zones, but most of the large plantations and processing plants are located on the alluvial lands of the Cauca Valley near Cali. Some of the sugar is exported, but domestic markets consume the bulk of the production.

Corn (maize), the traditional staple of rural peoples, especially those in the mountains, is grown everywhere except in the *páramos* zones. In some areas it is widely consumed as a form of beer called *chicha*. In the cooler highlands of the Nariño plateau and in the Cordillera Oriental, potatoes are a prominent crop. In the lowlands the production of rice has increased rapidly, most of it grown under irrigation. Cassava (yuca) in the *tierra caliente* and wheat on the *páramos* margins are other major food crops. Kidney beans and sorghum are also widely grown.

Cotton and textiles

Locally grown cotton supplies the requirements of the large Colombian textile industry. Other less significant crops are tobacco, sesame, African oil palms, cacao, peanuts (groundnuts), grapes, soybeans, and citrus fruit. Cut flowers are shipped by air freight in substantial volume from large greenhouses on the Sabana de Bogotá.

Stock raising is a major activity and source of wealth, especially in the lowlands. The Sinú and San Jorge river valleys, the savannas of the Atlantic Lowlands, and the Llanos are the regions with most of the beef cattle. Dairying is especially well developed on the high plateaus of the Cordillera Oriental. Poultry raising has expanded as a result of the application of modern techniques.

The rich resource of the forests has not been fully exploited because access roads to them are few. Where forests are accessible, cutting has been heavy and reforestation programs have been implemented both by the government and by private concerns such as paper manufacturers. The lumber industry is in the process of development, and by the late 20th century there were numerous factories for the manufacture of different types of plywood for domestic use and export.

Ocean fishing is little developed for a country that faces on two great seas. The maritime tradition is of minor proportion. River fish constitute the more abundant catch, although stocks have been reduced as a result of pollution and siltation.

**Industry.** The Instituto de Fomento Industrial (Institute of Industrial Development) has supplied the necessary capital for enterprises too large to be financed by private investment. The institute has invested large sums to strengthen the metalworking industry, to set up automotive assembly plants, to stimulate the construction of railroad cars and fishing vessels, and to encourage the manufacture of paper, vegetable oils, and petroleum derivatives. Despite these developments, the greater part of Colombian industrial activity continues to be carried on by small enterprises that produce consumer goods. An industry that is growing in importance is tourism, which provides much of the country's foreign exchange.

Emphasis on small enterprises

*Mining and quarrying.* Minerals have since colonial times played a key role in the Colombian economy. Colombia's gold production comes largely from dredges operating in the departments of Antioquia and Chocó. In some areas the gold-bearing gravels also contain silver and platinum. The export of ferronickel was initiated in 1985 from a major ore deposit, Mount Matoso in the upper reaches of the San Jorge River. A large copper deposit is located in western Antioquia. The Cordillera Oriental has long been an important source of rock salt, marble, limestone, and especially Colombia's highly prized emeralds, of which the country is the major world producer. Coal is mined in the Andean region for local markets, but production now centres on the great Cerrejón deposits in La Guajira, recently connected by rail to a new port at the extreme end of the peninsula.

High hopes were held for petroleum development in the early 1900s, but the results were disappointing. With the development of two major petroleum fields in the northern Llanos and in Amazonia, the outlook for growth has improved. Pipelines across the Andes linking these fields to ocean terminals have also raised the export potential. The older fields in the Magdalena River valley and in the Catatumbo River region facing Venezuela still produce

important quantities of petroleum. The industry is controlled as a government monopoly, but foreign concerns are partners in exploration and development. The major refinery is at Barrancabermeja on the Magdalena River.

Growth of manufacturing

*Manufacturing.* Manufacturing has been slowly increasing its value to the economy, having reached about 20 percent of the gross domestic product by the mid-20th century. The textile industry employs the largest share of workers and contributes a substantial part of the national income. In addition to supplying the national market, the larger concerns also export fabric and yarn. They are concentrated in Medellín. Food processing and chemical production rank with textiles as leading Colombian industries. Industrial chemicals, produced in part to supply the textile industry, have increased steadily in production, and there is also an important output of pharmaceuticals. The integrated iron and steel mill at Paz de Río, in Boyacá department, utilizes local raw materials and supplies a large share of the country's metal needs. Bogotá, Medellín, and Cali, along with the Caribbean coastal cities of Barranquilla and Cartagena, are the principal industrial centres. The interior location of the first three has placed them at a significant disadvantage in both processing of imported materials and producing for export, but the demands of the growing domestic market, coupled with substantial investments by foreign concerns in productive facilities, has enabled them to sustain substantial growth, especially since World War II. Cheap electric power distributed through a national grid has been an important developmental factor.

**Finance and trade.** The banking system is composed of a central bank, the Banco de la República, and more than 30 general banking institutions, some of which are partly foreign-owned. The Monetary Commission, created by the government in 1963, is the highest authority in matters involving the extension of credit. Such credit is extended through the central bank, which also issues currency, acts as banker for the government and other banks, serves as a guardian and administrator of the country's international reserves, and acts as a clearinghouse.

Foreign trade is concerned principally with the exportation of raw materials and the importation of machinery and manufactured goods. Colombia's single largest trading partner is the United States. Trade with the countries of the European Economic Community is significant, as is trade with neighbouring Andean countries.

Trading partners

Exports consist largely of coffee, crude petroleum and petroleum products, bananas, fresh-cut flowers, chemicals, and cotton. By the 1970s and '80s the illegal trade in Colombian marijuana and cocaine, especially with the United States, had become a major source of income, probably rivaling the value of legal exports. Imports consist mainly of machinery and transportation equipment, chemical products, crude petroleum and petroleum products, base metals and metal products, and paper and paper products.

**Transportation.** Transportation plays a particularly vital role in Colombia, where the problems of a diverse and difficult terrain are being overcome to unify the country. By far the most important means of surface transportation is the road system, about one-fourth of which is paved. Two parallel trunk roads extend toward the interior from the Caribbean ports, one following the Cordillera Oriental to Bogotá and Santa Marta, the other passing through Medellín, Cali, and Popayán to the Ecuadorian border. A branch from the first leads to Cúcuta and into Venezuela. There is, however, no overland communication with Panama and Central America, the difficult terrain of the Darién Gap, which separates Panama and Colombia, being the only separation in the Pan-American Highway. Road extension and improvement is a priority of the government, for most domestic cargo today moves by truck. Frequent landslides make highway maintenance difficult. One of the most important transverse routes is that linking Bogotá with the Cauca Valley and Buenaventura, the major Pacific port, through the Cordillera Central.

Role of air transport

Perhaps in no other country has air transport played so major a role as in Colombia. The government-controlled Avianca claims to be the oldest commercial airline operating in the Western Hemisphere. Frequent flights link all important cities, reducing travel times inordinately from those on the tortuous, indirect, and slow mountain highways. Most people travel by air in Colombia, which is claimed to have proportionally the highest rate of air travel in the world. The principal international airport is Bogotá's El Dorado, and there are others at Medellín, Cali, Cartagena, Barranquilla, and Isla San Andrés. The last serves the large tourist industry there.

The role of railroads has become increasingly secondary. The standard-gauge lines are owned by the government. The main line is that of the Ferrocarril del Atlántico, which runs north for 600 miles between Bogotá and the seaport of Santa Marta. At Puerto Berrío in the Magdalena Valley the main line connects with another that passes westward through Medellín and on southward to Cali and the port of Buenaventura. This and other regional lines are frequently closed by landslides.

The Magdalena River no longer plays the vital role in transportation that it once did although it still carries some bulk cargo, especially petroleum. Travelers en route to Bogotá in earlier times moved by river boat as far as La Dorada, where the trip to the interior capital continued overland. The Sinú, Atrato, and Meta rivers are also navigable, but these, too, are less used than formerly. Consideration has been given to the possibility of uniting the Caribbean with the Pacific by the construction of a canal between the Atrato and San Juan rivers.

Cargo ships ply the waters of both the Caribbean and the Pacific, which are joined to the north by the Panama Canal. The Caribbean ports of Cartagena, Barranquilla, and Santa Marta have relatively deep water and are equipped with modern port facilities and services. Silt deposited by the Magdalena River at its mouth requires constant dredging to maintain shipping access to the Barranquilla wharves. On the Pacific coast the port of Buenaventura, on a mangrove-lined embayment, is of easy access with modern installations. Tumaco to the south has been marked for development.

## GOVERNMENT AND SOCIAL CONDITIONS

**Government.** Under the constitution of 1886 Colombia is a republic, the public powers of which are divided between the executive, legislative, and judicial branches of government. The president, who may not succeed himself, is elected to a four-year term by universal suffrage. The executive is assisted by a ministerial Cabinet. The bicameral legislature is composed of a senate and a house of representatives, both of whose members are elected by universal suffrage to four-year terms.

Administrative structure

The country is divided for administrative purposes into 23 departments, four intendencies, five commissariats, and one special district, of Bogotá. The departments are headed by governors appointed by the president, and each has an elected legislature. The departments are subdivided into municipalities, which are headed by mayors appointed by the governors. The intendencies of Arauca, Casanare, Putumayo, and San Andrés y Providencia—as well as the commissariats of Amazonas, Guainía, Guaviare, Vaupés, and Vichada—are governed directly by representatives of the central government.

The Colombian political process had its origins during the formation of the republic. For more than 100 years, the two largest political parties—the Liberals and the Conservatives—maintained a struggle for power. In 1957, in order to overthrow the dictatorship of Gustavo Rojas Pinilla, the parties agreed upon a truce that continued for 16 years (see below *History*). In 1974 the country returned to its traditional system of competition between the two leading political parties.

Suffrage is extended to all citizens over the age of 18. Citizens are guaranteed civil rights, including the right to strike, to assemble, and to petition; freedom of the press is also guaranteed. All male citizens between the ages of 18 and 30 may be called for military service.

**Education.** The educational system includes kindergartens (preschool facilities), primary schools, and secondary schools and other educational facilities that offer training in industry, domestic science, veterinary science, business, nursing, theology, and art. The majority of the

country's universities are located in the capital city. Public institutions of higher learning include the National University of Colombia, the Francisco José de Caldas District University, the Central University Foundation, and the Women's National Pedagogical University—all of which are in Bogotá—as well as universities in such other major cities as Medellín, Barranquilla, Cartagena, Popayán, and Cali. Private universities in Bogotá include the University Foundation of Bogotá, the Xavieran Pontifical University, and the University of the Andes.

**Welfare and health.** Welfare services date to the 1930s. Social security programs include health and maternity benefits, workers' compensation, and allowances for those unable to work. As in most Latin-American countries housing is in short supply, a problem that is especially serious in large cities, which attract a large migrant class that settles in slums. The Housing Institute deals with the problem, directing the construction of housing for the low-income rural and urban population.

Public health advancements

The Ministry of Public Health seeks to arouse the interest of individual communities in seeking solutions to health problems through independent efforts. Projects include the construction of systems to supply drinking water; public education in the matters of basic sanitation, home maintenance, balanced diet, and personal cleanliness; and the control of industries and organizations whose operations might be hazardous to health. Malaria and dysentery are common health problems in the rural areas, particularly in the poorly drained lowlands. Hookworm is troublesome in the damp environments of the shaded *cafetales,* or coffee plantations. Yellow fever, once of serious concern in the port cities, has been eradicated. Although health conditions have improved, serious problems still exist, especially among the poor and in remote areas. Many health problems are caused by malnutrition.

## CULTURAL LIFE

Geography has caused local isolation to be an important factor in Colombian life, and cultural particularism is highly developed. People are often known by the department in which they live, and Antioqueños, Santandereanos, Tolimenses, Nariñenses, Bogotanos, and Boyacanses are recognized by their dress, diet, and speech. The most important social group is that of the Antioqueños, who migrated from Antioquia southward along the Cordilleras Central and Occidental during the 19th century. Numbering almost 5,000,000, the Antioqueños grow about three-fourths of the nation's coffee crop and control much of Colombia's trade, banking, and industry.

**Cultural origins.** Prior to the arrival of the first Europeans in the 16th century the aboriginal populations of the area that was to become Colombia had achieved impressive levels of cultural development. Because they built largely of wood and occupied a tropical area of generally moderate to high rainfall, they left little evidence of their achievements. All groups had some form of social organization, but except for the Chibcha of the Cordillera Oriental, they were organized in small chiefdoms (*cacigazcos*) under chiefs (*caciques*), whose authority was sharply limited geographically. Agriculture, pottery making, and weaving were all but universal. Some tribes, like the Chibcha, Quimbaya, Tairona, Sinú, and Calima had impressive skills in metalworking (especially goldsmithing), sculpture, and ceramics. The San Agustín culture, centred in the headwaters area of the Magdalena River, left giant anthropomorphic figures carved of stone that have been an enigma for archaeologists. While groups of Caribbean origin were warlike and practiced ritual cannibalism, others from the interior possessed a rich mythology and a religion that upheld ethical standards and norms on questions of private ownership and the prevention of crime.

Pre-Colombian cultures

Until the mid-1970s it was thought that no indigenous group had left any large architectural monuments—such as those erected by the Aztecs, Mayas, or Incas. The excavation, beginning in 1976, of a 1,500-acre city apparently built about AD 900 by the Tairona in the Santa Marta massif, however, marked a turning point in the study of Colombia's prehistory.

The Andean Indians, particularly the Chibcha, practiced sedentary agriculture and were able to offer but small resistance to the Spanish invaders. Instead, they became the great biological and cultural contributors to the process of racial amalgamation, or *mestizaje.* The low demographic density of the pre-Hispanic population, its swift destruction, and the relatively limited African immigration to Colombia led to the formation of a rather open society and to the substitution of Hispanic forms of culture for the indigenous ones. Since the 17th century the most widely used native language, Chibcha, has virtually disappeared.

Since colonial times, Bogotá—the Athens of South America—has been the nation's cultural centre, and most cultural institutions are located within the metropolitan area. Other cities of cultural importance include Cali, Medellín, Manizales, and Tunja.

**The arts.** The arts in Colombia are fostered and developed by conservatories and schools that function in several cities either in connection with the universities or independently and by the growing number of concert halls and galleries. Persons of middle income levels display considerable curiosity and the desire to be informed about contemporary artistic developments, and this same eclectic spirit is found among the artists themselves. There is no distinct national school of art.

Literature

The Nobel Prize for Literature awarded to Gabriel García Márquez in 1982 provided recognition of a national literary tradition that Colombians believe constitutes a basic element of the national character, as they boast that more poets than soldiers have occupied the presidency. Although historical themes dominated the earliest writing, the novel and poetry appeared in early colonial times. The European Enlightenment brought essays and the first scientific surveys in the 18th century; political and imaginative themes characterized the 19th century. Although poetry and the novel have remained strong in the 20th century, the focus has often been regional rather than national, personal as much as social.

Handicrafts suffered a decline during the colonial period and into the early years of the republic but since the early 1930s have experienced a revival, especially in the production of textiles. There was also a revival in the manufacture of ceramics and pottery, chiefly in the municipalities of Ráquira, El Espinal, and Malambo. Basket weaving, harness making, and passementerie (fancy edging or trimming on clothing or upholstery) are also popular.

Popular traditions concerning manners and customs, music, legends, and food preparation continue in somewhat attenuated form in their places of origin. Perhaps the most deeply rooted folkloric form of expression is that of music. The tunes and melodies of the indigenous tribes are sung only in limited geographical areas. The music of the mestizo can be divided into that of the Andes, the plains, and the Atlantic Lowlands and the Pacific coast. Some music forms of the colonial period also have survived.

**Cultural institutions.** The history and culture of Colombia's indigenous peoples are revealed in several museums of outstanding reputation. The Gold Museum of Bogotá contains a famous collection of goldwork, while the Bogotá Museum of Colonial Art has a rich collection of criollo (*i.e.,* by Spanish persons born in Colombia) religious sculpture and painting. The National Museum displays treasures and relics dating from prehistoric times to the present and possesses various collections of Colombian painting and sculpture. The July 20 Museum contains precious documents of the period of independence.

Major museums

No less important vehicles for the diffusion of culture are the National Library and the Bank of the Republic Library, which contains a vast amount of reading material, exposition and music halls, and a concert theatre. Outside Bogotá there are other institutions of this kind, including the Zea Museum in Medellín, and the House of Don Juan de Vargas, in Tunja.

**Recreation.** Since the 1960s regional fairs have been held in various parts of the country to celebrate occurrences of local importance. They are government-subsidized and, with the aid of modern means of communication, have promoted and preserved popular tunes and dances as well as traditional costumes. Fiestas in Colombia vary locally, but the pre-Lenten carnival is especially celebrated nation-

ally, reaching a particular intensity along the Caribbean coast, especially in Barranquilla.

Organized sports have grown steadily in popularity among the Colombians. Without question the most widely played and watched sport is football (soccer). Basketball and baseball draw an increasing number of fans, and golf, tennis, and skiing are enjoyed by the smaller numbers who can afford them. Automobile and bicycle racing are also attractions. Perhaps the only indigenous sport is *tejo,* a game similar to quoits, which was derived from the Chibcha Indians. Colombians seem peculiarly to enjoy gambling, including government-sponsored lotteries, which fund social programs. Like many other Latin-American peoples, Colombians enjoy bullfights, an inheritance of their Spanish culture.

Sports

**Press and broadcasting.** Although freedom of the press has generally been established in Colombia, the degree to which the press can exercise its rights has been somewhat dependent upon the government in power. Newspapers have traditionally been the most widely available source of political information and have been the least controlled, while radio and television, regarded more as entertainment media, have received stricter government control. Newspapers have often been the voices of particular political parties; two noted Bogotá newspapers, *El Tiempo* and *El Espectador,* for instance, have usually been identified with the Liberal Party philosophy.

For statistical data on the land and people of Colombia, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL. (C.Ga./Ja.J.P.)

## History

### PRECONQUEST

Even before the Spanish conquest, the western mountainous part of Colombia attracted the bulk of the population. The higher Indian cultures were found in this region, and the most favourable location for the growth of civilization was the high plateau in the Cordillera Oriental of the Colombian Andes. The present capital city of Bogotá is located near the southern terminus of the plateau, which extends northward to the mountains dividing it from the drainage of the César River. There the Spanish found the major concentration of the Chibchan-speaking peoples.

Pre-Columbian civilization

At the time of the Spanish conquest the Chibcha were in the process of consolidation by warfare. They had not achieved firm union and political institutions, and their other cultural traits had not attained the solidity and refinement needed to resist cultural conquest.

Except for the invading Carib peoples in the deep mountain valleys, there was a considerable similarity among the Chibcha, sub-Andean, and circum-Caribbean cultures of Colombia. All were characterized by intensive agriculture, fairly dense populations living in villages, organized religion, class divisions, and matrilineal inheritance of political and religious offices. The sub-Andean culture in the Cordillera Central and the narrower portions of the Cauca Valley generally lacked the feature of large villages because of the unsuitability of the terrain. The more advanced Chibcha made war for political ends, using large forces armed with the dart and dart thrower.

Geographic and climatic conditions placed limits to the development of the Chibcha and other cultures in Colombia. Of the total Indian population at the time of the conquest, probably about one-third were Chibcha. None of the larger domesticated animals and their wild related species found in the central Andes existed in Colombia. The Chibcha were adequate craftsmen, but their work shows more interest in utility or in the expression of ideas than in the attainment of the skilled workmanship striven for among some sub-Andean peoples.

### CONQUEST

European exploration of the Colombian coastline was the work of Rodrigo de Bastidas, who in 1500–01 sailed the Caribbean coast from Cabo de la Vela to Nombre de Dios in Panama, and of Francisco Pizarro, who sailed the Pacific coast in 1525. Effective conquest of Colombia began in 1525 with the founding of Santa Marta on the north coast by Bastidas. In 1533 Pedro de Heredia founded Cartagena, which became one of the major naval and merchant marine bases of the Spanish empire. By the end of 1539 all but one of the major inland colonial cities had been founded, as well as the most important communications centres along the routes connecting them. The capital city of Santa Fé de Bogotá was founded by Gonzalo Jiménez de Quesada in 1538. By mid-century the era of the conqueror drew to an end.

### COLONIAL PERIOD

**The audiencia.** Establishment of the audiencia (an administrative and judicial tribunal) of Santa Fé de Bogotá in 1549 opened the colonial era. The conquerors had organized local governments in accordance with the terms of their contracts with the crown. The crown then rapidly repossessed the broad powers granted the conquerors and formed its own institutions to rule the empire. The governments of Popayán, Antioquia, Cartagena, Santa Marta, Ríohacha, the New Kingdom of Granada (Bogotá), and the llanos of Casanare and San Martín were made subject to the new audiencia. The president of the audiencia was the executive head of government, subject to the viceroy of Peru in administrative matters. The difficulties of travel, however, impeded communications and checked centralized control. The area declined in population after the conquest as a result of disease and the economic demands made upon the Indians.

Dominance of the crown

As elsewhere in the empire, the downward trend seems to have reversed itself at the end of the 17th and the beginning of the 18th century. Acculturation and intermarriage rapidly destroyed most of the special cultural traits of many remaining Indians. Subordinate political jurisdictions developed strong regional characteristics as a result of isolation, which fostered intense local loyalties and rivalries. The economy was based on mining and agriculture, but a small yet important textile industry had grown up in Socorro, north of Bogotá, by the mid-18th century. Slavery was introduced during the conquest and became common in the placer-mining areas of the Chocó and western Antioquia and in the agricultural regions of the Cauca Valley, the lower Magdalena Valley, and the coastal lowlands. Indians were subject to the encomienda tribute, but by 1700 most of the privately held encomiendas had reverted to the crown, and they were rarely granted thereafter. During the era of the audiencia, from 1549 to 1740, the population was politically quiet. The Roman Catholic Church played an important role, providing most welfare services and operating most schools. The church was an effective instrument of the crown, since the latter controlled much of its activity.

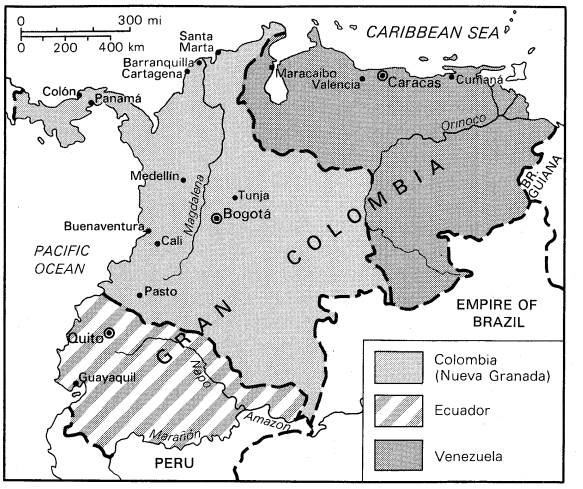
Role of the Roman Catholic Church

**Viceroyalty of New Granada.** Creation of the Viceroyalty of New Granada, which included present-day Colombia, Venezuela, and Ecuador, temporarily in 1717–23 and permanently in 1740, opened a new era. In the next decades the crown introduced political and economic measures to reorganize and strengthen the empire by greater centralization of authority, improved administration and communication, and freer development and movement of trade within the empire. Population grew, trade increased, and prosperity touched the colonial subjects. There was a spurt of intellectual activity and the formation of a corps of Creole intellectuals and professional men, many in government positions. The small Creole officer corps came into being when Charles III, then king of Spain, authorized militia defense units in the colonies. A relatively large group of wealthy landowners and merchants constituted the economic community that supported these new groups. Between 1785 and 1810 in New Granada the outlook of the Creole upper and middle groups changed from resistance toward political and economic change introduced by government to a quest for specific changes in imperial policies. Thus in the Comunero Rebellion of 1780–81 the Socorrans opposed change, while in 1809 they proposed new policies leading to the free-enterprise system, the abolition of slavery, restrictions on government, and worldwide freedom of trade.

Change in Creole politics

Educational reforms played an important role in the changing outlook of the Granadines. Archbishop Caballero

y Góngora as viceroy (1782–88) made education one of his main interests. He modernized the program of studies in the schools, opened a school of mines, and initiated the botanical expedition under the able guidance of naturalist José Celestino Mutis. The new institute trained many of the major figures of the independence movement. The 1790s were the decade in which the first newspaper and the theatre were introduced. A new interest in writing developed and intellectual gatherings for discussion were introduced. In 1808 the allegiance of the Granadines to the crown remained unquestioned except for a few individuals. The once warm loyalty of the Creole middle and upper classes, however, was cooling under the pressure of economic interests, scandals in the royal family, and persistent social tension between Creole and European Spaniards.




The division of Gran Colombia (1830).

Adapted from D. Worcester and W. Schaeffer; *The Growth and Culture of Latin America*, vol. 1, 2nd ed.; Oxford University Press, Inc.; New York (1970)

## REVOLUTION AND INDEPENDENCE

The French invasion of Spain in 1808 caused an outburst of loyalty to king and country and excited grave concern for the church. Profound Granadine anxiety over the fate of the empire and conflicting courses of action attempted by colonial and peninsular subjects over control of government during the captivity of the Spanish king Ferdinand VII led to strife in New Granada and to declarations of independence. In 1810 the subordinated jurisdictions in New Granada threw out their Spanish officials, except in Santa Marta, Ríohacha, modern Panama, and present-day Ecuador. The uprising in Bogotá on July 20, 1810, is commemorated as independence day in Colombia. These new governments swore allegiance to Ferdinand VII and did not begin to declare independence until 1811. Idealists and ambitious provincial leaders desired federation. Creole leaders sought to centralize authority over the new governments. A series of civil wars ensued, facilitating Spanish reconquest of the United Provinces of New Granada between 1814 and 1816. A remnant of republican forces fled to the llanos of Casanare, where they reorganized under Francisco de Paula Santander, a Colombian general who remained a prominent figure in Granadine politics until his death in 1840.

Bogotá uprising

Any remaining loyalty to the crown was alienated by the punitive arbitrary conduct of the European and partisan troops. Their conduct gave validity to the attack on Spanish civilization that began late in 1810 and continued through the 19th century. The forces in Casanare joined those of Simón Bolívar in the Orinoco Valley of Venezuela. By 1819 arrangements for a regular government were completed, and a constitutional convention met at Angostura (now Ciudad Bolívar, Venezuela) with delegates from Casanare and some Venezuelan provinces. In that same year Bolívar invaded New Granada and decisively defeated the Spanish forces on August 7 at Boyacá. There followed the decisive Battle of Carabobo, Venezuela, in 1821 and that of Pichincha, Ecuador, in 1822. Mopping-up operations were completed in 1823, while Bolívar led his forces on to Peru.

Congress of Angostura

The Congress of Angostura laid the foundation for the formation of the Republic of Colombia (1819–30), which was generally known as Gran Colombia because it included what are now the separate countries of Colombia, Venezuela, and Ecuador. The republic was definitively organized by the Congress of Cúcuta in 1821. Prior to that time the government was highly military and strongly centralized with direct executive power exercised by regional vice presidents while President Bolívar was campaigning. Organized as a centralized representative government, the republic had Bolívar as president and acting president Santander as vice president.

Gran Colombia had a brief, virile existence during the war. Subsequent civilian and military rivalry for public office and regional jealousies led in 1826 to a rebellion in Venezuela led by Gen. José Antonio Páez. President Bolívar returned from Peru to restore unity but secured only the acknowledgment of his personal authority. As discontent spread it became clear that no group loved the republic enough to fight for its existence. By 1829 Bolívar had divided the land into four jurisdictions under Venezuelan generals possessing civil and military authority. Meanwhile the convention of Ocaña had failed to reorganize the republic, and the brief dictatorship of Bolívar (1828–30) had no better success. Bolívar then convoked the Convention of 1830, which produced a constitution honoured only in Nueva Granada (Colombia). During this convention Bolívar resigned and left for the northern coast, where he died near Santa Marta on Dec. 17, 1830. By that time Venezuela and Ecuador had seceded from Gran Colombia. Nueva Granada, a country of 1,500,000 inhabitants in 1835, was left on its own (Colombia continued to exercise control over Panama until 1903).

(R.L.Ge./Ed.)

## THE REPUBLIC TO 1930

Santander, the vice president under Bolívar and then leader of the opposition to Bolívar's imperial ambitions in 1828, held the presidency from 1832 until 1837 and was the dominant political figure of that era. The 1830s brought some prosperity to the new nation, but a civil war that broke out in 1840 ended a nascent industrial development, disrupted trade, and discouraged local enterprise. The seeds of political rivalry between liberals and conservatives had already been sown, and they bore fruit in the bloody revolution and costly violence that ravaged the country in the years between 1840 and 1903.

Revolution and violence, 1840–1903

**Conservative–Liberal struggle, 1840–80.** Colombia's modern political history began in the late 1840s with the delineation of the Liberal and Conservative parties. Gen. Tomás Cipriano de Mosquera, a Conservative, during his first term as president (1845–49), replaced the government monopoly on tobacco sales with a private monopoly and expanded international trade. These changes increased the production and export of tobacco but reduced the tax income of the national government.

In 1849 Gen. José Hilario López, of the radical faction of the Liberal Party, became president. It later became his task to implement the reforms passed in 1850, which were to galvanize political sentiment and divide the country politically and economically for half a century. The guiding principle of the radical Liberals under General López was greater liberty for the people of Colombia. His government ended slavery, ended communal ownership of Indian lands, diverted tax resources from the central to local governments, and eliminated a number of taxes and monopolies held by the central government.

Reforms of 1850

Rather than eliminating the institutional barriers to self-fulfillment by the people, however, the reforms of 1850 tended to eliminate the traditional proscriptions that had stood as safeguards against the exploitation of the poor by the rich. The reforms, despite the liberal rhetoric that accompanied them, legalized, indeed encouraged, a redistribution of landed property and tended to strengthen the position of the wealthy landowners, merchants, and professionals against the mass of poor Indians, peasants, and artisans. Since there were only 25,000 slaves (in a

country of 2,000,000 in 1851), the effects of manumission were small compared to those of the breakdown of the Indian communal system, which affected a third of the population. The Indians were induced to give up their little plots of land and the small amount of independence they enjoyed. Within a few years the ownership of Indian lands was concentrated in a few hands; the Indian had become a tenant, his land used for grazing cattle.

But while class conflict seethed under the surface in Colombian society, the struggle between members and groups within the elite was more open. Two issues in particular divided the upper class: first, whether a centralist or federalist political system would be the best arrangement for Colombia, and second, what role was appropriate for the Roman Catholic Church and particularly for its clerics in Colombian society. Adherents of federalism were strongest in the years between 1863 and 1880, during which years the country was called the United States of Colombia. Subsequent government publications were to refer to that period as the "Epoch of Civil Wars." In 51 of the 240 months that passed in the 1860s and '70s, there was some form of civil conflict taking place within the country. The Colombian army was so small that public order could not be maintained.

The "Epoch of Civil Wars"

The power of the anticlerical faction reached a peak in the early 1860s. A revolutionary government headed by Mosquera expropriated church lands in 1861, and a constitution adopted in 1863 guaranteed freedom of religious practice, thus bringing to an end the traditional intimate relationship between church and state in Colombia.

**The return of the Conservatives, 1880–1930.** Both actions were reversed during the period of Regeneration (1880–95) under Rafael Núñez and the Conservatives who followed him. After further civil conflict in the 1880s Núñez was able to promulgate a new constitution in 1886, to reestablish relations with the Vatican via the Concordat of 1887, and to promote some internal improvements and industrial development. But the political struggle between Liberals and Conservatives was far from over. Armed conflict reached its peak in the War of a Thousand Days (1899–1903). The estimates of the number of deaths in that struggle range from 60,000 to 130,000.

The tragedy of civil war was followed by the loss of Panama. The Colombian Congress vacillated too long in considering a United States offer to build a canal across the isthmus, and in 1903 the Panamanians revolted against the government in Bogotá and arranged for a treaty agreement providing for U.S. sovereignty in the 10-mile-wide Canal Zone in exchange for an agreement by the United States to build the canal and to provide a regular annual payment to Panama. Although the U.S. government later agreed to pay $25,000,000 to Colombia, the episode embittered Colombian-U.S. relations for many years.

Colombia's internal development quickened after 1905, with coffee exports expanding by nearly 10 percent per year between 1909 and 1928. At the beginning of the 20th century Colombia supplied about 3 percent of world coffee exports; by 1923 its share had risen to nearly 10 percent. In the late 1920s coffee accounted for some 18 percent of the gross domestic product.

Colombia's internal development

## COLOMBIA SINCE 1930

The new dependence on exports was not without its pitfalls. In the late 1920s coffee, petroleum, and bananas accounted, respectively, for 69, 17, and 6 percent of total Colombian exports, and all three dropped precipitously in value during the Great Depression. This economic collapse had an immediate political result; the Conservatives were ousted from the presidency by Enrique Olaya Herrera (who served 1930–34).

**The era of the Liberals, 1930–46.** During the presidency of Alfonso López, from 1934 to 1938, a series of reforms, called the "Revolution on the March," was instituted. The most important social act of the López regime established effective occupancy as the legal basis for tenure (1936), thus upholding the rights of thousands of peasant squatters against the claims of landowners who had been holding land without using it productively. In the coffee-growing zone of Cundinamarca, west of Bogotá, thousands of families obtained recognition of their ownership by occupation. Subsequent governments took a more conservative stance toward the question of land rights of the poor, but in 1961 continuing social pressure finally resulted in legislation to create the Colombian Institute of Agrarian Reform (Instituto Colombiano de Reforma Agraria). By the mid-1970s, more than 135,000 land titles had been distributed by the institute.

Rapid industrial development started in the 1930s. Medellín became the principal producer of cotton textiles and other fabrics. The limited availability of imports because of the Depression enabled local manufacturing to get its start. The tendency toward substitution of home products for imports continued into the 1950s and 1960s, when Colombia became practically self-sufficient in production of consumer nondurables. By the early 1980s, however, manufacturing still accounted for the same one-fifth of the gross domestic product that it had in the early 1960s.

**Civil unrest, dictatorship, and democratic restoration, 1946–70.** Liberal hegemony continued through the 1930s and the World War II era. In the elections of 1946, however, two Liberal candidates, Gabriel Turbay and Jorge Eliécer Gaitán, stood for election and thus split the Liberal vote. A Conservative, Mariano Ospina Pérez, took office, and Conservatives instituted crude reprisals against Liberals. On April 9, 1948, Gaitán, leader of the left wing of the Liberal Party, was assassinated in broad daylight in downtown Bogotá. The resulting riot and property damage (estimated at $570,000,000 for the country as a whole) has come to be called the *bogotazo.*

Conservative resurgence

There is disagreement on whether the acts of violence began in 1930 when the Liberals came to power, in 1946 with the Conservatives, or in 1948 with the death of Gaitán. There is agreement that they sprang from a political feud between Liberals and Conservatives that had little to do with class conflict, foreign ideologies, or other matters extraneous to the Colombian scene. Authoritative sources estimate that some 200,000 persons lost their lives in the period known as La Violencia, between 1948 and 1962. The most spectacular aspect of the violence, however, was the extremes of cruelty perpetrated on the victims. The aggressive force unleashed by that political conflict has been a continuing problem of study for Colombians.

La Violencia

La Violencia intensified under the regime of Laureano Gómez (1950–53), who attempted to introduce a fascist state. His excesses brought his downfall by military coup—Colombia's first in the 20th century. Gen. Gustavo Rojas Pinilla assumed the presidency in 1953 and, aided by his daughter, María Eugenia Rojas, began an effort to end La Violencia and to stimulate the economy. Rojas was a populist leader who appealed to the masses and supported their call for the redress of grievances against the elite. Support for Rojas began to collapse when it appeared that he would not be able to fulfill his promises and when the economy faltered as a result of a disastrous fall in coffee prices in 1957. Rojas was driven from office in 1957 by a military junta.

The arrangement for the National Front (Frente Nacional) government—a coalition of Conservatives and Liberals—was made by Alberto Lleras Camargo, representing the Liberals, and Laureano Gómez, leader of the Conservative Party, in the Declaration of Sitges (1957). The unique agreement provided for alternation of Conservatives and Liberals in the presidency, an equitable sharing of ministerial and other governmental posts, and equal representation on all executive and legislative bodies. The agreement was to remain in force for 16 years—equivalent to four presidential terms, two each for Conservatives and Liberals. The question of what governmental structure would follow the National Front was left unsettled.

Terms of the National Front

It had been contemplated that a Conservative would be the first to occupy the presidency in 1958. When the Conservative Party could not agree on a candidate, however, the National Front selected Lleras Camargo, who had previously served in that office for 12 months in 1945–46. During Lleras Camargo's tenure an agrarian reform law was brought into effect, national economic planning for development got its start, and Colombia became the showcase of the Alliance for Progress (a U.S. attempt to further

economic development in Latin America). But severe economic difficulties caused by low coffee prices, domestic unemployment, and the apparent end of the possibilities of substituting industrialization for imports were only partially offset by Alliance aid. The Alliance increased Colombia's economic dependence on the United States, which, to some Colombians, had serious disadvantages. By 1962 economic growth had come almost to a standstill.

The precarious state of the economy and the degree of social tension were revealed when only about half of those eligible to vote did so in the 1962 presidential elections, which brought Guillermo León Valencia, a Conservative, to the presidency. During Valencia's first year in office internal political pressures led to devaluation of the peso, wage increases among unionized workers of some 40 percent, and the most rampant inflation since 1905. Extreme deflationary policies were applied in the next three years, raising the unemployment rates above 10 percent in the major cities and turning even more Colombians against the National Front. Less than 40 percent of the voters went to the polls in the 1964 congressional elections.

Government of Lleras Restrepo

Carlos Lleras Restrepo was the third National Front president (1966–70). He returned the economy to a sound footing, improved government planning for economic development, and pushed through political reforms essential to an orderly end to the National Front (which seemed increasingly to constitute a monopoly of power by the Conservative–Liberal oligarchy). Semiautonomous government corporations expanded their services to the private sector: the capital and reserves of the Institute of Industrial Development (Instituto de Fomento Industrial), for example, were increased from 6,600,000 pesos in 1967 to some 77,000,000 pesos in 1969. Colombia achieved its best rate of economic growth near the end of the Lleras administration, when the real gross domestic product increased by some 6.9 percent. These successes were in part due to high coffee prices, but effective government policy was of undeniable importance.

In the 1970 presidential election Misael Pastrana Borrero, the Conservative candidate backed by the National Front, nearly lost to former dictator Gustavo Rojas Pinilla because the urban vote went strongly against the Front. (For the first time Colombia's population was more than 50 percent urban. A rapid migration from country to city created new urban interest groups—particularly in the lower middle and working classes—that felt unrepresented by the traditional parties.) Nonetheless, the traditional parties prevailed and were not again successfully challenged.

**The Conservative–Liberal rule.** The process of change brought with it new political, economic, and social problems. In all three areas of national concern, the problems stemmed from uneven development, unequal gains, and a growing perception that the benefits of higher income on average were not widely shared. The transition from National Front to moderate political competition between Liberals and Conservatives in 1974 was reasonably smooth. Alfonso López Michelsen of the Liberal Party served his four-year term as president (1974–78) and handed power to Julio César Turbay Ayala, a centrist Liberal. Low rates of voter participation continued, keeping alive fears that military alternatives to democratic elections might be sought from the right or the left. In 1982, however, the Liberal vote was split, and Belisario Betancur Cuartas, the Conservative candidate, was elected president.

Return to political competition

The presidency of Betancur was marred by extremes of violence that tested Colombia's long-term commitment to democracy. Terrorists kidnapped and held captive a number of foreign ambassadors, and, in 1984, elements believed to be linked to the international drug trade assassinated the Minister of Justice. In 1985, guerrillas entered the Palace of Justice in Bogotá and held a number of captives before being expelled by the military with a loss of life estimated at about 100, including a number of Supreme Court judges. These events pointed to an ominous growth in the power of drug traffickers and to an inability of the government to control terrorist activities.

The presidency of Virgilio Barco Vargas, a former mayor of Bogotá, began in August 1986 with hopes for improvements in civil order. His main economic problem lay in how to reverse the long-term decline in the rate of economic growth. Gross domestic product had grown fastest in the years 1967–74 but had tapered off due to the jumps in oil prices in the 1970s and was declining into the 1980s. The petroleum discovery in 1985 had the potential of improving the economy and reducing Colombia's dependence on external energy sources.

After another severe Brazilian frost in 1975 destroyed that nation's coffee crop, Colombian exporters earned the highest prices ever for their coffee, and a new and endemic inflation, at rates above 20 percent per annum, occurred during three of the following four years. The real per capita national product grew between 3 and 4 percent annually, and the shift from an agricultural to an industrial economy continued. Manufactured exports in 1976 were almost 23 times greater in value than those of 1963, much of that growth attributable to opportunities for trade within the Andean Pact trade group of Bolivia, Chile, Colombia, Ecuador, and Peru. Colombia continued to receive foreign economic assistance, but the economy was insufficiently vigorous to draw a significant increase in private investment.

Demographic changes

The sum of the decades of change after the disturbances of La Violencia was a more modern, urbanized, and educated populace than Colombia had ever experienced. The remaining problem of dualism—the perpetuation of absolute poverty for a quarter of the population in the midst of economic growth—was brought into relief by the title of the government's 1975–78 development plan, *Para cerrar la brecha* ("To Close the Gap"). An elaborate plan for food production and child nutrition was also addressed to that problem. Many, however, were left behind in the process of economic and social change, although studies in the major cities showed that the number of families living in poverty had declined from 40 percent in the mid-1970s to about 20 percent by 1980 as a result of government programs, fertility decline, and increased participation of women in the labour force. A nagging economic problem continuing into the 1980s, however, was the weak growth in productivity and the inability to raise efficiency in manufacturing. Frustrations continued to result in outbreaks of antigovernment guerrilla activity, and the illegal drug trade remained as a major political and economic problem. (W.P.McG.)

In fact, the drug trade dominated the 1990 presidential campaign; three presidential candidates and hundreds of other people were killed by drug traffickers in a backlash against tougher drug-trade policies. Despite threats of terrorism, however, about 50 percent of the population voted in the peaceful May election, which was won by former finance minister and hard-line anti-drug-trade candidate César Gaviria Trujillo of the Liberal Party, with just under 50 percent of the vote. (Ed.)

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

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

BIBLIOGRAPHY

**Physical and human geography.** *General works:* Basic descriptive information is available in *The South American Handbook* (annual); HARVEY F. KLINE, *Colombia: Portrait of Unity and Diversity* (1983); HOWARD I. BLUTSTEIN *et al., Area Handbook for Colombia,* 3rd ed. (1977); and ERNESTO GUHL, *Colombia,* expanded ed. (1975– ), in Spanish. See also relevant sections of ARTHUR S. MORRIS, *South America,* 2nd ed. (1982); and PRESTON E. JAMES, C.W. MINKEL, and EILEEN W. JAMES, *Latin America,* 5th ed. (1986). A general atlas is INSTITUTO GEOGRÁFICO "AGUSTÍN CODAZZI," *Atlas de Colombia,* 3rd ed., rev. and enlarged (1977). Statistical information may be found in *Colombia estadistica* (annual). Current commerce and economic information is available in *Colombia Today* (monthly). Recent essays on politics, economics, and literature are found in MARIO ARRUBLA *et al., Colombia, hoy* (1978, reprinted 1981). An outline of prehistoric cultural developments in Colombia is given in G. REICHEL-DOLMATOFF, *Colombia* (1965); and in WARWICK BRAY, *The Gold of El Dorado* (1978).

*The land and the people:* WILLIAM F. JENKS (ed.), *Handbook of South American Geology: An Explanation of the Geologic*

*Map of South America* (1956), contains technical information on the physical features of the continent, including those of Colombia. See also VÍCTOR MANUEL PATIÑO, *Los recursos naturales de Colombia: aproximación y retrospectiva* (1980); and ERNESTO GUHL, *La sabana de Bogotá, sus alrededores y su vegetación* (1981). Distribution of plants and animals is discussed in E.J. FITTKAU (ed.), *Biogeography and Ecology in South America*, 2 vol. (1968–69). STEVEN L. HILTY and WILLIAM L. BROWN, *A Guide to the Birds of Colombia* (1986), is an authoritative work. Analyses of Colombia's agricultural progress include T. LYNN SMITH, *Colombia: Social Structure and the Process of Development* (1967); and DIETER BRUNNSCHWEILER, *The Llanos Frontier of Colombia: Environment and Changing Land Use in Meta* (1972). Studies of the people and geography of specific areas are found in ROBERT C. WEST, *The Pacific Lowlands of Colombia: A Negroid Area of the American Tropics* (1957), a physical, historical, and human geography of northwestern Colombia; ORLANDO FALS-BORDA, *Peasant Society in the Colombian Andes: A Sociological Study of Saucío* (1955, reissued 1976), a highly recommended work on social organization, culture, and ecology; B. LE ROY GORDON, *Human Geography and Ecology in the Sinú Country of Colombia* (1957, reprinted 1977), a regional study of northern Colombia; and JAMES J. PARSONS, *Antioqueño Colonization in Western Colombia*, 2nd rev. ed. (1968), and *Antioquia's Corridor to the Sea: An Historical Geography of the Settlement of Urabá* (1967).

*The economy:* Economic development and current policy are discussed in *Colombia: Economic Structure* (1985), a report issued by the Economic Research Department of Colombia's Banco de la República; WILLIAM PAUL MCGREEVEY, *An Economic History of Colombia 1845–1930* (1971), and "The Transition to Economic Growth in Colombia," in ROBERTO CORTÉS CONDE and SHANE J. HUNT (eds.), *The Latin American Economies: Growth and the Export Sector, 1880–1930* (1985), pp. 23–91; MIGUEL URRUTIA, *Winners and Losers in Colombia's Economic Growth of the 1970s* (1985); WORLD BANK, *Colombia: Economic Development and Policy Under Changing Conditions* (1984); R. ALBERT BERRY and RONALD SOLIGO (eds.), *Economic Policy and Income Distribution in Colombia* (1980); and R. ALBERT BERRY and MIGUEL URRUTIA, *Income Distribution in Colombia* (1976).

*Government and social conditions:* Government, social policy, and contemporary conditions are the focus of RICHARD E. HARTWIG, *Roads to Reason: Transportation, Administration, and Rationality in Colombia* (1983); DANIEL H. LEVINE, *Religion and Politics in Latin America: The Catholic Church in Venezuela and Colombia* (1981); ALEJANDRO ANGULO-NOVOA, "The Family in Colombia," ch. 4 in MAN SINGH DAS and CLINTON J. JESSER (eds.), *The Family in Latin America* (1980), pp. 84–105; MARK RUHL, *Colombia: Armed Forces and Society* (1980); MARCELO SELOWSKY, *Who Benefits from Government Expenditure?: A Case Study of Colombia* (1979); MICHAEL B. WHITEFORD, *The Forgotten Ones: Colombian Countrymen in an Urban Setting* (1976); and JAMES L. PAYNE, *Patterns of Conflict in Colombia* (1968). Educational policy and history are discussed in JUAN ANTONIO GÓMEZ (ed.), *A dónde va la educación colombiana* (1980); and FRANK SAFFORD, *The Ideal of the Practical: Colombia's Struggle to Form a Technical Elite* (1976).

*Cultural life:* JORGE ARANGO and CARLOS MARTÍNEZ, *Arquitectura en Colombia: arquitectura colonial 1538–1810, arquitectura contemporánea en cinco años 1946–1951* (1951), is a fine text in Spanish, English, and French covering these two important periods. GEORGE LIST, *Music and Poetry in a Colombian Village: A Tri-Cultural Heritage* (1983), is a study of the indigenous musical heritage. ERNESTO PORRAS COLLANTES, *Bibliografía de la novela en Colombia* (1976), includes plot summaries, excerpts from reviews, and lists of translations.

**History.** General works include ACADEMIA COLOMBIANA DE HISTORIA, *Historia extensa de Colombia* (1964– ), a multivolume work, covering all facets of Colombian history from precolonial to contemporary times, useful to the specialist; ROBERT H. DAVIS, *Historical Dictionary of Colombia* (1977), a convenient reference manual for people, events, and other aspects of Colombian history, with an excellent bibliography; and GABRIEL GARCÍA MÁRQUEZ, *One Hundred Years of Solitude* (1970, reissued 1982; originally published in Spanish, 1967), a novel based on aspects of Colombian history. Treatments of specific periods in Colombian history include MARCO PALACIOS, *Coffee in Colombia, 1850–1970: An Economic, Social, and Political History* (1980; originally published in Spanish, 1979), an outstanding resource; DAVID BUSHNELL, *The Santander Regime in Gran Colombia* (1954, reprinted 1970); CHARLES W. BERGQUIST, *Coffee and Conflict in Colombia: 1886–1910* (1978); STEPHEN J. RANDALL, *The Diplomacy of Modernization: Colombian-American Relations, 1920–1940* (1977); VERNON LEE FLUHARTY, *Dance of the Millions: Military Rule and the Social Revolution in Colombia, 1930–1956* (1957, reprinted 1975); JAMES D. HENDERSON, *When Colombia Bled: A History of the "Violencia" in Tolima* (1985); HERBERT BRAUN, *The Assassination of Gaitán: Public Life and Urban Violence in Colombia* (1985); PAUL OQUIST, *Violence, Conflict, and Politics in Colombia* (1980); and R. ALBERT BERRY, RONALD G. HELLMAN, and MAURICIO SOLAÚN (eds.), *Politics of Compromise: Coalition Government in Colombia* (1980). See also ROBERT H. DIX, *Colombia: The Political Dimensions of Change* (1967); and ORLANDO FALS-BORDA, *Subversion and Social Change in Colombia*, rev. ed. (1969; originally published in Spanish, 1967). Useful biographies include DAVID BUSHNELL (ed.), *The Liberator Simón Bolívar: Man and Image* (1970); JAMES WILLIAM PARK, *Rafael Núñez and the Politics of Colombian Regionalism, 1863–1886* (1985); and RICHARD E. SHARPLESS, *Gaitán of Colombia* (1978).

(C.Ga./W.P.McG./Ja.J.P.)

# Biological Coloration

Biological coloration is the general appearance of an organism as determined by the quality and quantity of light that is reflected or emitted from its surfaces. Coloration depends upon several factors: the colour and distribution of the organism's biochromes (pigments), particularly the relative location of differently coloured areas; the shape, posture, position, and movement of the organism; and the quality and quantity of light striking the organism. The perceived coloration depends also on the visual capabilities of the viewer. Coloration is a dynamic and complex characteristic and must be clearly distinguished from the concept of "colour," which refers only to the spectral qualities of emitted or reflected light.

Many evolutionary functions have been suggested for the effects of coloration on optical signaling. An organism with conspicuous coloration draws attention to itself, with some sort of adaptive interaction the frequent result. Such "advertising" coloration may serve to repel or attract other animals. While conspicuous coloration emphasizes optical signals and thereby enhances communication, coloration may, conversely, suppress optical signals or create incorrect signals and thereby reduce communication. This "deceptive" coloration serves to lessen detrimental or maladaptive interactions with other organisms.

Coloration may also affect an organism in ways other than its interaction with other organisms. Such nonoptical functions of coloration include physiological roles that depend on the molecular properties (*e.g.*, strength and type of chemical bonds) of the chemicals that create colour. For example, dark hair is mechanically stronger than light hair, and dark feathers resist abrasion better than light feathers. Coloration may also play a part in the organism's energy budget, because biochromes create colour by the differential reflection and absorption of solar engery. Energy absorbed as a result of coloration may be used in biochemical reactions, such as photosynthesis, or it may contribute to the thermal equilibrium of the organism. Nonoptical functions of coloration also include visual functions in which coloration or its pattern affects an animal's own vision. Surfaces near the eye may be darkly coloured, for instance, to reduce reflectance that interferes with vision.

Emitted light, the product of bioluminescence, forms a portion of the coloration of some organisms. Bioluminescence may reveal an organism to nearby animals, but it may also serve as a light source in nocturnal species or in deepwater marine animals such as the pinecone fishes (*Monocentris*). These fishes feed at night and have bright photophores, or bioluminescent organs, at the tips of their lower jaws; they appear to use these organs much like tiny searchlights as they feed on planktonic (minute floating) organisms.

Because many pigments are formed as the natural or only slightly modified by-products of metabolic processes, some coloration may be without adaptive function. Nonfunctional coloration can, for example, be an incidental effect of a pleiotropic gene (a gene that has multiple effects), or it can result from pharmacological reaction (as when the skin of a Caucasian person turns blue in cold water) or from pure chance. It seems unlikely, however, that any apparently fortuitous coloration could long escape the process of natural selection and thus remain totally without function.

Regardless of its adaptive advantages, a particular coloration or pattern of coloration cannot evolve unless it is within the species' natural pool of genetic variability. Thus a species may lack a seemingly adaptive coloration because genetic variability has not included that coloration or pattern in its hereditary repertoire.

Because humans are highly visual animals, we are naturally interested in and attentive to biological coloration. Human attention to coloration ranges from the purely aesthetic to the rigidly pragmatic. Soft, pastel colorations aid in increasing work efficiency and contribute to tranquil moods; bright, strongly contrasting colours seem to contribute to excitement and enthusiasm. These phenomena may be extensions of the basic human response to the soft blue, green, and brown backgrounds of the environment as opposed to sharply contrasting warning colorations found on many dangerous organisms. It is possible that much of the aesthetic value humans attach to coloration is closely related to its broad biological functions.

Human interest in coloration has led to biological studies. The classical work by the Moravian abbot Gregor Mendel on inherited characteristics, based largely on plant coloration, formed the foundation for modern genetics. Coloration also aids in the identification of organisms. It is an easily perceived, described, and compared characteristic. Related species living in different habitats, however, frequently have strikingly different colorations. Since coloration is susceptible to alteration in various functional contexts, it usually lacks value as a conservative characteristic for determining systematic relationships between all but the most closely related species.

Related articles of interest include BEHAVIOUR, ANIMAL; MIMICRY. (G.S.Lo./E.H.B.)

This article is divided into the following sections:


Structural and biochemical bases for colour 586
  Structural colours (schemochromes) 586
    Reflection
    Interference
    Scattering
  Pigments (biochromes) 586
    Chemical and biochemical features
    Carotenoids
    Quinones
    Flavonoids
    Tetrapyrroles, porphyrins, and their derivatives
    Indole pigments
    Phenoxazones and sclerotins
    Purines and pterins
    Flavins (lyochromes)
    Miscellaneous pigments
Control of coloration 589
  Genetic control 589
  Physiological control 589
The adaptive value of biological coloration 590
  Optical functions: deceptive coloration 590
    Camouflage
    Mimicry
  Optical functions: advertising coloration 591
    Attraction
    Repulsion
  Optical functions: combination of concealing and advertising coloration 592
  Optical functions: the roles of the selective agent and of illumination 592
    The selective agent
    Illumination
  Visual functions 592
  Physiological functions 593
Coloration changes 593
  Coloration changes in individual organisms 593
    Short-term changes
    Seasonal changes
    Age-related changes
  Coloration changes in populations 594
Bibliography 594

## Structural and biochemical bases for colour

Organisms produce colour physically, by submicroscopic structures that fractionate incident light into its component colours (schemochromes); or chemically, by natural pigments (biochromes) that reflect or transmit (or both) portions of the solar spectrum. Pigmentary colours, being of molecular origin, may be expressed independently of structural colour and are not altered by crushing, grinding, or compression. Structural colours are often reinforced by the presence of biochromes and are altered or destroyed by crushing, grinding, or compression.

### STRUCTURAL COLOURS (SCHEMOCHROMES)

The physical principles of total reflection, spectral interference, scattering, and, to some extent, polychromatic diffraction, all familiar in reference to inanimate objects, are also encountered among tissues of living forms, most commonly in animals. In plants these physical principles are exemplified only by the total reflection of white light by some fungi and bacteria and by the petals of some flowers and barks, and by some spectral interference in certain sea plants.

**Reflection.** Total reflection of light—which imparts whiteness to flowers, birds' feathers, mammalian hair, and the wings of certain butterflies—often results from the separation of finely divided materials by minute air spaces. Secretions or deposits in tissues may also contribute to the whiteness; for example, the fat and protein in mammalian milk and the calcium carbonate in the shells of mollusks, crustaceans, certain echinoderms, corals, and protozoans.

Iridescent colours

**Interference.** Fractionation of white light into its components occurs in organisms (chiefly animals) through interference: the incident light penetrates the animal structure and is reflected back through successive ultrathinly layered films, giving striking iridescence, even in diffuse light, as a result of the asynchrony between the wavelengths of visible light that enter and those that return.

Brilliant interference colours may display variety or be predominantly of one kind, depending upon the relative thicknesses of layers and interlaminar spaces giving rise to the colours. Such colours also are changeable with the angle of vision of the viewer.

Purely prismatic refraction of light (sometimes confused with interference iridescence) is probably rare in animals and is limited to instances in which direct beams of light impinge upon certain microcrystalline deposits. Polychromatic diffraction—*e.g.*, by natural, fine gratings or regular fine striations—may be observed among certain insects, but, like prismatic refraction, it is conspicuous only when a direct beam of light strikes such structures and they are viewed at an angle.

**Scattering.** A special instance of diffraction, often referred to as the Tyndall effect (after its discoverer, the 19th-century British physicist John Tyndall), results in the presence of blue colours in many animals. The Tyndall effect arises from the reflection of the shorter (blue) waves of incident light by finely dispersed particles situated above the dark layers of pigment, commonly melanin deposits. In these blue-scattering systems, the reflecting entities—whether very small globules of protein or lipid, semisolid substances in aqueous mediums, or very small vesicles of air—are of such small size as to approximate the shorter wavelengths of light (about 0.4 micron). The longer waves, such as red, orange, and yellow, pass through such mediums and are absorbed by the dark melanin below; the short waves, violet and blue, encounter bodies of approximately their own dimensions and consequently are reflected back.

Origin of Tyndall effect in animals

Two types of coloration may act in combination; in some instances, for example, structurally coloured and pigmented layers may be superimposed. Most of the greens found in the skin of fishes, amphibians, reptiles, and birds do not arise from the presence of green pigments (although exceptions occur); rather, they result from the emergence of scattered blue light through an overlying layer of yellow pigment. Extraction of the yellow pigment from the overlying cuticle of a green feather or of a reptilian skin leaves the object blue. (D.L.F./E.H.B.)

### PIGMENTS (BIOCHROMES)

Plants and animals commonly possess characteristic pigments. They range in plants from those that impart the brilliant hues of many fungi, through those that give rise to the various browns, reds, and greens of species that can synthesize their food from inorganic substances (autotrophs), to the colourful pigments found in the flowers of seed plants. The pigments of animals are located in nonliving skin derivatives such as hair in mammals, feathers in birds, scales in turtles and tortoises, and cuticles and shells in many invertebrates. Pigments also occur within living cells of the skin. The outermost skin cells may be pigmented, as in humans, or special pigment-containing cells, chromatophores, may occur in the deeper layers of the skin. Depending on the colour of their pigment, chromatophores are termed melanophores (black), erythrophores (red), xanthophores (yellow), or leucophores (white). (F.A.B.)

**Chemical and biochemical features.** The colour of a chemical compound depends on the selective absorption of light by molecules whose size or vibrational wavelengths or both lie between 3000 and 7000 angstroms (one angstrom equals $10^{-7}$ millimetre). Selective absorption of visible light results from retardation in the relative speed or vibrational frequency of the many rapidly vibrating electron pairs found in a compound. Sufficient modification in the frequency of vibration imparts to the whole molecule a special motion, or chemical resonance, that absorbs entering light rays of matching frequency with the evolution of heat; the residual, unabsorbed light is transmitted to the eye.

Selective absorption of light rays

If the molecular resonance involves short, rapid waves, the shorter visible light waves are absorbed (*i.e.*, violet and blue) and the compound appears yellow or orange; red-appearing substances, having slightly longer resonance values, absorb light from the blue and green regions; and blue and green compounds result from cancellation of light in the red or orange realms. Black substances absorb all light equally and completely; white compounds absorb no light in the visible spectrum. The colour reflected by a pigment usually includes all the wavelengths of visible light except the absorbed fraction; the observed colour of a compound thus depends upon the dominant wavelength reflected or transmitted.

Distinction between nitrogenous and nonnitrogenous pigments

The more important natural pigments may be grouped into (1) classes whose molecules lack nitrogen and (2) those that contain nitrogen. Of the nonnitrogenous pigments, by far the most important, conspicuous, and widely distributed in both plants and animals are the carotenoids. Naphthoquinones, anthraquinones, and flavonoids are other nitrogen-free pigments that occur in animals, all being synthesized originally in plants, as are the carotenoids. But unlike the carotenoids, the others have a limited distribution in animals, and little is known of their physiological attributes in either kingdom.

Prominent among the nitrogenous biochromes are the tetrapyrroles, including both the porphyrins (*i.e.*, the red or green heme compounds present in the blood of many animals and the green chlorophylls of many plants) and the bile pigments, which occur in many secretions and excretory products of animals and in plant cells. Equally prominent are the melanins, which are dark biochromes found in skin, hair, feathers, scales, and some internal membranes; they represent end products from the breakdown of tyrosine and related amino acids.

Below are outlined the basic colours, sources, and metabolic features of some representative biological pigments.

**Carotenoids.** The carotenoids constitute a group of yellow, orange, or red pigments of almost universal distribution in living things. Carotenoids generally are insoluble in water but dissolve readily in fat solvents such as alcohol, ether, and chloroform. They are readily bleached by light and by exposure to atmospheric oxygen and are also unstable in acids such as sulfuric acid.

Two basic types of carotenoids

Carotenoids occur as two major types: the hydrocarbon class, or carotenes, and the oxygenated (alcoholic) class, or xanthophylls. Some animals exhibit a high degree of selectivity for the assimilation of members of one or the

other class. The horse (*Equus caballus*), for instance, absorbs through its intestine only the carotenes, even though its green food contains mostly xanthophylls; the domestic hen (*Gallus domesticus*), on the other hand, stores only members of the xanthophyll class, as do many fishes and invertebrates. Other animals, including certain frogs, *Octopus* species, and humans, assimilate and store both classes in the liver and in fat deposits.

Carotenoids are synthesized by bacteria, fungi, algae, and other plants to highly evolved flowering forms, in which they are most conspicuous in petals, pollen, fruit, and some roots—*e.g.,* carrots, sweet potatoes, tomatoes, and citrus fruits. All animals and protozoans contain carotenoids, although the blood plasma of a number of mammals (*e.g.,* swine, sheep, goats, some carnivores) is almost entirely free of these pigments. The livers of animals often yield carotenoids; all animals depend upon a nutritional supply of vitamin A or one of its precursors, such as carotene, for maintenance of normal metabolism and growth. Carotenoids are relatively more concentrated in such structures as ovaries, eggs, testes (some animals), the liver (or the liver-like analogue of invertebrates), adrenal glands, skin, and eyes. In birds, carotenoid pigmentation may be conspicuous in the yellow tarsal (lower leg) skin, external ear, body fat, and egg yolk (especially in poultry) and in red-coloured feathers. Carotenoids are also found in the wings or wing covers of many insects and in the milk fat of cattle.

**Quinones.** The quinones include the benzoquinones, naphthoquinones, anthraquinones, and polycyclic quinones.

*Benzoquinones.* Benzoquinones occur in certain fungi and in roots, berries, or galls (abnormal growths) of higher plants, from which they can be recovered as yellow, orange, red, violet, or darker coloured crystals or solids. Small quantities of pale-yellow crystals of coenzyme Q, often called ubiquinones, are almost universally distributed in plants and animals. The ubiquinones impart no recognizable coloration to an organism because of their very small concentrations; they play an important role, however, as respiratory enzymes in catalyzing cellular oxidations.

Ubiquinones

*Naphthoquinones.* Naphthoquinones are encountered in some bacteria and in the leaves, seeds, and woody parts of higher plants. They can be recovered as yellow, orange, red, or purple crystals. They are soluble in organic solvents and have been used extensively as dyes for fabrics. Among the naphthoquinones of biochemical and physiological importance are the K vitamins. Another series within the naphthoquinone class manifests conspicuous red, purple, or sometimes green colours in a few animal types. These are the echinochromes and spinochromes, so named because they are conspicuous in tissues and in the calcareous tests (shells) of echinoids, or sea urchins.

*Anthraquinones.* The anthraquinones occur widely in plants but in only a few animals. These brilliantly coloured compounds have found wide application as dyes and as chemical indicators of acidity or alkalinity.

*Polycyclic quinones.* The polycyclic quinones occur in some bacteria, fungi, and parts of higher plants. One of the more interesting representatives is the aphin group, so called because of their initial recovery from the hemolymph (circulating fluid) of several coloured species of aphids; aphids parasitize plants, as do the other quinone-assimilating insects.

**Flavonoids.** The biochromes in the class of flavonoids, another instance of compounds lacking nitrogen, are extensively represented in plants but are of relatively minor and limited occurrence in animals, which rely on plants as sources of these pigments. Flavonoids consist of a 15-carbon skeleton compound called flavone (2-phenylbenzopyrone), in which one or more hydrogen atoms (H) is replaced either by hydroxyl groups (—OH) or by methoxyl groups ($-OCH_3$). Flavonoids occur in living tissue mainly in combination with sugar molecules, forming glycosides. Many members of this group, notably the anthoxanthins, impart yellow colours, often to flower petals; the class also includes the anthocyanins, which are water-soluble plant pigments exhibiting orange-reds, crimson, blue, or other colours.

*Anthoxanthins.* The variety of anthoxanthins is greater than that of anthocyanins, and new anthoxanthins are continuously being discovered. A prominent flavonoid is the pale-yellow flavonal quercitin, first isolated from an oak (*Quercus*) but widely distributed in nature. A weak acid, it combines with strong acids to form orange salts, which are not very stable and readily dissociate in water. Quercitin is a strong dyestuff; it yields more than one colour, depending on the mordant used. A yellow pigment isolated from the wings of the butterfly *Melanargia galatea* possesses chemical properties closely allied to those of quercitin. Other well-known anthoxanthins include chrysin, found in the leaf buds of the poplar (*Populus*), and apigenin, found in the leaves, stem, and seeds of parsley (*Petroselinum*) and the flowers of the camomile (*Anthemis*).

Quercitin

*Anthocyanins.* The anthocyanins are largely responsible for the red colouring of buds and young shoots and the purple and purple-red colours of autumn leaves. The red colour becomes apparent when the green chlorophyll decomposes with the approach of winter. Intense light and low temperatures favour the development of anthocyanin pigments. Some leaves and flowers lose anthocyanins on reaching maturity; others gain in pigment content during development. Often an excess of sugars exists in leaves when anthocyanins are abundant. Injury to individual leaves may be instrumental in causing the sugar excess in such cases. Anthocyanins also occur in blossoms, fruits, and even roots (*e.g.,* beets) and occasionally in larval and adult flies and in true bugs (Heteroptera).

A typical anthocyanin is red in acid, violet in neutral, and blue in alkaline solution. Thus, the blue cornflower, the bordeaux-red cornflower, the deep-red dahlia, and the red rose contain the same anthocyanin, the variation in colour resulting from the different degrees of acidity and alkalinity of the cell sap. More than one anthocyanin may be present in a flower or blossom, and the colours of many flowers are caused by the presence of both anthocyanins and plastid pigments in the tissues. Moreover, small genetic changes in varieties or species may be associated with the development of different anthocyanins.

No physiological functions seem to have been definitely established for the flavonoids in animals and plants. It has been pointed out, however, that flower colour is valuable in attracting bees, butterflies, and other pollen-transporting visitors that implement fertilization in plants; brightly coloured fruits have improved seed dispersal by animals attracted to them as food.

**Tetrapyrroles, porphyrins, and their derivatives.** A biologically important class of water-soluble, nitrogenous 16-membered ring, or cyclic, compounds is referred to as porphyrins. The elementary structural unit of all porphyrins is a large ring itself composed of four pyrrole rings, or cyclic tetrapyrroles. This basic compound is known as porphin.

Porphyrins combine with metals (metalloporphyrins) and protein. They are represented by the green, photosynthetic chlorophylls of higher plants and by the hemoglobins in the blood of many animals.

*Porphyrins.* Many invertebrates display in their skins or shells porphyrin pigments (or salts of them), some showing fluorescence (*i.e.,* the emission of visible light during exposure to outside radiation). In addition, various porphyrins occur in secretory and excretory products of animals, and some kinds, predominantly the phorbides, which result from the breakdown of chlorophyll, have been recovered from ancient natural deposits such as coal and petroleum and from muds of long-buried marine and lacustrine strata. Ooporphyrin is responsible for the red flecks on the eggshells of some plovers and many other birds. The African turacos (Musophagidae) secrete a copper salt of uroporphyrin III into their wing feathers. This deep-red pigment, turacin, is readily leached from the feathers by water containing even traces of alkali. The green plumes of these birds owe their colour to the presence of turacoverdin, a derivative of turacin.

Porphyrins in secretory and excretory products

*Hemoglobins.* Hemoglobins are present in the red blood cells of all vertebrate animals and in the circulatory fluids of many invertebrates, notably annelid worms, some arthropods, echinoderms, and a few mollusks. The hemoglobin molecule consists of a heme fraction and a

globin fraction; the former consists of four pyrrole moieties (porphin) with a ferrous iron ($Fe^{2+}$) atom in the centre. The globin fraction is a protein that may constitute more than 90 percent of the total molecular weight of hemoglobin. Hemoglobins have the capacity to combine with atmospheric oxygen in lungs, gills, or other respiratory surfaces of the body and to release oxygen to tissues. They are responsible for the pink to red colours observed in combs and wattles of birds and in the skin of humans and other primates. Particularly prominent are portions of the face, buttocks, and genital areas of baboons.

*Chlorophylls.* Chlorophyll is one of the most important pigments in nature. Through the process of photosynthesis, it is capable of channeling the radiant energy of sunlight into the chemical energy of organic carbon compounds in the cell. For a detailed account of this process, see PHOTOSYNTHESIS. A pigment very much like chlorophyll was probably the first step in the evolution of self-sustaining life. Chlorophyll exists in several forms. Chlorophylls *a* and *b* are the chief forms in higher plants and green algae; bacteriochlorophyll is found in certain photosynthetic bacteria.

Types of chlorophyll

The chlorophylls are magnesium porphyrin compounds in which a cyclic tetrapyrrole is attached to a single central magnesium atom. They contain two more hydrogen atoms than do other porphyrins. The various forms differ in minor modifications of side groups attached to the pyrrole groups. In higher plants, chlorophyll is bound to proteins and lipids aschloroplastin in definite and specific laminations in bodies called chloroplasts. The combination of chlorophyll with protein in chloroplastin is of special significance, because only as a result of the combination is chlorophyll able to remain resistant to light.

*Bilins.* Among the metabolic products of certain porphyrins, including the heme portion of hemoglobin, is a series of yellow, green, red, or brown nonmetallic compounds arranged as linear, or chain, structures rather than in the cyclic configuration of porphyrins. These are the so-called bilins, or bilichromes. Small quantities of the red waste compound, bilirubin ($C_{33}H_{36}O_6N_4$); a green product formed from it by the removal of two hydrogen atoms, biliverdin ($C_{33}H_{34}O_6N_4$); and various other chemically similar compounds occur in normal tissues and may be conspicuous in excretory or secretory materials under normal circumstances and certain pathological conditions. The bile pigments, although first identified in mammalian tissues or products (*e.g.,* in the bile of the gall bladder), are by no means confined thereto. Various members of the bilichrome series are encountered in invertebrates, lower vertebrates, and in red algae and green plants.

Linear tetrapyrroles

Although the bile pigments of animals arise in all probability from the catabolism of heme precursors, there is evidence that bilirubin, accompanied by iron salts, promotes the synthesis of new hemoglobin when injected into humans, dogs, or rabbits suffering from secondary anemia.

In addition to the chlorophylls, plants also contain linear bilichromes, which have especially important roles in green plants. Among them are the blue phycocyanins and the red phycoerythrins, which serve, in red algae, as accessory pigments in photosynthesis. Another example is phytochrome, a bilichrome pigment of blue colour, which, although present in very minute quantities in green plants, is indispensable in various photoperiodic processes.

Types of phytochrome

Phytochrome exists in two alternative forms: $P_{660}$ and $P_{730}$. Of these, $P_{730}$ triggers the germination and respiration of seeds (and of spores of ferns and mosses), the flowering of long-day plants (or inhibition of flowering in short-day plants), etiolation (growth in darkness), cuticle coloration, anthocyanin synthesis (*e.g.,* in apples, red cabbage, and turnips), and several structural and physiological responses. $P_{660}$ is capable of reversing many physiological reactions initiated by $P_{730}$. Even very brief exposures to light absorbed by $P_{660}$ delays flowering in some short-day plants otherwise geared to flower by previous exposure to light of such wavelength that only the $P_{730}$ phytochrome is involved. Much yet remains to be learned about the biochemistry of phytochromes and the reactions catalyzed or otherwise regulated by them.

**Indole pigments.** *Melanins.* These pigments produce buff, red-brown, brown, and black colours. Melanins occur widely in the feathers of birds; in hair, eyes, and skin of mammals, including humans; in skin or scales or both of many fishes, amphibians, and reptiles; in the ink of cephalopods (octopus, squid); and in various tissues of many invertebrates.

Melanins are polymers (compounds consisting of repeating units) of variable mass and complexity. They are synthesized from the amino acid tyrosine by progressive oxidation, a process catalyzed by the copper-containing enzyme tyrosinase. Extractable in very dilute alkali, melanins are also soluble when fresh and undried in very dilute acid solutions; they are bleached by hydrogen peroxide, which is sometimes applied to growing hair to create a blond effect, and by chlorine, chromate, and permanganate.

Composition of melanin

Pale-yellow, tawny, buff, reddish, brown, and black colours of hair and some feathers can arise from the presence of melanins in various phases of formation or subdivision in granules. The dark, light-absorbing sublayers of melanin that intensify reflected structural (Tyndall) blues or iridescent displays in feathers were mentioned above. Black melanins and brown melanoproteins occur in many invertebrate animals. Certain worms and many crustaceans and mollusks exhibit melanism in the skin.

The degree of natural melanization depends upon relative concentrations of copper and of the copper-containing enzyme tyrosinase. Dark hairs contain higher traces of copper than pale hairs do; should the intake of copper fall substantially below a fraction of a milligram per day, new fur emerges successively less dark. This trend is reversed by restoring sufficient copper to the diet.

All human skin except that of albinos contains greater or lesser amounts of melanin. In fair-skinned persons the epidermis, or outermost layer of the skin, contains little of the pigment; in the dark-skinned races epidermal deposits of melanin are heavy. On exposure to sunlight, human epidermis undergoes gradual tanning with increases in the melanin content, which helps to protect underlying tissues from injurious sun rays (see also INTEGUMENTARY SYSTEMS).

Tanning in man

*Indigoids.* Like melanins, the indigo compounds are excretory metabolic breakdown products in certain animals. But, in contrast to the melanins, their distribution as conspicuous pigmentary compounds is very limited, and they are not dark but red, green, blue, or purple.

One of the most common members of this group is indigo, or indigotin, which occurs as a glucoside (*i.e.,* chemically combined with glucose) in many plants of Asia, the East Indies, Africa, and South America. It has long been used as a blue dye.

**Phenoxazones and sclerotins.** Once confused with melanins, biochromes such as phenoxazones and sclerotins show a similar colour series (yellow, ruddy, brown, or black). Genetic research, notably with reference to eye pigments of the fruitfly, *Drosophila melanogaster,* has resulted in the description of a class of so-called ommochromes, which are phenoxazones. The ommochromes not only are conspicuous in the eyes of insects and crustaceans but have also been detected in the eggs of the echiurid worm *Urechis caupo* and in the changeable chromatophores in the skin of cephalopods. In addition to being responsible for the brown, vermilion, cinnabar, and other colours of insect eyes, ommochromes are also sometimes present in the molting fluid and integument. They are distinguished from the melanins by solubility in formic acid and in dilute mineral acids, by manifestation of violet colours in concentrated sulfuric acid, and by showing reversible colour changes with oxidizing and reducing agents. The ommochromes, which are derived from breakdown of the amino acid tryptophan, include ommatins and ommins. The ommatins, although complex in chemical structure, are relatively small molecules. The ommins are large molecules, in which the chromogenic moiety is seemingly condensed with longer chains, such as peptides (amino acids linked together).

Ommochromes in invertebrates

Sclerotins arise as a result of an enzyme-catalyzed tanning of protein. Certain roaches secrete a phenolase enzyme, the glucoside of a dihydroxyphenol, and a glycosidase. Mixing of these substances results in the release of the

phenolic compound from glucose and its combination, via a reaction catalyzed by the phenolase, with protein; the products are pink, ruddy, and ultimately dark-brown polymers that are incorporated into the insect's body cuticle and egg cases. Similar reactions take place in the carapace (the shell covering the body) of certain crustaceans.

**Purines and pterins.** Although the purine compounds cannot be classed as true pigments—they characteristically occur as white crystals—they often contribute to the general colour patterns in lower vertebrates and invertebrates. That purines are excretory materials is illustrated by the uric acid (or urates) and guanine found in the excrement of birds and of uric acid found in that of reptiles. Uric acid has also been detected in the mucus excreted by sea anemones, and urates are present in small amounts in the urine of higher apes and humans.

Manifestation of uric acid

The white, silvery, or iridescent chromatophores, both stationary iridocytes and changeable leucophores, of some fishes, amphibians, lizards, and cephalopods contain microcrystalline aggregates of the purine guanine; a layer of white skin on the underside of many fishes, called the stratum arginatum, is particularly rich in guanine.

Closely related to the purines and formerly classed among them are the pterins, so named from their notable appearance in and first chemical isolation from the wings of certain butterflies. Both purines and pterins contain a six-atom pyrimidine ring; in purines this ring is chemically condensed with an imidazole ring; pterins contain the pyrazine ring. Pterins occur as white, yellow, orange, or red granules in association with insect wing material.

**Flavins (lyochromes).** Flavins constitute a class of pale-yellow, greenly fluorescent, water-soluble biochromes widely distributed in small quantities in plant and animal tissues. The most prevalent member of the class is riboflavin (vitamin $B_2$).

Flavins are synthesized by bacteria, yeasts, and green plants; riboflavin is not manufactured by animals, which therefore are dependent upon plant sources. Riboflavin is a component of an enzyme capable of combining with molecular oxygen; the product, which is yellow, releases the oxygen in the cell with simultaneous loss of colour.

Metabolic role of riboflavin

**Miscellaneous pigments.** The chemical constitution of many pigments remains imperfectly known. Only a few of the more conspicuous examples are mentioned below.

*Hemocyanins.* Copper-containing proteins called hemocyanins occur notably in the blood of larger crustaceans and of gastropod and cephalopod mollusks. Hemocyanins are colourless in the reduced, or deoxygenated, state and blue when exposed to air or to oxygen dissolved in the blood. Hemocyanins serve as respiratory pigments in many animals, although it has not been established that they perform this function wherever they occur.

*Hemerythrins.* Iron-containing, proteinaceous pigments, hemerythrins are present in the blood of certain bottom-dwelling marine worms (notably burrowing sipunculids) and of the brachiopod *Lingula;* the pigments serve as oxygen-carriers.

*Hemovanadin.* Pale-green pigment, hemovanadin, is found within the blood cells (vanadocytes) of sea squirts (Tunicata) belonging to the families Ascidiidae and Perophoridae. The biochemical function of hemovanadin, a strong reducing agent, is unknown.

The vanadium-containing pigment of tunicates

*Actiniochrome.* A relatively rare pigment, actiniochrome occurs in red or violet tentacle tips and in the stomodeum (oral region) of various sea anemones. The pigment plays no recognized physiological role.

*Adenochrome.* Adenochrome is a nonproteinaceous pigment that occurs as garnet-red inclusions at high concentrations in the glandular, branchial heart tissues of *Octopus bimaculatus.* The compound contains small amounts of ferric iron and some nitrogen and gives a positive reaction for pyrroles. It is believed to be an excretory product.

(D.L.F./E.H.B.)

## Control of coloration

### GENETIC CONTROL

Coloration is in large measure determined genetically. As mentioned earlier, the inheritance of colour in garden peas provided part of the basis for the pioneering studies of heredity by Mendel. These studies led Mendel to postulate the existence of discrete units of heredity that segregate independently of one another during the formation of reproductive cells. The studies also led to his discovery of the phenomenon of dominance. The basic units of heredity are now known as genes, and the variant forms of a given gene are termed alleles. Among species that reproduce sexually, an individual normally possesses a pair of alleles for any gene—one inherited from the female parent and one from the male parent. These two alleles are situated at corresponding loci on the paired chromosomes found in diploid cells—*i.e.,* cells containing two similar sets of complementary chromosomes. Segregation of the alleles occurs during formation of reproductive cells, with the result that only one of the pairs enters each cell, which is called a haploid cell.

In his experiments Mendel crossed purple-flowered peas with white-flowered ones. The plants he used in these crosses were true-breeding for flower colour, meaning that the purple-flowered plants were descended for generations from only other purple-flowered plants, and that the white-flowered plants were likewise descended for generations from only other white-flowered plants. Because of these true-breeding characteristics, Mendel postulated that the original plants were homozygous for the trait of flower colour—in other words, that each plant carried a pair of identical heredity units (*i.e.,* alleles) for this trait. When he crossed purple-flowered peas with white-flowered ones, he obtained a first filial ($F_1$) generation in which all the offspring had purple flowers. He therefore deduced that the unit for purple (usually designated *R*) was dominant over the unit for white (*r*). Thus in the parental generation the purple-flowered plants can be designated *RR* (indicating that they are homozygous for the dominant allele), and the white-flowered plants can be symbolized as *rr.* The $F_1$ plants were heterozygous for flower colour (*Rr*), but they expressed purple colour because of the complete dominance of the allele *R* over *r.*

Inheritance of flower colour in peas

Dominance may be incomplete, however; a crossing between homozygous red Japanese four-o'clocks (*Mirabilis*) and homozygous white ones yields heterozygous *Rr* offspring, which are all pink. A cross of the heterozygous pink generation of four-o'clocks with each other yields a second generation with the colour ratio of 25 percent red (*RR*), 50 percent pink (*Rr*), and 25 percent white (*rr*). This is because each of the parent ($F_1$) plants produces equal numbers of *R*- and *r*-containing reproductive cells through segregation, and there is a random chance of either type of male haploid cell (gamete) fertilizing either of the two female types. For peas, on the other hand, the ratio resulting from a cross of parent ($F_1$) plants is three purple (one *RR* and two *Rr*) to one white (*rr*) because of the dominance of *R.*

Although the principle of inheritance of colour and coloration patterns in all organisms is like that for the two plants described above, it is usually far more complex. Within the species population, a particular gene may have multiple alleles instead of two; thus numerous combinations within any individual may be possible; in addition, the coloration may depend upon genes at several sites. In this case either all pairs may segregate simultaneously and more or less independently into the gametes, or the genes may be linked in their inheritance by location on the same chromosome. Such possibilities, together with different degrees of dominance, result in tremendously complex hereditary bases for the genetic control of colour and colour patterns within many species. For a fuller treatment of these principles, see GENETICS AND HEREDITY, THE PRINCIPLES OF: *Mendelian genetics.*

### PHYSIOLOGICAL CONTROL

The development of coloration often depends upon regulatory substances (hormones) secreted by endocrine glands. In birds the level of the hormone thyroxine determines the coloration of feathers and bill, although specific seasonal biochromes are often laid down under the influence of sex hormones, as in the beak of the starling, which turns from black to yellow in early spring. The variability in

control among bird species is so great, however, that generalizations are impossible. Hormonally controlled colour changes also occur in mammals; for example, swellings in the genital areas that become pink due to vascularization during the reproductive season. The species specificity of coloration patterns, however, always depends on a genetically determined responsiveness of various target tissues to certain hormones.

Chromatophores occur in cephalopods, crustaceans, insects, fishes, amphibians, and lizards and are responsible for the most rapid colour changes. They allow conspicuous display of a biochrome by dispersing it in the chromatophore-bearing surface, or they conceal the biochrome by concentrating it into small areas. Chromatophores are of three kinds. The chromatophoric organs of cephalopods consist of an elastic sac filled with biochrome and controlled by a ring of radiating muscle fibres. These fibres contract in response to neural stimulation, thereby stretching the sac into a broad, thin disk. Chromatophoric syncytia occur in crustaceans, the movement of biochrome being due to the ebb and flow of cytoplasm through fixed tubular spaces that collapse when the cell is contracted and fill when the cell expands. Chromatophoric syncytia are hormonally controlled. Cellular chromatophores, the third kind, are found in vertebrates. In these cells melanic granules flow in stable cellular processes that maintain a fixed position, unlike the contracting and expanding processes of the syncytia. Control among vertebrates is varied: chromatophores of bony fishes are controlled by the autonomic nervous system; those of elasmobranch fishes (sharks and rays) and lizards are controlled by hormones and nerves; those of amphibians are regulated by hormones alone.

Types of chromatophores

One animal may contain biochromes of several colours, commonly red, yellow, black, and reflecting white; prawns also have a blue biochrome. By appropriate migrations of biochromes, an animal can achieve substantial alterations in colour or shade for varying periods of time. In prawns, dispersion of blue and yellow yields green; unequal dispersion of biochromes over parts of the body produces patterns of coloration.

Rapid physiological colour changes are supplemented by morphological ones, the animal either gradually synthesizing or destroying biochromes, usually in an adaptive manner (see the section *Coloration changes*).

(F.A.B./E.H.B.)

## The adaptive value of biological coloration

Coloration and the pattern of coloration play a central role in the lives of plants and animals—even those species in which vision is lacking or not the dominant sense. For example, cryptic coloration often goes hand in hand with cryptic behaviour; nonreflective colours occur on the faces of birds that forage in bright sunlight; and abrasion-resistant coloration occurs more often among species that inhabit abrasive habitats than among species that inhabit nonabrasive habitats. The functions of biological coloration fall into three broad categories: (1) optical functions, in which coloration affects the animal's or plant's visibility to other animals; (2) visual functions, in which coloration affects the animal's own vision; and (3) physiological functions, in which the molecular properties of biochromes play a role unrelated to either optical signaling or vision.

### OPTICAL FUNCTIONS: DECEPTIVE COLORATION

Deceptive coloration depends on four factors: the coloured organism, hereafter referred to as the organism; its model, which may be the background against which it is concealed; the spectral quality of the illumination; and the visual sensitivity and behaviour of the animal or animals that the organism is deceiving. To some extent the following discussion considers the relationships among the four factors separately; but in reality the deceptive, opitcal effect results from the interaction of all four factors. There are two basic types of deceptive coloration: (1) concealing coloration, or camouflage, in which the organism blends into its surroundings; and (2) mimicry, in which the organism is not hidden but rather presents a false identity by its resemblance to another species.

**Camouflage.** *Background matching.* Background matching is probably the most common form of concealment. It makes little difference whether the background model is an animate or inanimate object since both involve the initial establishment and continued maintenance of the concealment. Not only coloration but also the form and the activities or behaviour of the organism in relation to its model are important.

The simplest examples of background matching are provided by the fish eggs and planktonic (free-floating) larval fishes that exist in the uniformly blue environment of the open sea—*i.e.,* those that are pelagic. They usually possess minimal pigmentation and are transparent.

In other organisms and environments the behaviour and form of the organism become more important as adjuncts to coloration. Evidence of the importance of the choice of a proper background is provided by three differently coloured species of lizards of the genus *Anolis,* which form mixed hunting groups over the same background. Many of the individuals are easily perceived on this background, but, when disturbed, they conceal themselves by segregating according to species over the appropriately coloured backgrounds. Camouflage may also be accomplished through a change in coloration. Many flatfishes, for example, show a remarkable ability to match the pattern of the surface on which they are resting. Some nudibranchs, a group of marine gastropods, such as *Phestilla melanobrachia,* manage to establish and maintain their resemblance to the background by ingesting portions of their model, which is the living coral on which they live. The pigments in the coral polyps are deposited in diverticulae (branches) of the gut and occasionally in the epidermis and show through as nearly perfect camouflage. The slow-moving nudibranchs are very difficult to see on their coral host, and when they move to differently coloured coral, their coloration changes as their food source changes.

Alteration of coloration to match changing backgrounds

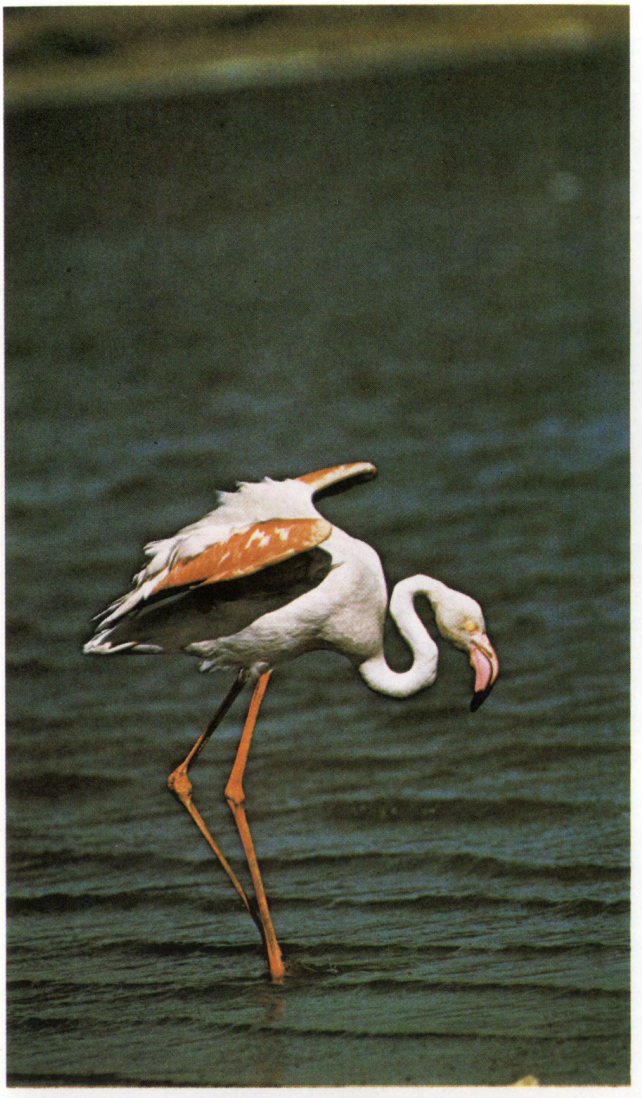
Some of the parasites that live on marine fishes conceal themselves in a similar manner. Flukes, or monogenean trematodes, gorge themselves on their hosts' tissues and biochromes and appear to remain within areas on the host that have similar pigmentation. The adaptive significance of the coloration is known to lie in escape from predation by the third party, cleaning organisms such as the fish *Labroides,* which feeds on the external parasites of other fishes. Several decorator crabs use portions of the model for concealment by picking up algae and sponges and placing them on the carapace (upper shell) to cover their own coloration; the algae and sponges continue to live as if in their normal habitat.

*Disruptive coloration.* Disruptive patterns, frequently a part of camouflage coloration, serve the function of visual disruption by forming a pattern that does not coincide with the contour and outline of the body. The blenny *Hypsoblennius sordidus,* for example, usually has a mottled coloration that crudely matches its background in terms of the size and colour of differently pigmented areas; it also has a series of darkly pigmented "saddles" that break up the outline contour of its back. This species also demonstrates the fact that the type of disruptive patterning may change when an individual shifts to another type of background. The saddled condition is found when the background is composed of disruptive elements of the same approximate size—*e.g.,* small sponges, barnacles, and patches of algae. But when the fish moves to an evenly coloured area, its coloration becomes stripes that run horizontally from head to tail.

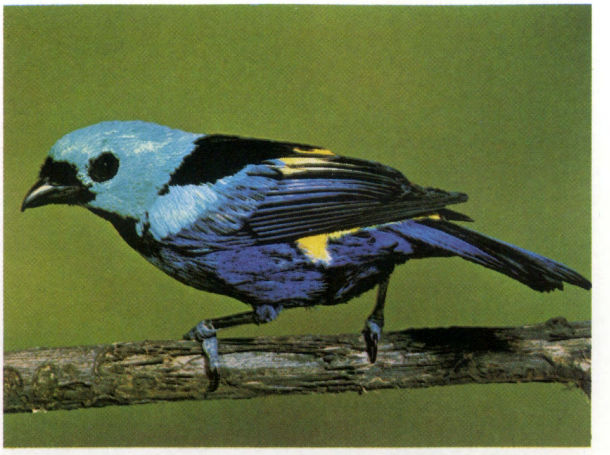
Disruptive patterns are found in the coloration of many fish that form schools over the reef during daylight hours for protection against predation. When a predator approaches, the fishes form dense schools in which all of the individuals orient in the same direction. The movement of many individuals, coupled with their similar disruptive coloration, presents an extremely confusing spectacle, presumably one that makes it difficult for a predator to fix upon and attack any one.

Behaviour that aids the function of coloration

Some forms of disruptive coloration also function to conceal movement. Forward movement in concentrically



Flamingo *(Phoenicopterus ruber roseus)* showing carotenoid pigmentation in the plumage and leg skin.



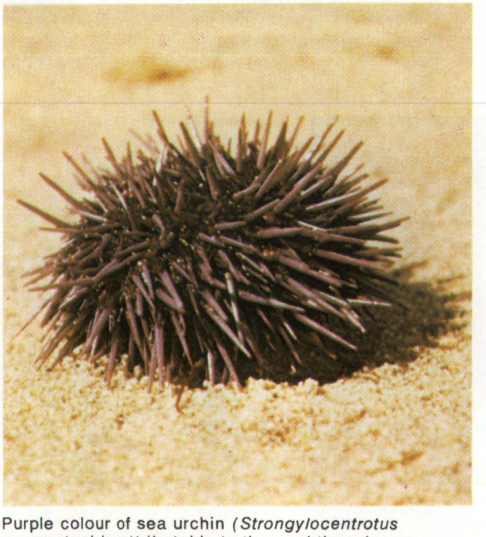
Structural colour of the superb tanager *(Tangara fastuosa)* is tyndall blue.



Red colour in wing of white-crested turacos *(Tauraco leucolophus)* results from turacin, a pigment derived from porphyrin.



The silvery appearance of the butterfly fish *(Chaetodon)* is due to a deposit of guanine, a colourless purine.



Purple colour of sea urchin *(Strongylocentrotus purpuratus)* is attributable to the naphthoquinone compound, echinochrome.



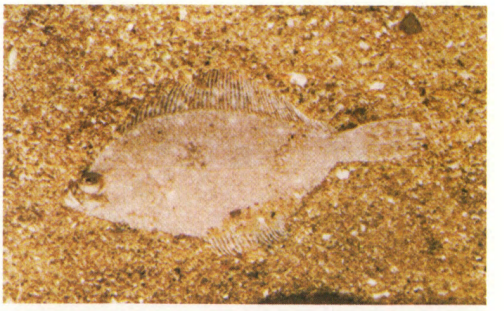
Rivoli's hummingbird *(Eugenes fulgens)* has iridescent structural colour.

## Pigmentation

Plate 1: (Top left) C. Laubscher—Bruce Coleman Inc., (top right) Eric Hosking, (centre right, bottom right) Root Resources—EB Inc., (centre right) Kenneth Fink, (bottom right) Earl Kubis, (bottom left) Bruce Barnetson from The Natural History Photographic Agency—EB Inc., (bottom centre) Cy LaTour



Zebras *(Equus burchelli)* at a waterhole, an example of coloration disruption.

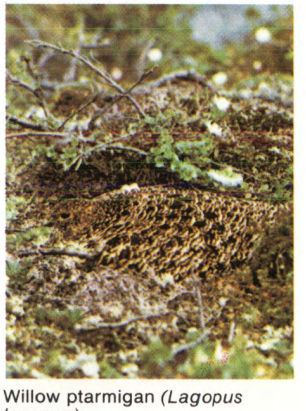




Background matching by two pleuronectiform fishes, (above) right-eyed flounder and (below) flatfish on sandy bottom floor.



European woodcock *(Scolopax rusticola)* incubating.



Willow ptarmigan *(Lagopus lagopus).*



Blacksmith plover *(Vanellus armatus)* showing disruptive markings.



The disruptive markings of the moorish idol *(Zanclus canescens).*



Uganda kobs *(Kobus kob thomasi)* exemplify countershading.

**Concealing and disruptive coloration**

Plate 2: (Top left) Gerald Cubitt, (top right above) J.A.L. Cooke, (top right below, centre left, centre, centre right) Bruce Coleman Inc., (top right below) Jen & Des Bartlett, (centre left) E. Soothill, (centre) Eric Hosking, (centre right) Tony Deane, (bottom left) Douglas Faulkner, (bottom right) Leonard Lee Rue III



Twig-like measuring worm (Geometridae).



Thorn treehopper *(Umbonia spinosa)*.



Walkingstick (Phasmatidae).



Swallowtail butterfly *(Papilio)* concealed in leaves.



Orchid mantis *(Hymenopus coronatus)* of Malay Peninsula.



Leaf butterfly *(Anaea)* on jungle leaf litter.



Ceylon leaf insect *(Philium)*.

**Concealing coloration**

Plate 3: (Top left, top right, centre right, bottom left) E.S. Ross, (centre left) Lilyan Simmons—EB Inc., (centre) F. Baillie from The Natural History Photographic Agency—EB Inc., (bottom right) S.C. Bisserot—Bruce Coleman Ltd.





Warning coloration: (above) the puss moth caterpillar *(Cerura)* at rest; when threatened (below), its raised head and tail thorns intimidate attackers.



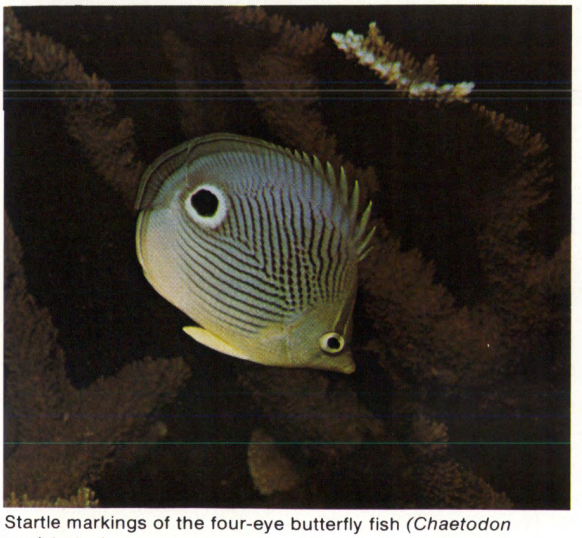
Startle markings: false eyes on noctuid moth (Noctuidae).



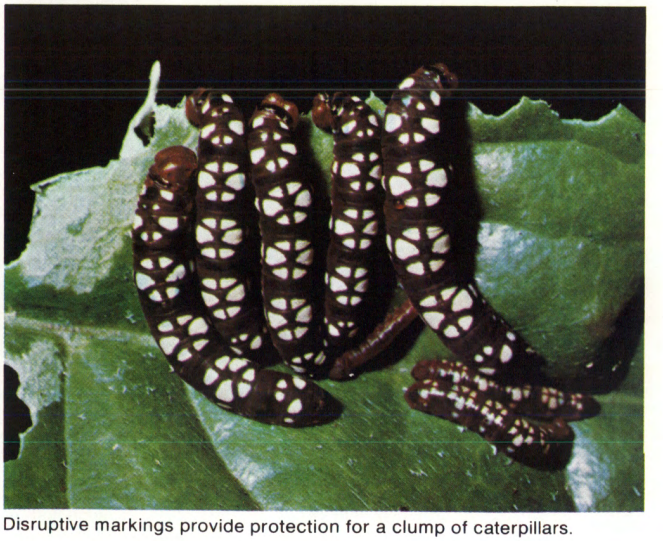
Startle markings of a mature swallowtail butterfly larva *(Papilio)*.



Shadow pattern: mottling of the fawn's coat *(Odocoileus virginianus)* provides protection.



Startle markings of the four-eye butterfly fish *(Chaetodon capistratus)*.



Disruptive markings provide protection for a clump of caterpillars.

**Warning coloration and protective and startle markings**

Plate 4: (Top left, centre left) S. Beaufoy, (top centre, top right, bottom right) E.S. Ross, (centre right) Leonard Lee Rue III, (bottom left) Douglas Faulkner



Alluring coloration: potential predators of the blue-tailed skink *(Eumeces skiltonianus)* are attracted to its tail, which can be shed at will.



Keel-billed toucan *(Ramphastos sulfuratus);* the bill is probably used for species recognition.



Flash colours: male great frigate bird *(Fregata minor)* with red throat patch inflated to attract a female.



Display coloration of cock of the rock *(Rupicola rupicola).*



Courtship coloration of a mammal: male mandrill *(Mandrillus sphynx).*

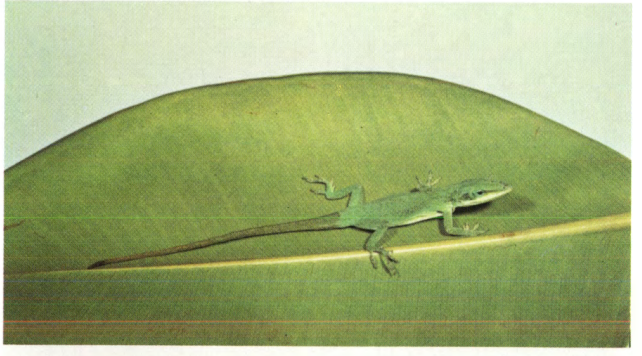
**Courtship and distraction coloration**

Plate 5: (Top left) E.S. Ross, (top right) © Kojo Tanaka, Animals Animals 1972, (centre) Verna R. Johnston, (bottom left) K.W. Fink—Bruce Coleman Inc., (bottom right) George H. Harrison from Grant Heilman—EB Inc.

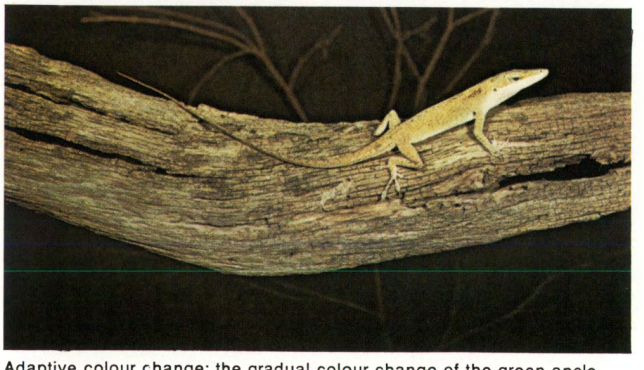




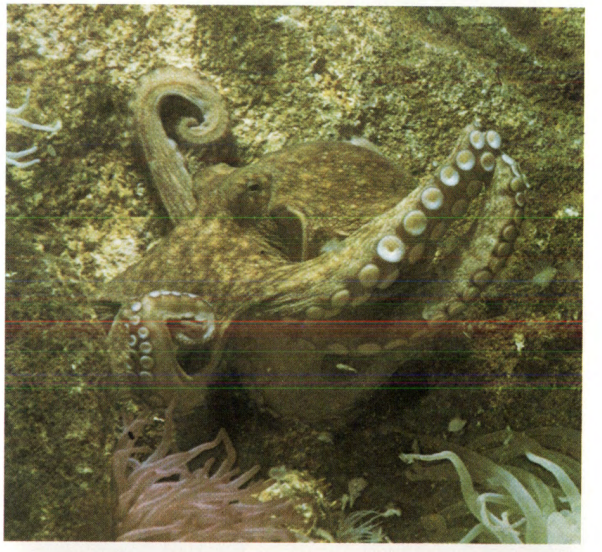
Seasonal colour change in the varying hare *(Lepus)*; (left) summer pelage and (right) winter pelage.







Adaptive colour change: the gradual colour change of the green anole *(Anolis carolinensis)* when moved from a green leaf to a brown branch.





Adaptive colour change: (top) at rest the common octopus *(Octopus vulgaris)* blends with its surroundings; when agitated (bottom), it blanches.

**Colour change**

Plate 6: (Top left) Rue—Annan Photo Features, (top right) Leonard Lee Rue III—Bruce Coleman Inc., (centre left above, centre left below, bottom left) John H. Gerard—EB Inc., (centre right, bottom right) Marineland of Florida

banded snakes, for example, is difficult to perceive when the animal moves between reeds or long blades of grass.

*Countershading.* Another clue can lead to the recognition of an organism: its three-dimensional form, which causes the unilluminated portion of the body to be in shadow. Countershading is a form of coloration in which the upper surfaces of the body are more darkly pigmented than the unilluminated lower areas, giving the body a more uniform darkness and a lack of depth relief. Widespread among vertebrates, countershading is frequently superimposed over camouflage and disruptive colorations.

Countershading by luminescent organs

The light-producing organs, or photophores, of many deepwater fishes provide a unique form of countershading. Photophores occur in bands along the lower parts of the sides and are directed downward. Deepwater fishes live in the twilight zone of the sea, in which the illumination is too weak to allow little more than a silhouette of prospective prey sighted by a predator from below. The downward-projecting photophores may provide countershading by obliterating the silhouette when it is viewed by a predator from below.

*The role of shape in relation to coloration.* The shape of an organism is important in determining the total configuration for protection. Both concealment and mimicry may depend strongly on imitation of both the shape and coloration of the model. Deep-bodied schooling fishes frequently show vertical banding, and elongated forms usually bear horizontal stripes. This dichotomy may be partially related to different swimming patterns: deep-bodied fishes perform frequent lateral turns; elongated forms show frequent horizontal movement and change of position.

**Mimicry.** As mentioned above, deception may be accomplished by providing false information through mimicry. Aggressive mimicry occurs when a predator resembles its prey or a harmless third party. For example, the American zone-tailed hawk (*Buteo albonotatus*) is nearly black and has long narrow wings, and it glides in the company of similarly coloured and shaped vultures. The vultures do not prey on small animals and therefore do not cause fright reactions among them. The zone-tailed hawk exploits this lack of fear by suddenly diving on its prey from among the group of circling vultures.

Some organisms provide false information as to their identity by mimicking dangerous or inedible species. When a third party, such as a predator, fails to distinguish between the mimic and its inedible model, the relationship is termed Batesian mimicry (see MIMICRY). Batesian mimicry can be contrasted to those forms of camouflage in which organisms show an "imitative resemblance" to inanimate objects in their environment, such as the leaves or twigs of a tree. Imitative resemblance is a true concealing coloration in that it usually disguises the organism sufficiently so it is not perceived as distinct from its background. The form and coloration of a Batesian mimic, on the other hand, usually ensures that the organism will be perceived by animals, including predators, but that it will be identified with the harmful or distasteful model species. Batesian mimicry is thus both a deceptive and an advertising coloration, and it is effective only because the model species itself has a warning coloration (see below).

Batesian mimicry

## OPTICAL FUNCTIONS: ADVERTISING COLORATION

Whereas concealing coloration reduces visual information, advertising coloration provides easily perceived information as to an organism's location, identity, and movement.

**Attraction.** The most commonly recognized forms of advertisement occur as intraspecific communication. Most important in such interactions are the organism, or signal sender, and the third party, or signal receiver; also important are illumination and the relationship between the organism and its background.

*Reproductive signals.* Courtship colorations function to attract and arouse a mate and to aid in the reproductive isolation of species. Although by no means universal, it is common, at least among vertebrates, to find that the male of the species has the brightest courtship colours. Bright colours are usually accompanied by movements and display postures that further enhance the display coloration. In some species a number of males form a communal display group in active competition for females. Examples among birds include manakins (Pipridae), cocks of the rock (*Rupicola*), and some grouse (Tetraonidae); similar communal displays occur in some giant species of fruit flies (*Drosophila*) found in the mountains of Hawaii. The male flies hold their variously adorned wings outstretched and perform a series of visual displays toward females.

Communal courtship systems

To be maximally efficient, courtship coloration should either be shown only by sexually ripe individuals or be unique to the individuals that are courting. In many birds this is accomplished by spreading coloured feathers that are otherwise largely concealed. In others, however, the coloration serves multiple functions or is present throughout the year, and courting individuals are rendered unique by other displays, perhaps of a visual or auditory nature. Many fishes show dramatic changes in coloration during courtship. In some species these changes are long-term, hormonally mediated alterations of coloration and frequently include a proliferation of the carotenoid (red and yellow) pigments. Other coloration changes in courting fishes are short-term alterations involving melanophores, which cause rapid colour changes. As a female approaches the male, his sexual arousal can be measured by the degree of coloration change. Luminescence is involved in courtship signals in a variety of animals; for example, different species of the common firefly (Lampyridae) show unique flashing codes.

Courtship flashing codes

*Schooling signals.* In gregarious animals, coloration, morphology, and general behaviour may identify an individual to others of its species and can aid in the formation of species aggregations throughout the year. This is seen in schooling fishes, in which the portion of the body moved by swimming motions frequently contrasts with the coloration of the rest of the body, apparently providing an attracting stimulus within the school.

*Interspecific signals.* Species that enter into symbiotic, or mutualistic, interactions may be brought together by advertising coloration. Many plants depend on insects and even certain birds and bats for pollination and the dispersal of seeds. The pollinator is attracted first to the flower of the plant from which it picks up pollen while feeding; then it visits another flower of the same species, transferring some of the pollen. The coloration and shape of the flowers attract the pollinators and provide information as to the species of the plant. The flowers of plants pollinated by insects usually have patterns of yellow, blue, and ultraviolet that evoke a strong response in the insect eye. They usually have a darkly coloured pattern near the centre of the flower, called the nectar guide, which orients the insect toward the proper pollinating location. Bees show a strong preference for flowers with intricate shapes and colorations. Intricate radial patterns seem to be the most attractive; in fact, bees cannot be trained to prefer a simple to an intricate pattern. Some orchids take advantage of the sexual behaviour of bees, the flowers being nearly perfect mimics of the female bees. A male bee attempting to copulate with the flower acquires the pollen capsules and transfers them to another flower (see MIMICRY).

The attraction of different species

**Repulsion.** *Territorial advertising.* During the reproductive season, many animals defend a particular area or territory that includes their nest or spawning site. Many other animals defend territories throughout the year. In either case, coloration is frequently important. In species in which the task of territorial defense is accomplished largely by one sex, strong sexual dimorphism usually exists, the more brightly coloured sex being the one that holds the territory. Both male- and female-territorial species are found within the diverse fish family Cichlidae. Species in which the male holds a territory are marked by large and colourful males, the females being smaller and camouflaged; in those species in which the female defends the territory the reverse is found. In still other species the fish pair and share the territory, and there is little sexual dimorphism.

Coloration frequently releases agonistic (flight or attack) behaviour in territorial animals and intimidates intruders. The flashing coloration displays of a dominant octopus are an excellent example of a visual battle in which the victor may be determined with little or no bodily contact.

Use of coloration to avoid fighting

Although similar advertising colorations may contribute to the spacing out of territorial animals, dissimilarity in coloration between members of a species may allow closer spacing. Many brightly coloured reef fishes, for example, defend territories or personal spaces. In many of these species the young and subadults, with radically different coloration from the adult, live within the territory of an adult but remain free from attack; after they assume adult coloration, however, they are driven away. The territories frequently function to ensure a food supply; because the juveniles utilize different food, they pose no threat to the adult's supply. As the juveniles age, their feeding habits overlap those of the adult, and spacing is necessary.

*Warning, or aposematic, coloration.* Certain advertising colorations warn a third party of dangerous or inedible qualities of the organism (aposematic colorations), such as spines, poisons, or other defensive weapons, allowing the possessor to avoid a potentially damaging interaction in which the weapon is used. Red, black, and yellow are common in this context and may represent aposematic colours recognized by many animals.

The advantage of warning

As discussed above, Batesian mimicry is the imitation of aposematic coloration by benign organisms, which thereby enjoy at least a portion of the protection of the model species. While Batesian mimicry involves deceptive coloration, resemblance in warning coloration need not provide false information. Müllerian mimicry refers to instances in which several noxious species display the same warning coloration, thus enabling potential predators to learn and generalize the signal easily. The black-and-yellow coloration of bees and wasps is a typical example.

### OPTICAL FUNCTIONS: COMBINATION OF CONCEALING AND ADVERTISING COLORATION

Most animals need both concealment and advertisement. An animal may need to conceal itself from predators and to advertise its presence to symbionts or to members of its own species for reproductive purposes.

Many birds that conceal courtship coloration when their feathers are held close to the body present a brilliant display upon erecting their feathers. Similar mechanisms are common in many animals, such as *Anolis* lizards, which have brightly coloured throat fans that are visible only when erected during courtship or threat behaviour.

Substitution of coloured objects for bright plumage

Many bower birds (Ptilonorhynchidae) have bright courtship colorations, although some males of *Amblyornis* species do not. Instead, they decorate an elaborate bower with leaves, flower petals, and other brightly coloured objects, which attract females but provide no clue to predators as to the exact location of the male.

Some predators deceive with advertising coloration. The frogfishes, or shallow-water anglerfishes, are extremely difficult to detect against their background. They have intricate and obvious lures that are waved near the mouth on a long stalk; prey fishes attracted to the lure are eaten.

Coloration change is another obvious mechanism that can restrict advertisement to times when it is needed for purposes of communication. Many animals change from cryptic to noncryptic colorations as they change from their normal resting coloration to a display coloration during social interactions. These changes are particularly common in fishes and cephalopods, which have efficient neural mechanisms of coloration change.

### OPTICAL FUNCTIONS: THE ROLES OF THE SELECTIVE AGENT AND OF ILLUMINATION

**The selective agent.** Of obvious importance in the evolution of coloration is the third party, which is the actual selective agent involved in the relationship between the organism and its background. Identification of the third party and the sensory and nervous system components used by it are important in order to understand thoroughly the adaptive nature of deceptive or advertising coloration.

In analyzing concealing coloration, the actual identification of the third party may have a profound influence on the interpretation of the coloration and behaviour. For example, the early stages of the green Scotch pine caterpillar (*Bupalus piniarius* and others) are found at the tips of pine needles, well camouflaged in this position. As they grow larger, they move into the bases of the needles and onto the branch. One explanation for the movement is that the older caterpillars are much larger than the background needle, thus rendering the camouflage less effective. Another factor appears to be a shift in the third party as the caterpillar ages; young caterpillars are preyed upon by spiders found on the twigs; larger caterpillars, by birds such as titmice (*Parus*).

Importance of the visual capabilities of the third party

After the initial identification of the third party, its visual capabilities must be investigated. The spectral sensitivity of its eyes must be determined, as must the way in which it perceives combinations of biochromes and their arrangements. The visual stimulus is subject to encoding and integrating steps as it passes from the eye to the cerebral cortex of the brain. Contrast and movement are amplified by some cells, while other properties, such as shape and intensity, are ignored. In humans, for example, contrast is greatly enhanced at the junction between a red and a blue stripe, producing the optical illusion that the two stripes never meet and are on different planes. Such phenomena may be of importance in disruptive coloration.

Advertisement is likewise subject to the visual capabilities of the third party, or signal receiver. Many species of plants have yellow flowers barely distinguishable to the human eye; when an ultraviolet camera is used to photograph such flowers, however, various bright patterns and nectar guides are revealed that appear to be highly species specific. The importance of strong contrast and contour in the attraction of insects to flowers is related to the perceptual qualities of the insect's compound eye, which shows maximal response to flickering stimuli and may depend upon similar qualities for much form discrimination.

In social signals, the visual system of a species is frequently maximally responsive to its own range of colorations. Butterflies of the genus *Dardanus,* for example, are maximally responsive to their own blue courtship coloration. The visual system and coloration are coadapted to provide an efficient signal mechanism. (G.S.Lo./E.H.B.)

**Illumination.** Most optical signals depend on sunlight reflected from the animal or plant. Therefore, the receiver's perception of the signal depends on the characteristics of the ambient illumination, which, in turn, depends on such variables as time of year, time of day, amount of cloud cover, amount of vegetation between the light source and the optical signal, and spectral reflectance of the habitat. Clear-sky sunlight with the Sun overhead is essentially white, but with the Sun low in the sky the light has a yellow or orange spectral emphasis. Light in broadleaf forests has a yellow-green emphasis, whereas light in conifer forests has a slight bluish emphasis. These small but consistent differences may affect the evolution of optical coloration.

### VISUAL FUNCTIONS

Biological coloration can play a variety of roles in an animal's visual system. For example, facial coloration can help determine the amount of light that is reflected into the eyes. Among animals living in brightly lit habitats, too much reflected light could have undesirable effects on vision. It could, for example, produce blinding glare or dazzle; it might result in high luminance in parts of the visual field, thereby diminishing contrast in other parts of the field; or it could cause adaptation to a higher illuminance level than is appropriate for the remainder of the visual field. Birds that forage in sunlight for aerial insects—a visually demanding task—have bills that are black. Apparently the black coloration reduces reflectance that interferes with their vision.

The role of biochromes in vision

Vision itself depends on a biochrome that consists of a protein, opsin, attached to a chromophore. The chromophore may be either retinal (vitamin $A_1$), in which case the molecule is called rhodopsin; or 3-dehydroretinal (vitamin $A_2$), in which case the molecule is called porphyropsin. When light enters the eye and strikes the visual biochrome, the molecule undergoes a chemical change that stimulates the receptor nerve and thereby produces a visual stimulus.

In addition to the visual pigments, the eyes of many invertebrates contain biochromes that affect the spectrum of light that reaches the photoreceptors. Similarly, oil

droplets in the retina and epithelium of vertebrate eyes contain carotenoids that may affect colour perception. More importantly, the epithelium contains melanin, which absorbs stray light that penetrates the retina without being absorbed by the visual pigments. In insect eyes a similar function is performed by ommochromes in secondary pigment cells surrounding the photoreceptors.

Among many nocturnal vertebrates the white compound guanine is found in the epithelium or retina of the eye. This provides a mirrorlike surface, the tapetum lucidum, which reflects light outward and thereby allows a second chance for its absorption by visual pigments at very low light intensities. Tapeta lucida produce the familiar eyeshine of nocturnal animals.

### PHYSIOLOGICAL FUNCTIONS

The discussion of biochromes earlier in this article touched upon the many important physiological roles of biological pigments, including that of the chlorophylls in photosynthesis and of the hemoglobins in oxygen transport. This section provides examples of other physiological effects of biological coloration.

Hair and feathers that contain melanin are more durable than those that lack this biochrome. Increased durability probably accounts for the dark, melanic wing tips of most birds. It may also be a contributing factor to the high proportion of black among birds that live in deserts, which are exceptionally abrasive habitats.

Absorption of solar energy

The absorption of solar energy by dark skin, scales, feathers, or hair is often associated with increased heat gain and reduced metabolic rates. Because birds lose a large amount of body heat through their uninsulated legs, dark leg coloration may help to warm the legs by absorbing solar energy, thereby reducing heat loss. Such reduced heat loss may explain why dark-legged North American woodwarblers (Parulidae) arrive in their northern breeding areas earlier than light-legged woodwarblers. Dark feathers, however, may actually reduce the amount of solar energy that penetrates to and is absorbed by a bird's skin. With fully erect plumage in moderate winds, a dark bird in full sunlight absorbs less heat into its body than a light bird does. This may also be a factor contributing to the high proportion of black among desert-dwelling birds.

Photoactivation of 7-dehydrocholesterol into vitamin D occurs throughout the epidermis of humans in the presence of ultraviolet light. The melanization of human skin may be an adaptation to optimize synthesis of vitamin D by permitting more or less ultraviolet radiation to penetrate the epidermis.

A widespread response to increased light levels is the addition of melanin, or darkening of the body—for example, tanning in humans. Such melanic shielding protects the tissues of the organism from potentially dangerous levels of ultraviolet radiation. Since the ultraviolet shield need protect only easily damaged cells in the nervous and reproductive systems, it does not necessarily have to lie in the skin but can instead be located internally, immediately around sensitive organs. When the ultraviolet shield is internal, external coloration may conform to other selection pressures.

Water is conserved by reducing evaporative loss and by reducing excretory water loss. Insects reduce evaporative water loss by adding melanin to the cuticle, melanin being more waterproof than other biochromes. The black-coloured beetle *Onymacris laeviceps* loses significantly less water than does the white-coloured beetle *O. brincki* when both species are kept without food under identical conditions. Quinones also darken insect exoskeletons, and in *Drosophila* quinones contribute to the low permeability of the exoskeleton. Some insects avoid excretory water loss by depositing nitrogenous wastes in the exoskeleton, which is shed periodically. In these species external coloration is a consequence of nitrogen excretion.

Some arthropods produce offensive odours as a means of defense against predators. These odours derive from *p*-benzoquinones in the exoskeleton and are correlated with the chromatic properties of the molecules. Consequently, coloration in these species may be a consequence of selection for chemical defense.

## Coloration changes

### COLORATION CHANGES IN INDIVIDUAL ORGANISMS

**Short-term changes.** Most rapid colour changes are chromatophoric ones that alter the colour of the organism through the dispersion or concentration of biochromes. Emotion plays a role in such changes among some cephalopods, fishes, and horned lizards (*Phrynosoma*). When excited, certain fishes and horned lizards undergo a transient blanching that probably results from the secretion of adrenaline (epinephrine), a hormone known to concentrate the dark biochrome of vertebrates. Excited cephalopods exhibit spectacular displays of colour, with waves of colour rippling across the body. Chromatophoric colour change is slower in vertebrates than in cephalopods. Although some fish may complete a colour change within a minute (compared to half a second or less for cephalopods), most vertebrates require several minutes to several hours.

Colour changes linked to external cycles

Colour changes extending over several hours are often entrained to external cycles. Fiddler crabs (*Uca*) that live in the intertidal zone show a complex pattern of cyclic chromatophoric colour change that is entrained not only to the local tidal cycle but also to the lunar and solar cycles. So important is this cyclic colour change that the response is innate to every part of the integument. The legs of a fiddler crab can be removed and sustained for a few days in saline solution; during this time melanophores in the legs continue to disperse and concentrate their melanin according to the cycle at the time they were removed from the body. (E.H.B.)

Changes in colour that extend over periods of several months may involve the synthesis or destruction of chromatophores or biochromes. The quantities of deposited guanine in some fishes vary in proportion to the relative lightness in colour of the background upon which they are living. Greenfish, or opaleye (*Girella nigricans*), kept in white-walled aquariums became very pale during a four-month period, storing about four times the quantity of integumentary guanine as was recoverable from the skins of individuals living in black-walled aquariums but receiving the same kind and amounts of food and the same overhead illumination. (D.L.F./E.H.B.)

Some chromatophores respond directly to relevant environmental stimuli, independent of the nervous system. Such response occurs in the young of some fish and of the clawed frog (*Xenopus*); but in older individuals the nervous system, which is by this time fully developed, controls responsiveness. More typically the chromatophore response is mediated by the sensorimotor system from the start. The eye plays a major role in cephalopods and most vertebrates, particularly in animals capable of matching complex backgrounds, but the pineal organ (a light-sensitive organ on top of the brain) and a generalized dermal light sense may also mediate the chromatophore response.

**Seasonal changes.** Seasonal changes of fields and forests include the annual colour changes involving foliage, flowers, fruits, and seeds of plants. Many birds and mammals undergo seasonal molts, replacing their plumage or pelage with differently coloured feathers or hair. Winter whitening of the willow ptarmigan (*Lagopus lagopus*) and varying hare (*Lepus*) are examples of a shift in camouflage coincident with a change in the background coloration. Many songbirds adopt a bright, contrasting nuptial plumage during the breeding season, reverting to a drabber winter plumage during the postnuptial molt.

Winter whitening

Seasonal colour changes are usually regulated by light (mediated by the visual or pineal systems) or by temperature. Decreasing day lengths initiate whitening in the willow ptarmigan, whereas falling temperatures initiate whitening in the weasel (*Mustela erminea*). The spring molt of the varying hare is stimulated by the lengthening day, but the rate of molt depends on temperature. Seasonal changes in coloration may occur without a molt as a result of bleaching or wear, for example, the bleaching of human hair in the summer sun and birds that have bright colours based on carotenoids.

**Age-related changes.** Colour changes during the life of an individual are common. Graying hair is a familiar

badge of the elderly, both in humans and, to varying degrees, in other mammals. Among primate groups, particularly gorillas and chimpanzees, silver hairs indicate both age and dominance. Young birds of many species have a juvenile plumage that gives way to either an adult plumage in short-lived birds or a series of immature plumages in longer-lived species. Most gulls, for example, are deep gray or brown during their first year and become increasingly white thereafter. Changes of colour are also associated with age and size in many fish; for example, the blue parrot fish changes from a vertically barred pattern to all blue in association with increasing age and size.

(F.A.B./E.H.B.)

## COLORATION CHANGES IN POPULATIONS

Coloration changes occur not only in individuals but in populations as well. These latter result from evolutionary pressures—*i.e.,* agents of natural selection—that act upon the natural variations in colour types (morphs) found among the population. As a result of such pressures, certain colour morphs have increased odds of surviving and passing on their coloration pattern. Depending on the nature of the selection pressures, the population may come to include substantial numbers of individuals of different colour morphs; or one morph may become dominant, largely supplanting an earlier dominant colour form.

Polychromatism

When individual colour variation is discontinuous within a species, that species is said to be polychromatic. The white-throated sparrow (*Zonotrichia albicollis*) of North America, for example, has individuals with white-and-black head stripes and other individuals with tan-and-brown head stripes. The different colorations are not associated with age, sex, or geographic region. Polychromatism may evolve in response to predation. A predator that successfully takes one prey type may then concentrate its search on others of this type and hence may overlook differently coloured prey of the same species. The phenomenon—known as a perceptual set or a search image—is exemplified by the predator of the European snail *Cepaea.* Predators encounter one morph and form a search image; they continue to hunt for that one form until its increasing rarity causes the predator to hunt randomly, encounter a different morph, and form a new search image. In this way, oscillating selection pressures maintain several contemporaneous colour morphs among the snail population.

Evolutionary colour changes dictated by shifting selection are suggested by many populations that show geographical or temporal clines (graded series of morphological characters). For example, the common flicker (*Colaptes auratus*) has yellow markings in eastern North America and red markings in western North America, suggesting a change in selection pressure as one moves from east to west. The best documented temporal shift in selection is the industrial melanism of noctuid and geometrid moths in England and Europe. The proportion of melanic, or darkly coloured, individuals in about 70 species of moths has increased dramatically since the 1850s. This increase correlates with the Industrial Revolution and the associated pollution of the countryside. Prior to that time, tree trunks, the normal daytime resting place of these nocturnal moths, had been covered by scattered whitish lichens. The trunks have turned dark in areas of industrial development because the lichens have been killed by pollutants and the trunks have been dirtied by soot. Blotched gray moths, previously protected from predation by birds, are now vulnerable, while the dark moths are less conspicuous. The shift to melanic populations in the United States lagged behind that in England and Europe, as did the industrialization process; but in Michigan by the early 1970s darkly coloured individuals formed up to 97 percent of some populations in regions where melanism was unknown before 1927. Since the 1970s in England there has been a reversal in the number of melanic individuals of some species, a sign that efforts to curb air pollution are having far-reaching effects.

Industrial melanism in moths

The many diverse functions discussed above lead to the inevitable conclusion that no single function can explain the coloration of living things. While biologists are far from a comprehensive theory that predicts the hues and patterns of coloration of plants and animals, such a theory will have to integrate the optical, visual, and physiological functions of biological coloration. (G.S.Lo./E.H.B.)

### BIBLIOGRAPHY

*Structural and biochemical bases for colour:* DENIS L. FOX, *Animal Biochromes and Structural Colours,* 2nd ed. (1976); and H. MUNRO FOX and GWYNE VEVERS, *The Nature of Animal Colours* (1960), are technical but readable works on pigments and schemochromes; ARTHUR E. NEEDHAM, *The Significance of Zoochromes* (1974), is a technical analysis of the chemistry, control, and function of biochromes. DENIS L. FOX, *Biochromy: Natural Coloration of Living Things* (1979), is a study of chemical and physical aspects of the coloration of flora and fauna; and J.N. LYTHGOE, *The Ecology of Vision* (1979), is a summary of research in the influence of colour chemistry on the life of marine organisms. ERSTON V. MILLER, *The Chemistry of Plants* (1957); T.J. MABRY, K.R. MARKHAM, and M.B. THOMAS, *The Systematic Identification of Flavonoids* (1970); T.W. GOODWIN (ed.), *Chemistry and Biochemistry of Plant Pigments,* 2nd ed., 2 vol. (1976); and THEODORE A. GEISSMAN, "Anthocyanins, Chalcones, Aurones, Flavones and Related Water-Soluble Plant Pigments," in KARL PAECH and M.V. TRACEY (eds.), *Modern Methods of Plant Analysis,* vol. 3 (1955), are technical dissertations on plant pigments. See also THEODORE A. GEISSMAN, *The Chemistry of Flavonoid Compounds* (1962). F. BLANK, "Anthocyanins, Flavones, Xanthones," in WILHELM RUHLAND (ed.), *Encyclopedia of Plant Physiology,* vol. 10 (1958), provides insight into the formative processes of plant pigments. SYLVIA FRANK, "Carotenoids," *Scientific American,* 194:80–86 (1956); and SARAH CLEVENGER, "Flower Pigments," *Scientific American,* 210:84–92 (1964), are well-illustrated articles for the lay reader. See also OTTO ISLER, HUGO GUTMANN, and ULRICH SOLMS (eds.), *Carotenoids* (1971); and JOHN PROCTOR and SUSAN PROCTOR, *Color in Plants and Flowers* (1978).

*Control of coloration:* M. FINGERMAN, *The Control of Chromatophores* (1963), is a good, readable account of the knowledge of physiological colour change. "Chromatophores and Color Changes," *American Zoologist,* 23(3):461–592 (1983), is a symposium of papers on hormonal and neural control of colour change. FRANK B. SMITHE, *Naturalist's Color Guide* (1975), is a pocket-size, loose-leaf book containing 182 named colours. A.H. STURTEVANT, *A History of Genetics* (1965), is a vivid description of the beginnings and development of classical genetics. C. DONNELL TURNER and JOSEPH T. BAGNARA, *General Endocrinology,* 6th ed. (1976), contains a treatment of hormonal regulation of animal coloration, with a selected bibliography. See also JOSEPH T. BAGNARA and MAC E. HADLEY, *Chromatophores and Color Change: The Comparative Physiology of Animal Pigmentation* (1973); PAUL A. JOHNSGARD, *The Hummingbirds of North America* (1983), which explains the physics of changing plumage colour; and WILLYS K. SILVERS, *The Coat Colors of Mice: A Model for Mammalian Gene Action and Interaction* (1979), a study of the genetic phenomena of interaction of colour factors.

*The adaptive value of biological coloration:* HUGH B. COTT, *Adaptive Coloration in Animals* (1940, reprinted 1966), is a detailed and scholarly treatment; EDWARD H. BURTT, JR. (ed.), *The Behavioral Significance of Color* (1979), is a technical but readable treatment of nonoptical and optical functions of coloration, with discussion of visual psychology and the physics of light; EDWARD H. BURTT, JR., *An Analysis of Physical, Physiological, and Optical Aspects of Avian Coloration with Emphasis on Wood-Warblers* (1986), looks at the evolution of colour and pattern in a single subfamily of birds; JACK P. HAILMAN, *Optical Signals: Animal Communication and Light* (1977), is a thought-provoking analysis of colour and behaviour as they affect optical signaling; BERNARD KETTLEWELL, *The Evolution of Melanism: The Study of a Recurring Necessity* (1973), presents an analysis of the role of colour in natural selection; SALLY FOY, *The Grand Design: Form and Colour in Animals* (1983), is a study of animal anatomy and morphology, with excellent illustrations; WILLIAM J. HAMILTON III, *Life's Color Code* (1973), is a popular account of some functions of coloration; and GERALD H. THAYER, *Concealing Coloration in the Animal Kingdom,* new ed. (1918), is a classic work on the theories of camouflage. Other works of interest for the general reader include MICHAEL FOGDEN and PATRICIA FOGDEN, *Animals and Their Colours: Camouflage, Warning Coloration, Courtship and Territorial Display, Mimicry* (1974), a comprehensive scientific treatment with excellent photographs; DENIS OWEN, *Camouflage and Mimicry* (1980), a popular account full of interesting anecdotes and photographs; and WOLFGANG WICKLER, *Mimicry in Plants and Animals* (1968; originally published in German, 1968), a thorough and reliable survey.

(D.L.F./F.A.B./G.S.Lo./E.H.B.)

# Colour

Vision is obviously involved in the perception of colour. A person can see in dim light, however, without being able to distinguish colours. Only when more light is present do colours appear. Light of some critical intensity, therefore, is also necessary for colour perception. Finally, the manner in which the brain responds to visual stimuli must also be considered. The colour green has a quite different meaning for a resident of a tropical rain forest than it has for a desert dweller. Even under identical conditions, the same object may appear red to one observer and orange to another. Clearly, the perception of colour depends on vision, light, and individual interpretation, and an understanding of colour involves physiology, physics, and psychology.

An object appears coloured because of the way it interacts with light. The analysis of this interaction and the factors that determine it are the concerns of the physics of colour. The physiology of colour involves the eye's and the brain's responses to light and the sensory data they produce. The psychology of colour is invoked when the mind processes the visual data, compares them to information stored in memory, and interprets them as colour.

This article concentrates on the physics of colour. For a discussion of colour as a quality of light, see the articles LIGHT and ELECTROMAGNETIC RADIATION. For the physiological aspects of colour vision, see the article SENSORY RECEPTION. See also the article PAINTING, THE ART OF for a discussion of the psychological and aesthetic uses of colour.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* Part One, Division II, especially Section 128.

This article is divided into the following sections:


Colour and light 595
  The nature of colour 595
  The visible spectrum 596
  The laws of colour mixture 596
The measurement of colour 597
  Tristimulus measurement and chromaticity diagrams 597
  Colour atlases 598
Physical and chemical causes of colour 598
  Simple excitations, vibrations, and rotations 599
    Incandescence
    Gas excitation
    Vibrations and rotations
  Ligand fields 599
    Transition metal impurities
    Transition metal compounds
  Molecular orbitals 600
    Organic compounds
    Charge transfer
  Energy bands 600
    Metals
    Pure semiconductors
    Doped semiconductors
    Colour centres
  Geometrical and physical optics 601
    Dispersion and polarization
    Scattering
    Interference
    Diffraction
The perception of colour 602
  Colour effects 602
  Colour vision 603
The psychology of colour 603
Bibliography 604


## Colour and light

### THE NATURE OF COLOUR

Aristotle viewed all colour to be the product of a mixture of white and black, and this was the prevailing belief until 1666, when Sir Isaac Newton's prism experiments provided the scientific basis for the understanding of colour. Newton showed that a prism could break up white light into a range of colours, which he called the spectrum (see Figure 1), and that the recombination of these spectral colours re-created the white light. Although he recognized that the spectrum was continuous, Newton used the seven colour names red, orange, yellow, green, blue, indigo, and violet for segments of the spectrum by analogy with the seven notes of the musical scale.

From K. Nassau, *Physics and Chemistry of Color* (1983); John Wiley & Sons, Inc.

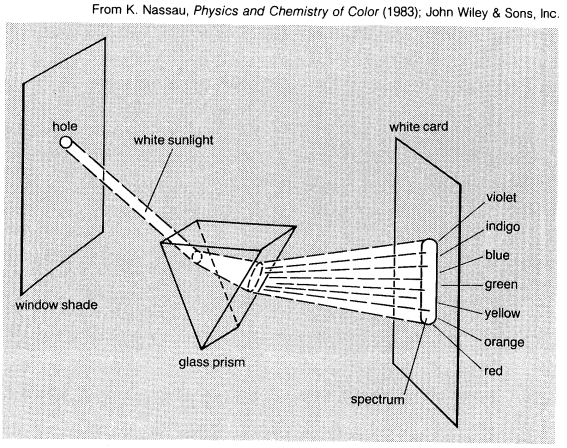



Figure 1: Newton's prism experiment of 1666.

Newton realized that colours other than those in the spectral sequence do exist, but noted that

> all the colours in the universe which are made by light, and depend not on the power of imagination, are either the colours of homogeneal lights [*i.e.,* spectral colours], or compounded of these.

Newton also recognized that

> rays, to speak properly, are not coloured. In them there is nothing else than a certain power . . . to stir up a sensation of this or that colour.

Nevertheless, terms such as "blue light" are commonly used and normally do not produce any confusion.

The unexpected difference between light perception and sound perception clarifies this curious aspect of colour. When beams of light of different colours, such as red and yellow, are projected together onto a white surface in equal amounts, the resulting perception of the eye signals a single colour (in this case, orange) to the brain, a signal that may be identical to that produced by a single beam of light. When, however, two musical tones are sounded simultaneously, the individual tones can still be easily discerned; the sound produced by a combination of tones is never identical to that of a single tone. A tone is the result of a specific sound wave, but a colour can be the result of a single light beam or a combination of any number of light beams.

Hue, saturation, and brightness

A colour can, however, be precisely specified by its hue, saturation, and brightness—three attributes sufficient to distinguish it from all other possible perceived colours. The hue is that aspect of colour usually associated with terms such as red, orange, yellow, and so forth. Saturation (also known as chroma, or tone) refers to relative purity. When

a pure, vivid, strong shade of red is mixed with a variable amount of white, weaker or paler reds are produced, each having the same hue but a different saturation. These paler colours are called unsaturated colours. Finally, light of any given combination of hue and saturation can have a variable brightness (also called intensity, or value), which depends on the total amount of light energy present.

### THE VISIBLE SPECTRUM

Newton demonstrated that colour is a quality of light. To understand colour, therefore, it is necessary to know something about light. As a form of electromagnetic radiation, light has properties in common with both waves and particles. It can be thought of as a stream of minute energy packets radiated at varying frequencies in a wave motion. Any given beam of light has specific values of frequency, wavelength, and energy associated with it. Frequency, which is the number of waves passing a point in a unit of time, is commonly expressed in units of hertz (1 Hz = 1 cycle/second). Wavelength is the distance between corresponding points of two consecutive waves and is often expressed in units of nanometres (1 nm = $10^{-9}$ metres). The energy of a light beam can be compared to that possessed by a small particle moving at the velocity of light, except that no particle having a rest mass could move at such a velocity. The name photon, used for the smallest quantity of light of any given wavelength, is meant to encompass this duality, including both the wave and particle characteristics inherent in wave mechanics and quantum theory. The energy of a photon is often expressed in units of electron volts; it is directly proportional to frequency and inversely proportional to wavelength.

Light is not the only type of electromagnetic radiation—it is, in fact, only a small segment of the total electromagnetic spectrum—but it is the one form the eye can perceive. Wavelengths of light range from about 400 nm at the violet end of the spectrum to 700 nm at the red end (see the Table). (The limits of the visible spectrum are not sharply defined but vary among individuals; there is some extended visibility for high-intensity light.) At shorter wavelengths the electromagnetic spectrum extends to the ultraviolet region and continues through X rays, gamma rays, and cosmic rays. Just beyond the red end of the spectrum are the longer wave infrared rays (which can be felt as heat), microwaves, and radio waves. Radiation of a single frequency is called monochromatic. When this frequency falls in the range of the visible spectrum, the colour perception produced is that of a saturated hue.

**The Range of the Visible Spectrum**

| colour* | wavelength (nm) | frequency (Hz × $10^{14}$) | energy (eV) |
|---|---|---|---|
| Red (limit) | 700 | 4.29 | 1.77 |
| Red | 650 | 4.62 | 1.91 |
| Orange | 600 | 5.00 | 2.06 |
| Yellow | 580 | 5.16 | 2.14 |
| Green | 550 | 5.45 | 2.25 |
| Cyan | 500 | 5.99 | 2.48 |
| Blue | 450 | 6.66 | 2.75 |
| Violet (limit) | 400 | 7.50 | 3.10 |

*Typical values only.

### THE LAWS OF COLOUR MIXTURE

Colours of the spectrum are called chromatic colours; there are also nonchromatic colours such as the browns, magentas, and pinks. The term achromatic colours is sometimes applied to the black-gray-white sequence. According to some estimates, the eye can distinguish some 10,000,000 colours, all of which derive from two types of light mixture: additive and subtractive. As the names imply, additive mixture involves the addition of spectral components and subtractive mixture concerns the subtraction or absorption of parts of the spectrum.

Additive colour mixture

Additive mixing occurs when beams of light are combined. The colour circle, first devised by Newton, is still widely used for purposes of colour design and is also useful when the qualitative behaviour of mixing beams of light is considered. Figure 2 shows a colour circle in which the spectral colours red, orange, yellow, green, cyan, indigo, and blue-violet are joined by the nonspectral colour magenta (a mixture of blue-violet and red light beams). White is at the centre and is produced by mixing light beams of approximately equal intensities of complementary colours (colours that are diametrically opposed on the colour circle), such as yellow and blue-violet, green and magenta, or cyan and red. Intermediate colours can be produced by mixing light beams, so mixing red and yellow gives orange, red and blue-violet gives magenta, and so on.

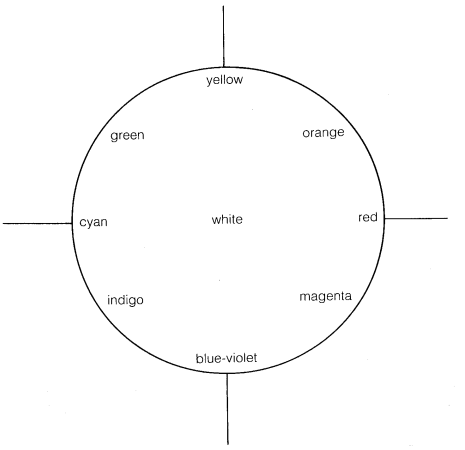



Figure 2: One form of Newton's colour circle.

The three additive primary colours are red, green, and blue; this means that by additively mixing the colours red, green, and blue in varying amounts almost all other colours can be produced, and when the three primaries are added together in equal amounts, white is produced.

Additive mixture can be demonstrated physically using three slide projectors fitted with filters so that one projector throws a beam of saturated red light onto a white screen, another a beam of saturated blue light, and the third a beam of saturated green light. Additive mixing occurs where the beams overlap (and thus are added together). Figure 3 (left) illustrates the results of such an experiment. As can be seen, where the red and green beams overlap, yellow is produced. If more red light is added or if the intensity of the green light is decreased, the light mixture becomes orange. Similarly, if there is more green light than red light, a yellow-green is produced.

Subtractive colour mixture

Subtractive colour mixing involves the absorption and selective transmission or reflection of light. It occurs when colorants (such as pigments or dyes) are mixed or when several coloured filters are inserted into a single beam of white light. For example, if a projector is fitted with a deep red filter, the filter transmits red light and absorbs other colours. If the projector is fitted with a strong green filter, red is absorbed and only green light is transmitted. If, therefore, the projector is fitted with both red and green filters, all colours are absorbed and no light is transmitted, resulting in black. Similarly, a yellow pigment absorbs blue and violet light while reflecting yellow, green, and red light (the green and red additively combining to produce more yellow). Blue pigment absorbs primarily yellow, orange, and red light. If the yellow and blue pigments are mixed,

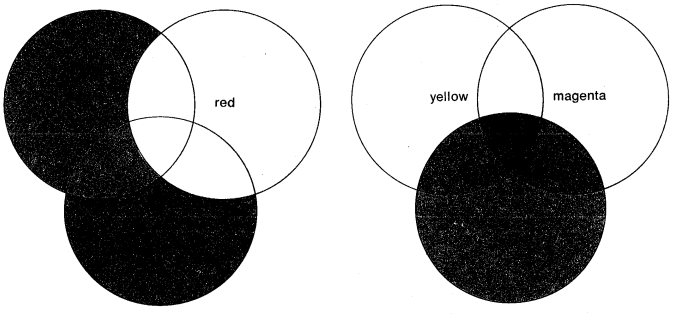



Figure 3: (Left) The additive mixing of red, green, and blue. (Right) The subtractive mixing of magenta, yellow, and cyan.

green is produced since it is the only spectral component that is not strongly absorbed by either pigment.

Because additive processes have the greatest gamut when the primaries are red, green, and blue, it is reasonable to expect that the greatest gamut in subtractive processes will be achieved when the primaries are, respectively, red-absorbing, green-absorbing, and blue-absorbing. The colour of an image that absorbs red light while transmitting all other radiations is blue-green, often called cyan. An image that absorbs only green light transmits both blue light and red light, and its colour is magenta. The blue-absorbing image transmits only green light and red light, and its colour is yellow. Hence, the subtractive primaries are cyan, magenta, and yellow (see Figure 3, right).

No concepts in the field of colour have traditionally been more confused than those just discussed. This confusion can be traced to two prevalent misnomers: the subtractive primary cyan, which is properly a blue-green, is commonly called blue; and the subtractive primary magenta is commonly called red. In these terms, the subtractive primaries become red, yellow, and blue; and those whose experience is confined for the most part to subtractive mixtures have good cause to wonder why the physicist insists on regarding red, green, and blue as the primary colours. The confusion is at once resolved when it is realized that red, green, and blue are selected as additive primaries because they provide the greatest colour gamut in mixtures. For the same reason, the subtractive primaries are, respectively, red-absorbing (cyan), green-absorbing (magenta), and blue-absorbing (yellow).

## The measurement of colour

Colorimetry

The measurement of colour is known as colorimetry. A variety of instruments are used in this field. The most sophisticated, the spectrophotometers, analyze light in terms of the amount of energy present at each spectral wavelength. The emittance curves for the light sources shown in Figure 4 are typical spectrophotometer results, as is the reflectance curve of the paint pigment known as emerald green, shown in Figure 5.

It is difficult to describe the colour of a specific spectral energy distribution. Since the eye perceives only a single colour for any given energy distribution, it is necessary to express colour measurements in a perception-related way. Several systems exist and some are outlined below.

### TRISTIMULUS MEASUREMENT AND CHROMATICITY DIAGRAMS

The tristimulus system is based on visually matching a colour under standardized conditions against the three primary colours, red, green, and blue; the three results are expressed as $X$, $Y$, and $Z$, respectively, and are called tristimulus values. The tristimulus values of the emerald-green pigment of Figure 5 are $X = 22.7$, $Y = 39.1$, and $Z = 31.0$. These values specify not only colour but also visually perceived reflectance, since they are calculated in such a way that the $Y$ value equals a sample's reflectivity (39.1 percent in this example) when visually compared to a standard white surface by a standard (average) viewer under average daylight. The tristimulus values can also be used to determine the visually perceived dominant spectral wavelength (which is related to the hue) of a given sample; the dominant wavelength of the emerald-green pigment is 511.9 nm.

Such data can be graphically represented on a chromaticity diagram (see Figures 6 and 7). Standardized by the Commission Internationale d'Éclairage (CIE) in 1931, the chromaticity diagram is based on the values $x$, $y$, and $z$, where $x = X/(X+Y+Z)$, $y = Y/(X+Y+Z)$, and $z = Z/(X+Y+Z)$. Note that $x+y+z = 1$; thus if two values are known, the third can always be calculated and the $z$ value is usually omitted. The $x$ and $y$ values together constitute the chromaticity of a sample. Light and dark colours that have the same chromaticity (and are therefore plotted at the same point on the two-dimensional chromaticity diagram) are distinguished by their different $Y$ values (luminance or visually perceived brightness).

Luminance

When their $x$ and $y$ coefficients are plotted on a chromaticity diagram, the spectral colours from 400 nm to 700 nm follow a horseshoe-shaped curve; the nonspectral violet-red mixtures fall along the straight line joining the 400 nm point to the 700 nm point. All visible colours fall within the resulting closed curve, as shown in Figure 6. Points along the circumference correspond to saturated colours; pale unsaturated colours appear closer to the centre of the diagram. The achromatic point is the central point at $x = 1/3$, $y = 1/3$ (shown as $W$ in Figure 7), where visually perceived white is located (as well as the pure grays and black, which vary only in the magnitude of the luminance $Y$).

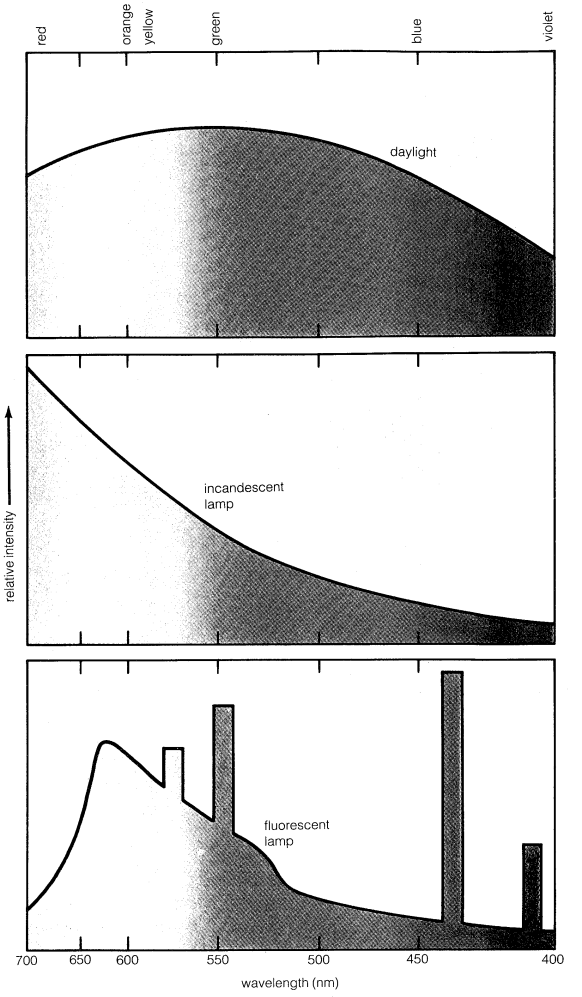



Figure 4: The energy distribution in light from daylight, an incandescent lamp, and a fluorescent lamp.
After K. Nassau, *Physics and Chemistry of Color* (1983); John Wiley & Sons, Inc.

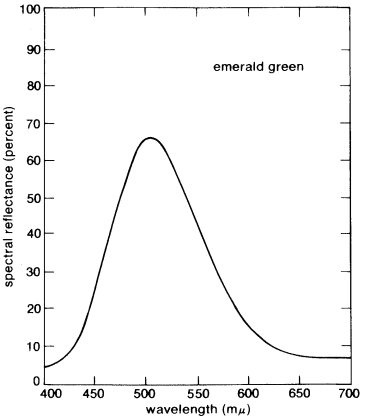



Figure 5: Spectral reflectance curve of an artist's pigment.

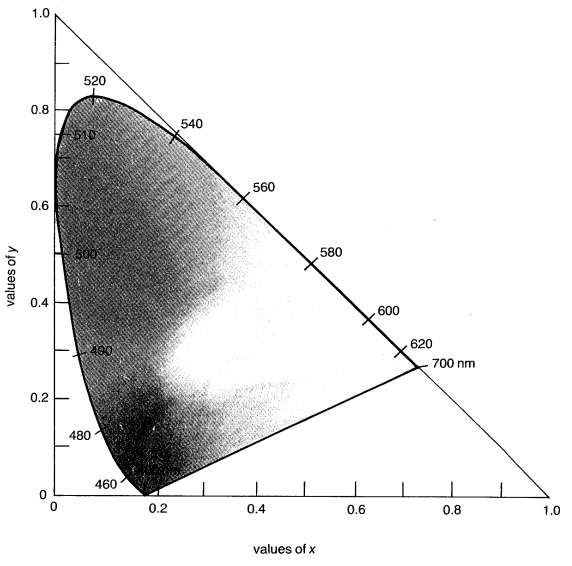



Figure 6: Standard chromaticity diagram.
By courtesy of RCA New Products Division, Lancaster, Pennsylvania

A straight line connecting any two points representing beams of light includes all the points representing colours formed by adding various amounts of the two beams. If the line passes through the achromatic point, the colours represented by its endpoints, when additively combined in the appropriate amounts, must form white; therefore all lines passing through the achromatic point terminate on the closed curve in saturated complementary colours.

Dominant wavelength and purity

By plotting the calculated $x = 0.245$ and $y = 0.421$ of the emerald-green pigment at point $E$ on the chromaticity diagram in Figure 7 and extending a line through it from the achromatic point $W$ to the saturated spectral boundary, it is possible to determine the dominant wavelength of the pigment colour, 511.9 nm. The colour of the pigment is the visual equivalent of adding white light and light of 511.9 nm in amounts proportional to the lengths $n$ (the distance between points $E$ and $W$) and $m$ (the distance between $E$ and the point of the dominant wavelength) in Figure 7. The purity equals $100n/(m + n)$ percent—in this case, 22.8 percent. A purity of 100 percent corresponds to a pure saturated spectral colour and 0 percent to the achromatic colours (white, gray, and black).

The colour of a specific red apple of $Y = 13.0$, $x = 0.460$, $y = 0.287$ has its $x$ and $y$ values plotted at $R$ in Figure 7. The line from the achromatic point $W$ intersects the chromaticity diagram boundary at a saturated nonspectral purple-red at $P$. The dominant colour designation is then obtained by extrapolating the line in the opposite direction to a saturated spectral colour and is given as "complementary dominant wavelength 495 nm" or 495$c$. The colour of this apple is therefore the visual equivalent of a mixture of white light and the 495$c$ saturated purple-red in the intensity ratio of the distances $p$ to $q$ shown in Figure 7 with a purity of $100p/(p + q)$ percent.

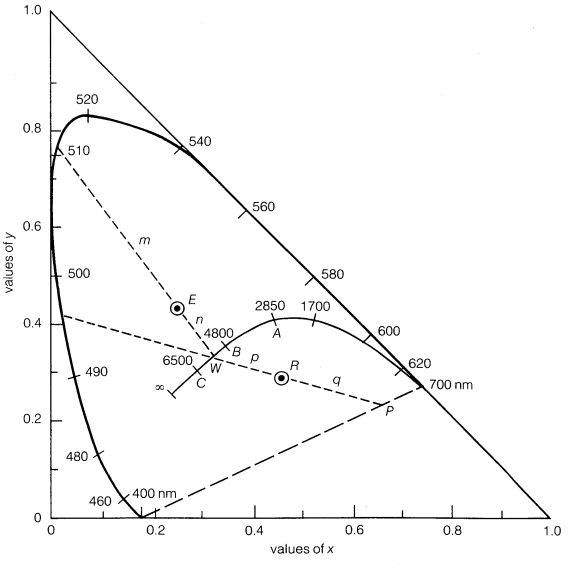



Figure 7: The location on the chromaticity diagram of the achromatic point *W*, the emerald-green pigment *E*, a red apple *R*, the incandescence curve with temperatures in kelvins, and the standard CIE illuminants *A*, *B*, and *C*.

Light from incandescent sources, further described below, falls on the solid curve marked with temperatures in this figure, following the sequence saturated red to saturated orange to unsaturated yellow to white to unsaturated bluish white for an infinite temperature indicated as ∞. The points *A*, *B*, and *C* on the curve are CIE standard illuminants that approximate, respectively, a 100-watt incandescent filament lamp at a colour temperature of about 2,850 K, noon sunlight (about 4,800 K), and average daylight (about 6,500 K).

### COLOUR ATLASES

Calculating chromaticity and luminance is a scientific method of determining a colour, but for the rapid visual determination of the colour of objects, a colour atlas such as the *Munsell Book of Color* is often used. In this system colours are matched to printed colour chips from a three-dimensional colour solid whose parameters are hue, value (corresponding to reflectance), and chroma (corresponding to purity). These three parameters are illustrated schematically in Figure 8. The central vertical axis provides a 10-step value scale extending from black at the bottom to white at the top. There are 100 hues divided into 10 groups around the vertical axis; each group has a colour name and consists of 10 subdivisions assigned a number from 1 to 10. The chroma scale starts at 0 at the vertical axis and extends radially outward from 10 to 18 steps depending on hue and value. The red apple discussed earlier would be designated 10RP 4/10 in the Munsell system, indicating a specific reddish purple hue 10RP, a value of 4, and a chroma of 10. Interpolated values are used to give more precise designations, so the emerald-green pigment can be specified as 5.0G 6.7/11.2.

Munsell system

A system that is useful when such precision is not required is the ISCC–NBS (Inter-Society Color Council–National Bureau of Standards) Centroid Color Charts. It has 267 numbered colour designations and uses descriptive terms such as very pale purple, light yellowish brown, and grayish blue; the red apple is 258 (moderate purplish red) in this system, the emerald-green pigment is 139 (vivid green). Other colour atlases include the Ostwald colour system, based on mixtures of white, black, and a high chroma colour; the Maerz and Paul dictionary of colour; the Plochere colour system; and the Ridgway colour standards.

## Physical and chemical causes of colour

According to the law of energy conservation, energy can be converted from one form to another, but it cannot be created or destroyed. Consequently, when a photon of light is absorbed by matter, usually by an atom, molecule, or ion, or a small grouping of such units, the photon disappears and its energy is gained by the matter. Similarly, when matter emits light, it loses the energy carried away by the photons. A given atom or molecule cannot emit light of any arbitrary energy, since quantum theory explains that only certain energy states are possible for a given system.

An example of permitted energy levels is shown at the left in Figure 9 for the trivalent chromium ion present in a crystal of aluminum oxide; this is the colorant that provides the red colour of the gemstone ruby. Present in this energy level scheme is the ground state, designated $^4A_2$; this is the energy state of the chromium ion in ruby when in the dark. When illuminated by white light, either a photon of energy 2.2 eV or a photon of energy 3.0 eV can be absorbed, raising the system to the $^4T_2$ or $^4T_1$

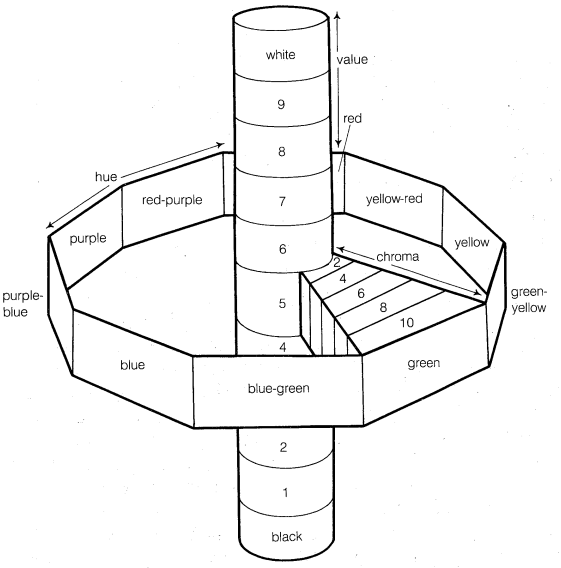



Figure 8: The hue, value, and chroma coordinates of Munsell's colour solid.

Energy transitions and the corresponding absorption of light

energy levels, respectively (in this system light cannot be absorbed into the level $^2E$ because of certain quantum limitations designated selection rules). These two energy transitions, broadened by the thermal atomic vibrations at room temperature into absorption bands, correspond to absorption of the violet and green-yellow parts of white light passing through the ruby, as shown at the centre in Figure 9. The remaining transmitted light consists of the strong red and weak blue parts of the spectrum, resulting in the deep red ruby colour with a slight purple overtone.

The chromium ion in ruby now contains excess energy, but the selection rules permit return to the ground state only through the intermediate $^2E$ energy level as shown at left in Figure 9. Part of the absorbed energy appears as a slight warming of the ruby. The other part is emitted as a photon producing a bright red fluorescence (best seen when the ruby is illuminated with ultraviolet radiation in the dark). The ruby has now returned to the ground state and energy has been conserved. This is just one explanation of the occurrence of colour. Although all occurrences or causes of colour involve the excitation of electrons, to simplify explanation, this article classifies the physical and chemical causes of colour into 15 groups. The first three involve transitions among the energy levels of excitations, vibrations, and rotations as explained by quantum theory. The next four involve modifications of this approach covered by the ligand field and molecular orbital theories. The following four involve the energy band formalism of solid state physics, and the final four are explained by geometrical and physical optics theory.

## SIMPLE EXCITATIONS, VIBRATIONS, AND ROTATIONS

**Incandescence.** Incandescent light is produced when hot matter releases parts of its thermal vibration energy as photons. At medium temperatures, say 800° C, the object's radiation energy reaches a peak in the infrared, with only a small intensity at the red end of the visible spectrum. As the temperature is raised, the peak moves toward and finally into the visible region. At successively higher temperatures the object becomes "red-hot," then orange, yellow, and finally "white-hot"; the very hottest of stars have a bluish-white colour. This sequence of colours is known as the blackbody radiation sequence and is included in the chromaticity diagram of Figure 7. Examples of incandescence include daylight (see Figure 4), candlelight, and light from tungsten filament lamps (see Figure 4), flashbulbs, the carbon arc, and pyrotechnic devices such as flares and fireworks.

Blackbody radiation sequence

**Gas excitation.** Gas excitation involves the emission of light by a chemical element present as a gas or vapour. When a gas such as neon or a vapour such as sodium or mercury is excited electrically, the electrical energy raises the atoms into high energy states from which they decay back to ground state with the emission of photons. This leads to the red light seen in neon tubes and the yellow and blue light seen in sodium and mercury vapour lamps, respectively. The same yellow sodium light is emitted when sodium atoms are thermally excited by being heated in a gas flame. In addition to being produced electrically or by chemical reactions, gas excitations can also result from interaction with energetic particles, as in auroras, where energetic particles emitted in solar storms excite gases high in the Earth's atmosphere to produce various colour effects.

**Vibrations and rotations.** All molecules have some vibration or rotation energy as a result of chemical bonding, but the energy involved is too low to interact directly with visible light. The frequency of vibration can be increased, however, by strengthening the chemical bonding involving very light atoms. For example, the bond between hydrogen and oxygen is stronger in liquid water and solid ice than in an isolated $H_2O$ molecule. The corresponding increase in vibration frequencies allows some absorption at the red end of the spectrum and produces the pale blue colour characteristic of pure water and ice when seen in bulk. Similar vibrations as well as rotations contribute to the purple colour of iodine, the brown of bromine, and the blue-to-green colours seen in the oxygen-rich gas flame of a kitchen range.

## LIGAND FIELDS

**Transition metal impurities.** Most chemical compounds are colourless when pure; examples include sodium chloride (ordinary table salt), aluminum oxide, naphthalene (moth flakes), and diamond. In these compounds all electrons are present in pairs. Such paired electrons are particularly stable and need very high energies to become unpaired and form excited energy levels. Only ultraviolet light is energetic enough to be absorbed, which explains the absence of visible light absorptions and the absence of colour. The compounds of a number of metals, most commonly iron, chromium, nickel, cobalt, and manganese, do, however, produce coloured salts. These metals are the transition elements, which contain unpaired electrons in their compounds. Excited energy levels are readily formed by these unpaired electrons, resulting in the absorption of photons and the production of colour.

Aluminum oxide, $Al_2O_3$, also known as corundum or colourless sapphire when pure, can serve as an example. In this compound each trivalent aluminum ion is surrounded by six oxygens in the configuration of a slightly irregular octahedron. The electric field at the aluminum site of this octahedral arrangement of oxygens is known as the ligand field (an older term, implying a simpler approach, was crystal field). If aluminum oxide contains chromium as an impurity, so that one out of every 100 aluminums is replaced by a chromium, which has unpaired electrons, then the ligand field produces a change in the energy levels

From K. Nassau, *Physics and Chemistry of Color* (1983); John Wiley & Sons, Inc.

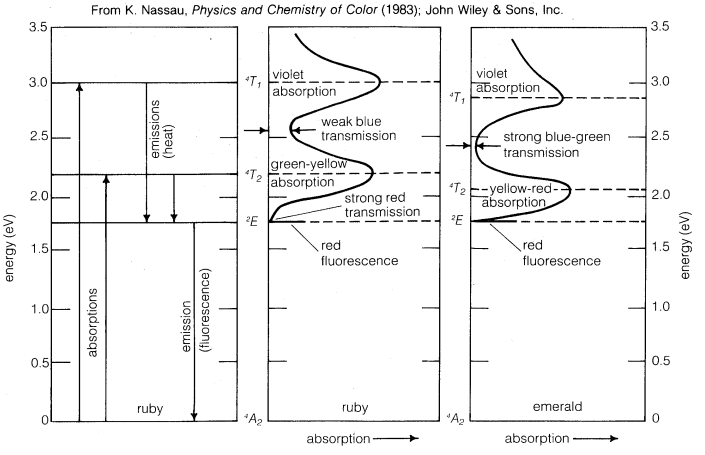



Figure 9: (Left) The energy level diagram of ruby with allowed transitions and (centre) the resulting absorption and fluorescence spectra. (Right) The absorption and fluorescence spectra of emerald.

that an isolated chromium ion would have. This gives the specific energy level scheme shown at the left in Figure 9, which leads to the light absorption curve at the centre of the Figure and produces the red colour (as well as the red fluorescence) of the chromium-containing aluminum oxide, also known as the gemstone ruby, as described above.

If chromium similarly replaces 1 or 2 percent of the aluminum in the compound beryllium aluminum silicate, a combination also known as the gemstone emerald, then the ligand field has the same geometry but is somewhat weaker, a result of the effect of the berylliums and silicons on the strength of aluminum-oxygen bonding. This produces small shifts in some of the absorption energy levels compared to ruby and results in the absorption spectrum shown at the right in Figure 9. These shifts have resulted in the almost total elimination of the red transmission and an intensification of the blue-green transmission, leading to an emerald-green colour. The $^2E$ energy level of Figure 9 has not shifted; accordingly, red ruby and green emerald show the same red fluorescence.

Other such transition metal impurities cause the colours of red iron ore and the gemstones yellow citrine and blue-to-green aquamarine (all coloured by a small percentage of iron impurity).

**Transition metal compounds.** The transition metal may be present not as an impurity but as an essential part of the substance. An example is chromium oxide, $Cr_2O_3$, also known as the pigment chrome green, in which the relatively weak ligand field of the chromium-oxygen bonding at the chromiums produces colour in a similar manner to that in the emerald discussed above. Additional examples are the copper-containing blue-to-green gem materials malachite, azurite, and turquoise, as well as the patina on copper statues, the red iron ore hematite, and the cobalt-glass pigment smalt.

## MOLECULAR ORBITALS

**Organic compounds.** All dyes and most pigments, whether natural or synthetic, are complex organic compounds whose molecular structures include a "colour-bearing" group known as a chromophore, usually a short conjugated system (a chain of atoms connected by alternating single and double bonds). The bonding electrons holding the molecule together can be viewed as belonging to the whole molecule. Simple conjugated chains have electronic transitions that absorb radiation only in the ultraviolet range of the spectrum. If, however, the chain is long, the resulting transitions between molecular orbital energy levels require less energy, and absorption shifts to longer wavelengths. The carotenes are naturally occurring examples of extended conjugated systems; they absorb some light in the violet or blue range of the spectrum and therefore appear yellow or orange in colour. The same effect occurs if the number of electrons present on a conjugated chain is modified by the addition of groups of atoms known as auxochromes. Auxochromes can be either electron acceptors or electron donors. Nitrophenylenediamine compounds contain both types of auxochromes. They absorb in the blue part of the spectrum and are often used in hair dyes because the small size of the molecules allows them to penetrate into hair easily.

Organic dyes occur widely in the plant and animal kingdoms as well as in the modern synthetic dye and pigment industry. Just as with ligand-field energy levels, some of the absorbed energy may be re-emitted in the form of fluorescence.

**Charge transfer.** Aluminum oxide containing a few hundredths of 1 percent of titanium is colourless. If it contains a similar amount of iron, a very pale yellow colour may be seen. If both impurities are present together, the aluminum oxide has a magnificent deep blue colour and is known as the gemstone sapphire. The colour is the result of charge transfer, in which the absorption of light energy allows an electron to move from one ion to another, resulting in a temporary change in the valence state of both ions:

$$Fe^{2+} + Ti^{4+} \longrightarrow Fe^{3+} + Ti^{3+}.$$

This process requires energy; since the energy corresponds to an absorption in the yellow region of the spectrum, the complementary colour blue results.

Other forms of charge transfer lead to the black of the iron oxide magnetite; the brilliant blue colour of potassium ferric ferrocyanide, the pigment Prussian blue; the yellow-to-orange chromates and dichromates; and the deep blue gemstone lapis lazuli, which has the same composition as the pigment ultramarine.

## ENERGY BANDS

**Metals.** The valence electrons, which in other substances produce bonding between individual atoms or small groups of atoms, are shared equally by all the atoms in a piece of a metal. These delocalized electrons are thus able to move over the whole piece of metal and provide the metallic lustre and good electrical and thermal conductivities of metals and alloys. Band theory explains that in such a system individual energy levels, as at the left in Figure 9, are replaced by a continuous region called a band, as in the density of states diagram for copper metal shown in Figure 10. This diagram shows that the number of electrons that can be accommodated in the band at any given energy varies; in copper the number declines as the band approaches being filled with electrons. The number of electrons in the copper fill the band to the level shown, leaving some empty space at higher energies.

Band theory

When a photon of light is absorbed by an electron near the top of the energy band, the electron is raised to a higher available energy level within the band as shown by the arrow in Figure 10. The light is so intensely absorbed that it can penetrate to a depth of only a few hundred atoms, typically less than a single wavelength. Because the metal is a conductor of electricity, this absorbed light, which is, after all, an electromagnetic wave, induces alternating electrical currents on the metal surface. These currents immediately re-emit the photon out of the metal, thus providing the strong reflection of a polished metal surface.

The efficiency of this process depends on certain selection rules. If the efficiency of absorption and re-emission is approximately equal at all optical energies, then the different colours in white light will be reflected equally well, leading to the "silvery" colour of polished silver and iron surfaces. In copper the efficiency of reflection decreases with increasing energy; the reduced reflectivity at the blue end of the spectrum results in a reddish colour. Similar considerations explain the yellow colour of gold and brass.

**Pure semiconductors.** In a number of substances a band gap appears in the density of states diagram (see Figure 11). This can happen, for example, when there are an average of exactly four valence electrons per atom in a pure substance, resulting in a completely full lower band, called the valence band, and an exactly empty upper band, the conduction band. Because there are no electron energy levels in the gap between the two bands, the lowest

From K. Nassau, *Physics and Chemistry of Color* (1983); John Wiley & Sons, Inc.

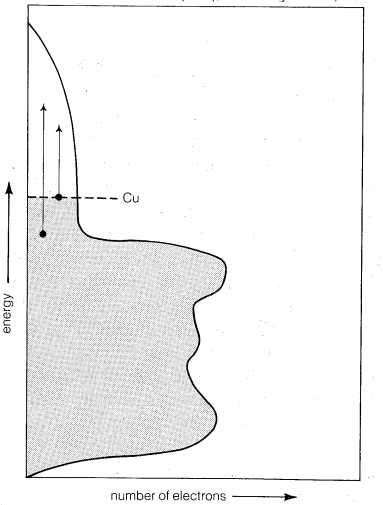



Figure 10: The density of states diagram of copper metal.

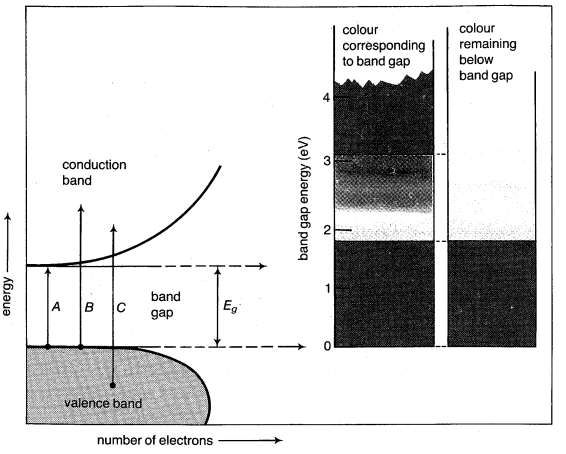



Figure 11: (Left) The absorption of light in a band-gap material. (Right) The variation of colour with the size of the band gap $E_g$.
After K. Nassau, *Physics and Chemistry of Color* (1983); John Wiley & Sons, Inc.

Band-gap energy

energy light that can be absorbed corresponds to arrow *A* in Figure 11; this represents the excitation of an electron from the top of the valence band up to the bottom of the conduction band and corresponds to the band-gap energy designated $E_g$. Light of any higher energy can also be absorbed as indicated by the arrows *B* and *C*.

If the substance represented by this Figure has a large band gap, such as the 5.4 eV of diamond, then no light in the visible spectrum can be absorbed and the substance appears colourless when pure. Such large band-gap semiconductors are excellent insulators and are more usually treated as ionic or covalently bonded materials.

The pigment cadmium yellow (cadmium sulfide, also known as the mineral greenockite) has a smaller band gap of 2.6 eV, which permits absorption of violet and some blue but none of the other colours. This leads to its yellow colour as can be deduced from the colour scale at the right in Figure 11. A somewhat smaller band gap that permits absorption of violet, blue, and green produces the colour orange; a yet smaller band gap as in the 2.0 eV of the pigment vermillion (mercuric sulfide, the mineral cinnabar) results in all energies but the red being absorbed, which leads to a red colour. All light is absorbed when the band-gap energy is less than the 1.77 eV (700 nm) limit of the visible spectrum; narrow band-gap semiconductors, such as the lead sulfide galena, therefore absorb all light and are black. This sequence of colourless, yellow, orange, red, and black is the precise range of colours available in pure semiconductors.

**Doped semiconductors.** If an impurity atom, often called a dopant, is present in a semiconductor (which is then designated as doped) and has a different number of valence electrons from the atom it replaces, extra energy levels can be formed within the band gap. If the impurity has more electrons, such as a nitrogen impurity (five valence electrons) in a diamond crystal (consisting of carbons, each having four valence electrons), a donor level is formed. Electrons from this level can be excited into the conduction band by the absorption of photons; this occurs only at the blue end of the spectrum in nitrogen-doped diamond, resulting in a complementary yellow colour. If the impurity has fewer electrons than the atom it replaces, such as a boron impurity (three valence electrons) in diamond, a hole level is formed. Photons can now be absorbed with the excitation of an electron from the valence band into the hole level. In boron-doped diamond this occurs only at the yellow end of the spectrum, resulting in a deep blue colour as in the famous Hope diamond.

Some materials containing both donors and acceptors can absorb ultraviolet or electrical energy to produce visible light. For example, phosphor powders, such as zinc sulfide containing copper and other impurities, are used as a coating in fluorescent lamps to convert the plentiful ultraviolet energy produced by the mercury arc into fluorescent light (see Figure 4). Phosphors are also used to coat the inside of a television screen, where they are activated by a stream of electrons (cathode rays) in cathodoluminescence, and in luminous paints, where they are activated by white light or by ultraviolet radiation, which causes them to display a slow luminous decay known as phosphorescence. Electroluminescence results from electrical excitation, as when a phosphor powder is deposited onto a metallic plate and covered with a transparent conducting electrode to produce lighting panels.

Fluorescent lamps

Injection electroluminescence occurs when a crystal contains a junction between differently doped semiconducting regions. An electric current will produce transitions between electrons and holes in the junction region, releasing energy that can appear as near-monochromatic light, as in the light-emitting diodes (LEDs) widely used on display devices in electronic equipment. With a suitable geometry, the emitted light can also be monochromatic and coherent as in semiconductor lasers.

**Colour centres.** A colour centre often involves a solid that is missing an atom, such as sodium chloride, an ionic crystal that consists of a three-dimensional array of positively charged sodium ions and negatively charged chloride ions. When a negative chloride ion is missing from the crystal, a way electrical neutrality can be maintained is if a free electron occupies the spot vacated by the chloride ion, forming an F-centre (after the German *Farbe,* "colour"). This replacement electron can be viewed as providing a trapping energy level within the large band gap.

F-centres

Some form of relatively high energy, such as ultraviolet radiation or high-energy X rays or gamma rays, can then be used to promote an electron from the valence band into the trap, which contains excited energy levels such as that designated $E_a$ in Figure 12. The $E_a$ level for the sodium chloride F-centre occurs at 2.7 eV and can absorb blue light, leading to a yellow-brown colour; such a defect is called a colour centre. The electron in this excited energy level is still within the trap. Only by supplying energy corresponding to $E_b$ can the electron leave the trap and return via the conduction band directly to the valence band. This can happen if the crystal is heated, resulting in bleaching of the colour centre. If $E_b$ is about the same size as $E_a$, bleaching can occur merely while the material is being illuminated, leading to optical bleaching. If $E_b$ is sufficiently small, the material may even fade in the dark at room temperature. This occurs in self-darkening sunglasses: the ultraviolet energy present in sunlight produces darkening, and room temperature leads to fading as soon as ultraviolet light is no longer present.

## GEOMETRICAL AND PHYSICAL OPTICS

**Dispersion and polarization.** In his 1666 experiment, shown in Figure 1, Newton discovered what is now called

From K. Nassau, *Physics and Chemistry of Color* (1983); John Wiley and Sons, Inc.

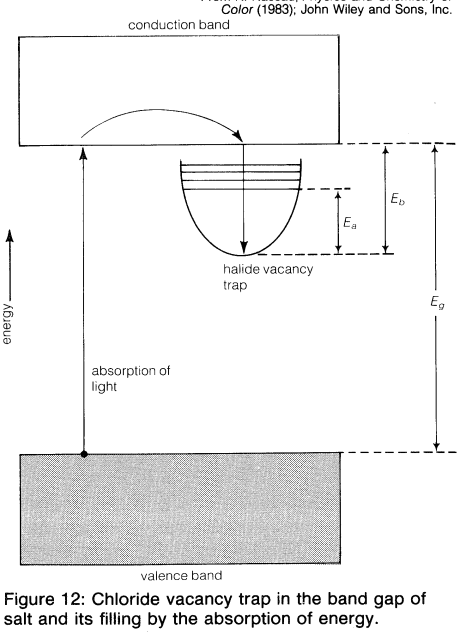



Figure 12: Chloride vacancy trap in the band gap of salt and its filling by the absorption of energy.

dispersion or dispersive refraction. He showed that a light beam is bent, or refracted, as it passes from one medium to another—*e.g.*, from air into glass. The natures of the two media as well as the wavelength of the light involved determine the degree of refraction, with shorter wavelengths bending more than longer wavelengths. Dispersion in a faceted diamond produces coloured flashes of light, in drops of water in the atmosphere it produces primary and secondary rainbows, and in ice crystals in thin clouds it produces a variety of halos and arcs around the Sun and Moon.

Absorption related to dispersion

Dispersion has its origin in absorption. Even a colourless, transparent substance, such as glass, absorbs electromagnetic radiation in the ultraviolet (derived from the unpairing of paired electrons and their further excitation) and in the infrared (from the vibrations of atoms, molecules, and larger structural units). It is a combination of these two effects that produces dispersion: only a vacuum has no absorptions and therefore no dispersion.

A rope can be snapped so that a wave movement travels from one end to the other; the motion of the wave can be from side to side, up and down, or in any direction perpendicular to the rope. Similarly, an unpolarized light wave travels in a single direction but vibrates in random directions perpendicular to its travel. When a light wave vibrates in only one direction, it is called polarized.

Light can be polarized in passing through certain substances (such as a crystal of calcium carbonate, the mineral calcite, or a sheet of polarizing film) that block out all waves except those vibrating in a particular direction. Polarized white light can interact with various doubly refracting materials (ones in which the index of refraction varies according to the direction in which the light waves passing through it vibrate) to produce colour. This technique is often used to view rocks or structural models; the colours produced are then studied to determine mineral composition or to analyze stress.

**Scattering.** When light strikes fine particles or an irregular surface, it is deflected in all directions and is said to be scattered. When the scattering particles are very small compared to the wavelength of light, the intensity of the scattered light is related to that of the incident light by the inverse fourth power of the wavelength (Rayleigh scattering). As a result, light at the blue end of the spectrum is scattered much more intensely than that at the red end.

Rayleigh scattering

The light from the Sun is scattered by dust particles and clusters of gas molecules, and the scattered blue rays seen against the dark background of outer space cause the sky to appear blue. At sunrise and sunset, when sunlight travels the farthest, almost all of the blue rays are scattered and the light that reaches the Earth directly is seen as predominantly red or orange. Scattering also causes that epitome of rare occurrences, the blue Moon (seen when forest fires produce clouds composed of small droplets of organic compounds). Most blue and green bird feathers involve scattering, as do many animal and some vegetable blues. Scattering also produces the blue colour of eyes, particularly the intense blue eyes of most infants, whose yellow-to-dark-brown pigments such as melanin have not yet all been formed so that only blue is seen against the dark interior of the eye.

If the size of the scattering particles approaches the wavelength of light or exceeds it, the complex Mie scattering theory applies and explains colours other than blue; white is scattered at the largest sizes, as in fog and clouds.

**Interference.** Two light waves of the same wavelength can interact under appropriate circumstances so as to reinforce each other if they are in phase or to cancel each other if they are out of phase. If a beam of light falls on a thin film, such as an oil slick on a puddle of water, part of the beam is reflected from the front of the oil film and part from the back. Depending on the thickness of the film, the two reflected beams can reinforce or cancel.

When monochromatic light falls on a film of tapering thickness, a series of dark and light bands, known as interference fringes, is produced. With white light the sequence of overlapping light and dark bands from the spectral colours leads to Newton's colours. The film appears black or gray where it is thinnest and the light waves cancel; as it becomes progressively thicker, it appears white, then yellow, orange, red, violet, blue, green, yellow, orange-red, violet, and so on. Newton's colours can also be seen in cracks in glass or in crystals, in a soap bubble, and in antireflection coatings such as on camera lenses.

Newton's colours

A large number of structural colorations in biologic systems also derive from thin film interference. These structures usually feature multiple layers and are frequently backed by a dark layer of melanin, which intensifies the colour by absorbing the nonreflected light. Such colorations are usually iridescent; the colours appear metallic and change with orientation. Examples include pearl and mother-of-pearl, the transparent wings of houseflies and dragonflies, the scales on beetles and butterflies, and the feathers of hummingbirds and peacocks. The eyes of many nocturnal animals contain multilayer structures that improve night vision and can produce iridescent reflections in the dark.

**Diffraction.** Interference is also involved in diffraction, another phenomenon that produces colour. Diffraction is the term used to describe the spreading of light at the edges of an obstacle and the subsequent interference that occurs. When a monochromatic beam of light falls on a single edge, a sequence of light and dark bands is produced; and with white light a sequence of colours much like the Newton colour sequence appears.

A diffraction grating consists of a regular two- or three-dimensional array of objects or openings that scatter light according to its wavelength over a wide range of angles. As these deflected waves interact, they reinforce one another in some directions to produce intense spectral colours. This effect can be seen by looking at a distant streetlight or flashlight through a black cloth umbrella. Diffraction arrays that reveal spectral colours in direct sunlight exist on the wings of some beetles and the skins of some snakes. Perhaps the most outstanding natural diffraction grating, however, is the gemstone opal. Electron microscope photographs reveal that an opal has a regular three-dimensional array of equal-size spheres, about 250 nm (0.00001 inch) in diameter, which produce the diffraction.

## The perception of colour

### COLOUR EFFECTS

When a person views an opaque coloured object, it is only the light deflected from the object that can activate the visual process in the eye and brain. Because different illuminants have different spectral energy distributions, as shown in Figure 4, a given object in these illuminations will reflect different energy distributions. Yet the eye and brain are such superb systems that they are able to compensate for such differences, and normal-appearing colours are perceived, a phenomenon called colour constancy.

Colour constancy

Colour constancy does not apply, however, when there are subtle differences in colour. If, for example, two orange objects, one coloured by an orange pigment, the other by a combination of red and yellow pigments, match precisely in daylight, in the light of a tungsten lamp one may appear more reddish than the other. Because of this effect, called metamerism, it is always necessary to follow precisely the illumination and viewing conditions specified when comparing a sample colour to one in a colour atlas.

The intensity of illumination also affects colour perception. At very low light levels, blue and green objects appear brighter than red ones compared to their relative brightness in stronger illumination, an effect known as the Purkinje shift. At higher levels of illumination, there is a related shift in hues, called the Bezold-Brücke effect, such that most colours appear less red or green and more blue or yellow as the intensity of illumination increases.

If a bright spot of white light is projected onto a screen uniformly illuminated with a pale blue light, the effect known as simultaneous colour contrast makes the white light appear pale yellow and the blue light seem grayer than if the two were viewed separately. The complementary hue is induced by the adjacent illumination. Successive colour contrast, which occurs when a person stares at one colour and then shifts to another, produces the same effect. A person who stares at a pattern of colours for some

time and then looks at a white area sees a negative afterimage of the pattern in complementary hues. This effect, also called chromatic adaptation, is what causes browns to appear reddish to someone who has just viewed a green lawn. Thus, even when the colour of a given object is measured and its physical cause identified, visual effects can prevent the precise perception of that colour from being specified. Some of these effects can be explained fairly simply by changes in the sensitivity of the eye's receptors to different colours as intensity changes, by fatigue in specific receptors, or by receptor inhibition; others are not understood. In fact, scientists did not know the process by which the eye and brain perceive colour until the early 1960s and even now do not understand all the details.

### COLOUR VISION

The trichromatic theory

One of the most successful theories of colour vision, the trichromatic theory, was first proposed around 1801 by Thomas Young, an English physician, and refined about 50 years later by the German scientist Hermann von Helmholtz. Based on experiments in colour matching, this theory postulates three types of colour receptors in the eye. The actual existence of such receptor cells, known as cones (from their shapes), was finally confirmed in the early 1960s. The three types of cones have maximum sensitivities in the blue, green, and red regions of the spectrum, with absorption peaks near 445 nm, 535 nm, and 565 nm, respectively. These three sets are often designated as S, M, and L for their sensitivity to short, medium, and long wavelengths. The trichromatic theory explains that colour vision results from the relative intensity of response of the S, M, and L cones. (Equal stimulation of all three gives the perception of white.) There is obviously a close connection between this trichromatic theory and the tristimulus value system.

One of the trichromatic theory's strengths is that the existence of several kinds of colour blindness can be simply explained as the lack of function of one or more sets of the cones. If one set of cones does not function, dichromatism results. People with deuteranopia (M set missing) or protanopia (L set missing) perceive only blue and yellow. In the much rarer tritanopia the S cones are missing and only green and red are perceived. Persons who have no functioning cone system suffer from the extremely rare monochromatism and can perceive only grays.

The opponent process theory

Although the trichromatic theory seems to explain much about colour vision, other theories have also been supported and studied, especially the opponent process theory. First proposed by the German physiologist Ewald Hering in 1878, this approach presumes that colour vision involves three mechanisms, each responding to a pair of opposites, namely, light–dark, red–green, and blue–yellow. It is based on many psychophysical observations, including the fact that blue and yellow (and also red and green) cannot coexist in any perceived colour; there are no bluish yellows (or reddish greens). Several of the contrast and afterimage effects can be explained very simply by this approach.

It is now recognized that the trichromatic and opponent process theories are not incompatible. They have been combined in a number of zone theories, which postulate that the cones function in a trichromatic manner in one zone, while in another zone the signals from the cones are combined in neural cells so as to produce one achromatic (white–black) and two chromatic (blue–yellow and green–red) signals, which are then interpreted in the brain. Although it is clear that zone theories, encompassing both trichromatic and opponent colour theories, are fully successful in explaining the many aspects of colour perception, there are still details that remain to be worked out.

## The psychology of colour

The most important aspect of colour in daily life is probably the one that is least defined and most variable. It involves aesthetic and psychological responses to colour and influences art, fashion, commerce, and even physical and emotional sensations. One example of the link between colour and emotion is the common perception that red, orange, yellow, and brown hues are "warm," while the blues, greens, and grays are "cold." The red, orange, and yellow hues are said to induce excitement, cheerfulness, stimulation, and aggression; the blues and greens security, calm, and peace; and the browns, grays, and blacks sadness, depression, and melancholy. It must be remembered, however, that the psychological perception of colour is subjective, and only general comments about its features and uses can be made.

Colours are not universal. Some languages do not contain separate words for green and blue or for yellow and orange, while Eskimos use 17 words for white as applied to different snow conditions. When colour terminology in different cultures is compared, certain patterns are observed consistently. All languages have designations for black and white. If a third hue is distinguished, it is red; next comes yellow or green, and then both yellow and green. Blue is the sixth colour named, and brown is the seventh. Finally, in no particular sequence, the colours gray, orange, pink, and purple are designated.

The role of cultural conditioning

Like colour terminology, colour harmony, colour preferences, colour symbolism, and other psychological aspects of colour are culturally conditioned, and they vary considerably with both place and historical period. One cross-cultural study showed that American and Japanese concepts of warm and cold colours are essentially the same, but that in Japan blue and green hues are perceived to be "good" and the red-purple range as "bad," while in the United States the red-yellow-green range is considered "good" and oranges and red-purples "bad." The colour of mourning is black in the West, yet other cultures use white, purple, or gold for this purpose. Many languages contain expressions that use colour metaphorically (common examples in English include "green with envy," "feeling blue," "seeing red," "purple passion," "white lie," and "black rage") and therefore cannot always be translated literally into other languages because the colour may lose its associated symbolic meaning.

Colour symbolism serves important roles in art, religion, politics, and ceremonials, as well as in everyday life. Its strong emotional connotations can affect colour perception so that, for example, an apple- or heart-shaped figure cut from orange paper may seem to have a redder hue than a geometric figure cut from the same paper because of the specific psychological meaning that is associated with the shape.

In addition to emotional associations, factors that affect colour perception include the observer's age, mood, and mental health. People who share distinct personal traits often share colour perceptions and preferences. For example, schizophrenics have been reported to have abnormal colour perception, and very young children learning to distinguish colours usually show a preference for red or orange. Many psychologists believe that analyzing an individual's uses of and responses to colour can reveal information about the individual's physiological and psychological condition. It has even been suggested that specific colours can have a therapeutic effect on physical and mental disabilities.

Physical and emotional reactions to colour

Although these medical benefits are still in question, colour has been shown to cause definite physical and emotional reactions in humans and in some animals. Rooms and objects that are white or in light shades of "cool" colours may appear to be larger than those that are in intense dark or "warm" colours; black or very dark colours have a slimming, or shrinking, effect, as is well known to designers and decorators. A "cool" room decorated in a pale blue requires a higher thermostat setting than a "warm" room painted a pale orange in order to achieve the same sensation of warmth. People who view a display of unusual colours produced by special illumination may experience headaches and nervous disorders; tasty wholesome food served under such conditions appears repulsive and may even induce illness. Some colours induce a feeling of pleasure in the observer. When an affectively positive, or pleasurably perceived, colour is viewed after a less pleasant colour, it produces more pleasure than when viewed by itself, an effect known as affective contrast enhancement.

The effect of combinations of colours on an observer depends not only on the individual effects of the colours but also on the harmony of the colours combined and the composition of the pattern. Artists and designers have been studying the effects of colours for centuries and have developed a multitude of theories on the uses of colour. The number and variety of these theories demonstrates that no universally accepted rules apply; the perception of colour depends on individual experience.

**BIBLIOGRAPHY**

*General works:* ISAAC NEWTON, *Optics; or, A Treatise of the Reflexions, Refractions, Inflexions, and Colours of Light,* 4th ed. (1730, reissued 1979), the beginnings of the scientific study of colour; JOHANN WOLFGANG VON GOETHE, *Goethe's Theory of Colours* (1840, reissued 1975; originally published in German, 1810), with excellent observations explained by an untenable theory; DAVID L. MACADAM (ed.), *Sources of Color Science* (1970), covering theories developed in all periods but omitting Goethe and including only a little Newton; RALPH MERRILL EVANS, *An Introduction to Color* (1948, reissued 1965), an authoritative, highly readable introduction, with emphasis on technical applications; and ENID VERITY, *Color Observed* (1980), a readable general introduction. A bibliography of colour studies is given in MARY BUCKLEY, *Color Theory: A Guide to Information Sources* (1975). Current research on the subject, together with discussions of applications, is found in the magazines *Color Research and Application* (quarterly), *Inter-Society Color Council News* (bimonthly), and *Journal of the Optical Society of America; Part A, Optics and Image Science* (monthly).

*Colorimetry:* W.D. WRIGHT, *The Measurement of Colour,* 4th ed. (1969), an authoritative outline of the trichromatic system; DEANE B. JUDD and GÜNTER WYSZECKI, *Color in Business, Science, and Industry,* 3rd ed. (1975), an authoritative work for the specialist; and FABER BIRREN, *A Grammar of Color: A Basic Treatise on the Color System of Albert H. Munsell* (1969), a revision of the 1921 work. Sets of colour chips used as identifiers are collected in *Munsell Book of Color* (1929– ), a loose-leaf publication. See also KENNETH L. KELLY and DEANE B. JUDD, *Color: Universal Language and Dictionary of Names* (1976), a system of simple colour names with cross-references to thousands of commonly used names; FRED W. BILLMEYER, JR., and MAX SALTZMAN, *Principles of Color Technology,* 2nd ed. (1981), a highly technical but readable text with an annotated bibliography; GÜNTER WYSZECKI and W.S. STILES, *Color Science: Concepts and Methods, Quantitative Data and Formulae,* 2nd ed. (1982), an advanced text; DEANE B. JUDD, *Contributions to Color Science* (1979); and DAVID L. MACADAM, *Color Measurement: Theme and Variations,* 2nd rev. ed. (1985).

*Physics and chemistry of colour:* R. DANIEL OVERHEIM and DAVID L. WAGNER, *Light and Color* (1982), a brief survey; FRANCIS A. JENKINS and HARVEY E. WHITE, *Fundamentals of Optics,* 4th ed. (1976); LESLIE E. ORGEL, *An Introduction to Transition-Metal Chemistry: Ligand-Field Theory,* 2nd ed. (1966); and KEITH MCLAREN, *The Colour Science of Dyes and Pigments* (1983), authoritative intermediate to advanced texts. KURT NASSAU, *The Physics and Chemistry of Color: The Fifteen Causes of Color* (1983), is a comprehensive up-to-date treatment.

*Perception of colour:* G. HUGH BEGBIE, *Seeing and the Eye: An Introduction to Vision* (1969, reprinted 1973), a survey for the general reader; GERALD S. WASSERMAN, *Color Vision: An Historical Introduction* (1978); and TOM N. CORNSWEET, *Visual Perception* (1970), more detailed treatments; RALPH MERRILL EVANS, *The Perception of Color* (1974); and ROBERT M. BOYNTON, *Human Color Vision* (1979), comprehensive advanced texts. EDWARD C. CARTERETTE and MORTON P. FRIEDMAN (eds.), *Handbook of Perception,* vol. 5 (1975), contains 12 chapters, written at the advanced level, on all aspects of colour perception.

*Colour in art:* SAMUEL J. WILLIAMSON and HERMAN Z. CUMMINS, *Light and Color in Nature and Art* (1983), a readable, wide-ranging intermediate-level textbook; JOHANNES ITTEN, *The Art of Color: The Subjective Experience and Objective Rationale of Color* (1961, reprinted 1973; originally published in German, 1961), an exposition of an influential aesthetic theory; M.E. CHEVREUL, *The Principles of Harmony and Contrast of Colors and Their Applications to the Arts* (1854, reissued 1981; originally published in French, 1839), with notes by Faber Birren; FABER BIRREN, *Principles of Color: A Review of Past Traditions and Modern Theories of Color Harmony* (1969), an introduction, and his *History of Color in Painting: With New Principles of Color Expression* (1965), an authoritative treatment; and GEORGE A. AGOSTON, *Color Theory and Its Application in Art and Design* (1979), a broad review valuable as a reference work.

(Ku.N.)

# Columbus

Opening the New World to Renaissance Europe, Christopher Columbus (in Spanish Cristóbal Colón; in Italian Cristoforo Colombo) was one of the greatest mariners of all time, but his despotic and capricious administration of the West Indies, which he claimed for Spain, eventually resulted in his recall. The combination of an outstanding and highly imaginative mind with a temper so high-strung and autocratic that his thoughts, writings, and actions at times suggested a man just this side of delusion initially brought him honour and riches, yet in the course of time led to a tragic reversal of his fortunes. From his next-to-last voyage to the New World the Admiral of the Ocean Seas and Viceroy of the Indies was returned in chains. After spending the remaining six years of his life in pitiable attempts to regain the splendour and vast possessions that had eluded him, he died in obscurity.

He was born in 1451, probably between August 26 and October 31, the son of Domenico Colombo, a weaver living then at the Italian port city of Genoa and later at Savona, and of Suzanna Fontanarossa. The evidence concerning the early life of Columbus is sparse and admits of many interpretations. One view is that he at first followed his father's business and resided at Genoa and Savona, but in one account Columbus says that he went to sea when he was 14. In any event, he was a pirate in the service of René d'Anjou early in the 1470s, and a few years later he went to the Greek island of Chios. He fought in a battle off Cape St. Vincent, under his namesake and probable relative Guillaume de Casenove-Coullon, on August 13, 1476, when his ship took fire and he swam to the Portuguese coast with the help of an oar. Some scholars believe that Columbus had been living in Savona in 1476 and had joined the convoy of ships that was organized by Genoa and attacked by the Portuguese. He never wrote in any form of Italian—not even to his brothers or to Genoese persons and institutions—but always in Spanish. Columbus signed himself successively Colombo, Colomo, Colom, and Colón. The last was the form he himself came to prefer and wished to be used. He never took the traditional form Columbus, not even when writing Latin. His brothers, also, in Spain and in the Indies, called themselves Colón and Hispanicized their Christian names (Bartholomeo becoming Bartolomé and Giacomo, Diego). One explanation is that Columbus came from a Spanish-Jewish family settled in Genoa. It has also been suggested that the family spoke a Genoese dialect and that Christopher, nearly illiterate upon leaving home, adopted the Spanish language of his new associates.

By courtesy of the Metropolitan Museum of Art, New York, gift of J. Pierpont Morgan, 1900



Columbus, oil painting by Sebastiano del Piombo, 1519. In the Metropolitan Museum of Art, New York.

**Columbus' plan.** Columbus discovered America by prophecy rather than by astronomy. "In the carrying out of this enterprise of the Indies," he wrote to King Ferdinand and Queen Isabella in 1502, "neither reason nor mathematics nor maps were any use to me: fully accomplished were the words of Isaiah"—referring to a more or less apposite passage in Isaiah 11:10–12—and, in fact, any writing became prophetic in his eyes when it could be interpreted as a forecast of his discovery. Nor is it any wonder that he felt sure of having been divinely selected for a mission, an assurance that he often sets forth in his writings and that is the source of both his pride and his humility, for his was a life rich in dramatic scenes.

Near disasters on early voyages

His very arrival in Portugal, his being miraculously saved from the wreck of his ship, and his landing so close to the rock of Sagres, where Prince Henry the Navigator had established his academy of seamanship, seem to justify his sense of having been chosen.

Portugal was then the westernmost end of the known world and Lisbon the natural meeting place of mariners bent on discovery. From there the Portuguese had discovered Madeira, turned the forbidding Cape Bojador on the Atlantic coast of Africa, and reached the Tropic of Cancer; and it was in Lisbon that these and future exploits were talked about, prepared, financed, and manned. The city was already a busy centre for the arts and sciences, particularly cosmography and astronomy. It was probably there that Christopher came across his younger brother Bartolomé, who had also left home for the sea and was working as a cartographer.

The idea of reaching the East by sailing westward seems first to have been mooted by the Florentine cosmographer Paolo Toscanelli. A concession that may be interpreted as contemplating such a quest was granted by the King of Portugal to one Fernão Telles on November 20, 1475. Soon after landing in Portugal, Columbus sailed to Iceland and during this voyage seems already to have been thinking of his voyage to "Cathay" through the West. In 1478 he married Felipa Perestrello e Moniz, a member of one of the first families of Portugal; by this marriage, Columbus had a son, Diego, born in 1479 or 1480. He then settled for a time on the island of Porto Santo, Madeira Islands, of which his brother-in-law had inherited the captaincy. From this base he acquired a considerable sailing experience of the South Atlantic, making one or perhaps even two voyages to the Portuguese trading post of Mina on the Gold Coast, the southernmost extremity of the then known world. During these voyages he also received a number of hints (big canes, pine-tree trunks, pieces of wrought wood) of the probable existence of lands beyond the western horizon.

Back in Lisbon, Columbus read assiduously in Cardinal Pierre d'Ailly's *Image of the World* and Marco Polo's account of the East. But though he studied Ptolemy—whose misconception of the extent of Asia further strengthened his belief that Asia could be reached by travelling westward—and must have known Toscanelli's opinion, the system of ideas he elaborated owed most to the apocryphal biblical Second Book of Esdras (see, for example, II Esd. 3:18). These ideas were: (1) the Earth is round; (2) the distance by land between the edge of the West (Spain) and the edge of the East ("India"—*i.e.*, Asia) is very long; (3) the distance by sea between Spain and "India" is therefore very short; (4) the length of a degree is $56^{2}/_{3}$ miles. These "miles" were not Arabic (1,975.5 metres), which would have made the figure remarkably accurate, but Italian (1,477.5 metres), which made his Equator about

one-quarter too small. Columbus calculated that the land distance between Spain and "India" was 282°; he was therefore left with only 78° for the sea distance, which he further reduced by his method of reckoning the degree.

Error in computing location of "India"

The outcome of all these errors was that "India" would be about 3,900 miles (6,275 kilometres) from the Canaries—*i.e.,* more or less where America happens to be. This tallied tolerably well with II Esdras 6:42, where it is asserted that the Earth is six parts dry land and one part sea.

This system set Columbus' imagination afire with the untold avenues for discovery that the western way opened up. It is unlikely that he was ever clear as to the exact nature of such avenues: new lands? new ways to sail to the old lands? Although vague about his plan, Columbus was clear about what he wanted in exchange: honour and wealth, definitely in that order. The proposal was finally put to the King of Portugal, who rejected it (1484). Columbus then went to Spain. He applied first for help to the powerful Duque de Medina-Sidonia, who was not interested, then to the Conde (later Duque) de Medina Celi, who housed and sheltered Columbus from 1484 to 1486; but though the Conde conceived the idea of letting him have some caravels, he decided in the end that the enterprise was too big for even as big a subject as he was and sent the would-be discoverer to King Ferdinand and Queen Isabella. The Queen handed the matter over to her chief accountant; eventually Columbus was granted an audience with Ferdinand and Isabella in Córdoba in the spring of 1486.

First meeting with Ferdinand and Isabella

The monarchs then decided to set up a special commission of "learned men and mariners" to study Columbus' proposals. This commission, under the chairmanship of the Queen's confessor, Hernando de Talavera, then bishop of Ávila, made him wait four years. This was not, as is often asserted, because the commission was either incompetent or backward in its views but because Columbus was vague and secretive as well as incoherent. For four years he went from city to city, following the court. In Salamanca he became acquainted with Diego de Deza, a professor at the university there, who, in 1486, became the tutor of Don Juan, the heir to the throne. Thanks no doubt to Deza, who throughout his life was to be his main standby, Columbus received several sums of money from the royal treasury in 1487, and on May 12, 1489, he was granted the privilege of being lodged and fed at public expense on his travels to court. During this period he became entangled with Beatriz Enríquez, who was to be the mother of his son Fernando.

In 1490 the Talavera commission reported unfavourably; and though the King and Queen did not abandon him, Columbus may have felt that his Spanish quest was at an end. He went to La Rábida, near Huelva, where he met two men who were to restore his faith: the friar-astronomer Antonio de Marchena, and a pilot and shipowner from Palos, Martín Alonso Pinzón. Pinzón told him that a year earlier, in Rome, he had been informed by a papal cosmographer "of those lands which were still undiscovered." Columbus asked him to become his partner, and Pinzón accepted. The evidence available suggests that Marchena and another friar from La Rábida succeeded in having Columbus recalled to the court—not to discuss his plan this time but to name the price he set on it. The price was exorbitant. He was to be knighted, appointed grand admiral and viceroy (these titles to remain in his family forever), and he was to receive 10 percent of the transactions within his admiralty. The King and Queen were stunned by his audacity. He held his ground so firmly that no compromise was possible, and he was dismissed. He left Santa Fé (near Granada) early in January 1492, possibly for France and England, but friends at court persuaded the King and Queen to recall him, and all his requests were met.

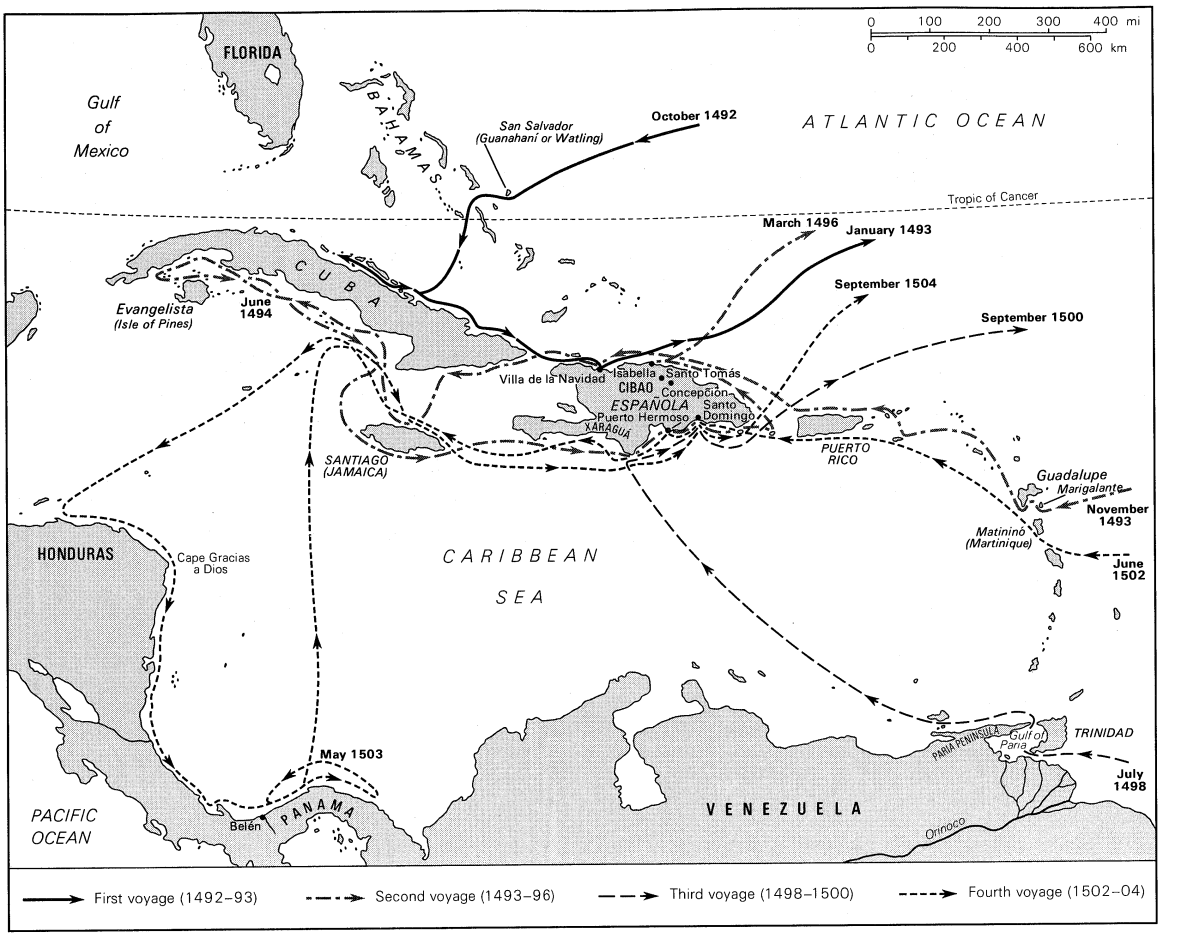
**First voyage.** Thanks to his own and Pinzón's initiative, all obstacles and the customary inertia enveloping official enterprises were swept away, and soon Columbus beheld at quay, ready to sail, his three vessels—the decked ship "Marigalante" (officially renamed "Santa María") and two caravels, "Pinta" and "Niña." The first was about 117 feet (36 metres) long, the other two about 50 feet (15 metres); the first two had castles fore and aft, the "Niña," only aft. They were armed with four-inch (10-centimetre) *bombardas* for heavy granite balls and *espingardes* for smaller lead projectiles. As Columbus did not know whether he was to come across new savages or old civilizations, he loaded his ships with cheap merchandise to relieve aboriginals of their gold but also took on board one Luis de Torres—"who had been a Jew and knew Hebrew and Chaldean and a little Arabic"—in case he met the "grand khan." The story that the Queen pawned her jewels to fund the expedition is quite untrue. More than 1,000,000 maravedis came from the funds of the Santa Hermandad (Holy Brotherhood). The sum of 500,000 maravedis was supplied by Columbus himself (advanced by Pinzón). The "Santa María" sailed under Columbus; the "Pinta" under Martín Alonso Pinzón with his brother Francisco Martín Pinzón as his pilot; and the "Niña" under Vicente Yáñez Pinzón, another brother, later to prove one of the finest sailors of the era.

On August 3, 1492, half an hour before sunrise, the small fleet sailed out of Palos. On August 12 it reached the Canary Islands, where it stayed until September 6, when at last it sailed resolutely westward. Once the three ships finally lost sight of land (September 9) the greatness of Columbus began to reveal itself, for it was at this moment that he conquered by faith and willpower the resistance of the unbelieving and faint-hearted members of his crew. He showed a typical mixture of strategic resolution and tactical caution, for he quickly handed out an "instruction" to his men that after 700 leagues (2,415 miles) they were not to navigate at night because land would certainly be near. On September 13 he observed that the relation of the magnetic north pole to the true North Pole changes with the position of the observer. A whole month passed in alternation of good and bad weather, good and bad mood. "There will be no wind to come back," the unbelievers sighed; and a strong contrary wind came to the rescue of their leader. On September 25, Columbus himself was infected by doubt and sought the help of Pinzón, who that very evening "saw" land. Columbus altered course to southwest in search of it but it vanished. Disbelief led to a conspiracy that might have degenerated into a disastrous mutiny but for the spirited intervention of Pinzón. On October 6 Pinzón suggested altering course again to the southwest; but Columbus was too proud to listen. The next day the "Niña" again "saw" land, but it was another illusion. Flights of birds all southwestward, however, made Columbus swallow his pride and follow Pinzón's advice (October 7). "All the night they heard birds passing," and carved sticks and reeds, picked up, gave them comfort and hope. Two hours after midnight (October 12) a sailor named Rodrigo de Triana or Juan Rodríguez Bermejo saw land from the prow of the "Pinta."

First sighting of the New World

At dawn, a land of virginal beauty and fresh colour revealed itself to the delighted Spaniards. With his two captains and the officials accompanying the expedition, Columbus went ashore carrying the royal banner. He planted it on the shore and took possession of the land in the name of Ferdinand and Isabella. This land was Guanahaní, one of the Bahama Islands, which the Spaniards renamed San Salvador. But everything he saw persuaded Columbus that he was among "the islands which are set down in the maps at the end of the Orient." So, although he saw signs of gold on the noses of the natives, he left in a hurry to see "whether I can come across the Island of Cipango" (*i.e.,* Japan).

For a fortnight he wandered among lovely islands to which he gave Spanish names, hoping "the Lord would show him where gold is born," yet already alive to the possibilities of what would now be called economic development. He even thought of slavery. In his mind there seemed to combine and struggle the two different and antagonistic strains that, curiously enough, appear also in his name: Christbearer and colonizer. Columbus raised crosses everywhere, but he also kept an eye on the material value of things even to the extent of seeing men as goods for sale. His honeymoon with the islanders may have ended on the day when he removed by force seven of the inhabitants of Guanahaní. They began then and there




The voyages of Christopher Columbus.

to think of the Spaniards as only a shade less tyrannical than the *caniba* or "cannibals," who from islands further south came to take them away and eat them. Columbus was convinced that these *caniba* were the subjects of the great can or khan. This "fact" persuaded him that a big island, so often mentioned by the islanders as Cuba, was Cipango itself. This became the more certain in his mind when he found that in Cuba, when he asked where gold "was born," he was told: "In Cubanacan" (*nacan* meaning the centre of the island). He sent Luis de Torres to interview the khan, but Torres found neither the khan nor the source of gold. He did, however, discover a greater source of wealth—tobacco. The fleet sailed on. On November 21 Pinzón and his "Pinta" vanished before a strong east wind. Columbus was worried. Had Pinzón gone to discover the source of gold and to talk to the great khan before him? Or worse still, had he sailed to Spain to steal from Columbus the glory of the discovery? In search of his vanished second-in-command he arrived in what is now Haiti.

Columbus was so struck by the beauty of this island that he renamed it Española (Hispaniola), "the Spanish island" (he usually thought in terms of Spain as a whole, not of Castile or Aragon). He praised the natives for their intelligence, industry, and handsomeness and thought that "they are good to be ordered and made to work, sow and do all that is necessary and to adopt our ways." At this point the "Santa María" ran aground in a strong wind and became a total loss. It happened exactly at midnight on Christmas Day, 1492. The "Pinta" was not available, and Columbus lost no time persuading himself that God meant him to leave behind an establishment with all the men and stores he could not carry back in the "Niña." He named his establishment Villa de la Navidad and chose 38 men to remain there under Diego de Arana, a relative of his mistress Beatriz Enríquez. Columbus left them food and stores for a year, including ammunition—so that the natives should obey "with love and fear"—and sailed for Spain on January 4, 1493.

Return to Spain

On January 6 the "Pinta" joined the "Niña." Columbus sailed back 40 miles (64 kilometres) to seek a safe anchorage, and he and Pinzón went through a stormy scene on board the "Niña." Columbus rejected Pinzón's explanations and threatened to have him hanged from the castle door, which, Pinzón declared, "is what I deserve for having raised you to the honour in which you stand." The quarrel was patched up, although the two men remained estranged. Columbus was worried about the men he had left behind—a decision of which Pinzón strongly disapproved—and also by astrological considerations, as well as by his own struggle between haste to return and report and curiosity to explore. At last, on January 16, 1493, he sailed for Europe.

Columbus deliberately chose a more northerly course, roughly on the parallel of Cape St. Vincent, for his return voyage. The first weeks were happy and carefree, but in mid-February the ships met storms that threatened to drown the explorers. During the night of February 14 both captains seem to have resigned themselves to let the wind do as it pleased; they signalled throughout the night but by dawn they had lost sight of each other. The storm raged on. Tormented by the thought that his discovery might remain unknown, Columbus wrote out its story and entrusted it to a cask, which he threw overboard. For four days he went without sleep, but on Monday, February 18, he at last cast anchor at Santa Maria, the southernmost island of the Azores. He was so pleased that he admitted in his journal that he had kept secret the distances covered daily in order to put other sailors off the track of "his" Indies. On February 24, 1493, he sailed for Spain. Worse storms assailed him, but at dawn on March 4 he recognized the Rock of Sintra. He decided to enter the Tagus River. The Portuguese king received him on March 9 with full honours (though some of his courtiers, it appears, would gladly have killed him then and there) and suggested he go to Castile by land; but Columbus preferred his less unsafe element and on March 15 the

"Niña" was welcomed enthusiastically as it entered Palos.

Pinzón arrived at Palos soon after having sought, on the same day, refuge in Bayona, near Vigo, from which he wrote to the King and Queen a letter loyal to Columbus. He died, exhausted, on March 20. Columbus also wrote to the King and Queen, who, upon receipt of his letter, wrote to him from Barcelona on March 30 asking him to come to court at once. They feared lest the King of Portugal should steal a march on them, and they wished to send a second expedition as soon as possible; they also obtained two papal bulls granting them the Indies discovered and to be discovered and apportioning the undiscovered parts of the West between the two Iberian powers by a line from pole to pole 100 leagues (345 miles) west of the Azores. The man who had started this keen rivalry between two powerful kingdoms travelled in triumph from Seville to Barcelona, where at the end of April he was received by the King and Queen in a scene of solemnity and splendour. The monarchs rose to greet him and offered him a stool. This was but the beginning of a flow of privileges and honours that staggered the court, including the right to display a castle and a lion (royal symbols) on his arms. Wealth and honour were, indeed, his. He nevertheless insisted on being paid the prize of 10,000 maravedis promised to the first man of the crew to see land. The humble sailor who actually had first sighted land went over to Morocco in disappointment and became a renegade. This episode sheds some light on a side of Columbus' character that was to bring him much unhappiness.

**Second voyage.** Trouble began soon enough with the man the King and Queen had put in charge of "Indian" affairs, Juan de Fonseca, afterward bishop of Burgos but far more gifted as a soldier than as either an administrator or an ecclesiastic. Royal letters urged both Columbus and Fonseca to make haste; but the plan for the second expedition was ambitious: 17 vessels in all, manned by 1,000 to 1,500 persons. The material obstacles must therefore have been formidable enough, but much of the trouble came from personal differences between Columbus, Fonseca, and two other officials, Juan de Soria and Francisco Pinelo. Soria was reluctant to treat an upstart like Columbus as a magnificent lord and had to be reprimanded for it by his royal masters, while Columbus himself was impatient of all supervision.

The second fleet was organized on the basis of a memorandum Columbus had prepared more as "colonizer" than as "Christ-bearer." Populating and developing came to the fore, while Christianizing took what was left. But Columbus' ideas were sensible, and he even suggested that gold seeking was to be forbidden at certain periods of the year so that settlers could turn their attention to the land. The King and Queen, however, added their own ideas: Christianization and welfare of the Indians; assertion of royal sovereignty over the new territories and the men who were sailing; greater stress on land cultivation than on gold. Several priests were sent, a doctor, a surgeon, peasants, hidalgos, and Columbus' own brother Giacomo, who became known as Don Diego. The expedition was largely financed by items confiscated from *conversos* and expelled Jews. The fleet sailed from Cádiz on September 25, 1493.

After the usual stay in the Canary Islands to restock food and water and live animals for the New World, Columbus began his second crossing on October 13. He chose a southwesterly course aiming at the mainland; this enabled him to benefit by the trade winds, and on November 3 the expedition sighted land. They had the choice of Dominica and Marigalante (Marie-Galante), and they anchored at the second. The next day they discovered the islands of Guadalupe (Guadeloupe) and, later, Puerto Rico. But Columbus was looking for the mainland. He was also worrying about the men he had left in Española. When he at last arrived there, he found that the Spaniards had been wiped out by the local chiefs. On hearing of the disaster, he remained cool and ordered a search for any gold the massacred men might have buried. He sailed away in search of a better base on the island and kept moving for a month, at the end of which he chose a site and founded the first European city in the New World, Isabella (now a ruined settlement having the same name on the northern coast of the Dominican Republic).

Quarrels and exploration

From Isabella, Columbus sent to Spain his first messenger, Antonio de Torres, of the royal household, with a number of explanations and excuses, as well as proposals, all of which were eventually approved except his scheme for a regular exchange of Spanish cattle for cannibal slaves. The news of the disaster at La Navidad, however, shook faith in him at court. Was the discovery worth the trouble and expense? His credit in the New World was also falling. The chief accountant of the fleet wrote a report on Columbus for Juan de Soria, hiding it in a wooden buoy. Columbus ferreted it out, imprisoned the accountant, and dealt no less severely with other "conspirators." On March 12, 1494, he left for Cibao, a region of the island where gold was believed "to be born," leaving Isabella in the care of his inexperienced brother Diego. In Cibao he founded Santo Tomás, a settlement with a fortress, where he left 56 men under the command of a Catalan captain, Mosén Pedro Margarite.

On his return to Isabella on March 29 Columbus found the food putrified by the heat, the men hungry and angry. He tried to make them build a mill for the wheat he had brought and punished those who refused. Bernardo Buil, a Catalan friar who had come in charge of spiritual affairs, objected to his harsh ways (Columbus had several men hanged), and a conflict arose between the secular and the spiritual powers. The friar struck the viceroy with interdicts, and the viceroy countered by stopping the friar's rations. He then sent a strong force to relieve Margarite and to explore the island.

Although the situation in Isabella and Santo Tomás was by no means good, Columbus sailed away again (April 24, 1494), taking with him three caravels of the five that remained after Antonio de Torres' departure for Spain with 12; and he again left Isabella in the care of his brother, though now assisted by a council of men all of whom were better qualified than their chairman. Columbus was anxious to explore Cuba, which he sighted on April 29, but, persuaded by a converted Indian, he veered southward and on May 5 discovered Jamaica, which he named Santiago. All of Jamaica's charms, however, could not lure him away from his westward urge, and within eight days he sailed back toward Cuba, determined to find out whether or not it was the mainland. His determination came suddenly to an end on June 13, when he was about 150 miles (241 kilometres) west of the island. He decided to return to Isabella just because he made up his mind that Cuba was the mainland, which he forced everyone in his squadron to swear to on oath, under pain of having their tongues torn out should they recant. He sailed eastward, lingered at the Isla de Juventud (formerly the Isle of Pines, which he called Evangelista), and sailed around the southern coast of Jamaica. He then fell ill and remained near death for days. The fleet returned to Isabella on September 29.

Columbus' brother Bartolomé had now arrived and was put in command with the title of *adelantado* (governor), which, when it became known at court, was rightly considered an encroachment on royal authority. This step increased the discontent prevailing at Isabella. Margarite and Buil had left for Spain with many complaints against Columbus and his brothers. Torres returned with royal letters for the Admiral congratulating him on his success and enjoining all and sundry to obey him. But, alleging that the natives were restive, Columbus and Bartolomé left Isabella on March 24, 1495, to "pacify" them—in fact, to take prisoners who could be justified as slaves. After the natives were defeated, Columbus sent his trusted messenger Torres back to Spain with a report, together with his brother Diego and 500 prisoners. His human cargo arrived in Spain when Buil and Margarite had spread their unfavourable views of him. When Juan de Aguado, sent to Hispaniola by the King and Queen to report on the situation, landed (October 1495), he found the natives on the point of rebellion because Columbus had imposed a gold tax. A duel of jurisdictions began between Aguado and Columbus; it lasted five months and deeply humiliated the proud viceroy. He decided to leave

and appointed Bartolomé as governor and Diego, who had returned from Europe, as his substitute, and he sailed for Spain on March 10, 1496, on one of two caravels (the other had Aguado on board) that had been built in the island. Columbus lingered among his islands for weeks and finally began his crossing on April 20. On June 11 his ships—the first two American vessels to arrive in Europe—cast anchor at Cádiz.

**Third voyage.** Columbus chose to arrive in Franciscan habit as a protection against any further humiliations. He left at once for Seville and Burgos. His adversaries had the ear of the King and Queen, but in July he received a royal letter of welcome. He prepared an efficient colonialist, but not very Christian, memorandum for the future organization of the Indies. He still possessed his fire and magnetism, however. He won over his adversaries, had all his privileges confirmed and some of the more material ones substantially increased, and was authorized to populate Española with convicts. He declined a royal offer of an estate in Española, located 50 leagues (172 miles) east–west and 25 leagues (86 miles) north–south, with the title of duke or marquess, no doubt feeling that the estate to which he was entitled was no less than the whole New World.

Financial difficulties delayed the third expedition, which did not sail from Sanlúcar until May 30, 1498. It was composed of six ships carrying 200 men, excluding sailors. On June 19 Columbus arrived in Gomera in the Canary Islands, where he rescued some Spanish ships from a French corsair. He then decided to send three of his ships directly to his brother and sail southwest with the other three. With one ship and two caravels he left Gomera on June 21, 1498, and called at the Cape Verde Islands. While anchored in São Tiago, he was told of islands to the south. Southward he sailed in haste until July 13. He then turned west. Had he kept to his course, he would have discovered the Amazon within four days. By July 28 his course was parallel to the mainland he sought and he was only saved from a long frustration by a land crowned with three summits, which was sighted just as he ordered a change of course northward toward Dominica. He named the new land Trinidad. He sailed round the island to the Gulf of Paria, set foot on the mainland but deemed it an island, and left in a hurry, drawn by the call of his "mainland"—*i.e.,* Cuba. On a later landing on the Paria Peninsula, the sight of native women wearing pearl necklaces caused great excitement among the Spaniards, who saw in it proof that they were indeed in the Orient. Columbus had, however, time to observe that the gulf received huge and powerful currents of fresh water from what were, in fact, several mouths of the Orinoco River. On this observation he built up his fantastic theory that he was at the mouth of one of the four rivers of paradise, situated at the top of the Earth, the Earth being not round but pear-shaped.

When Columbus arrived in Española he found that Francisco Roldán, whom he had appointed mayor of Isabella, had rebelled with about 70 followers against Columbus' brothers in Xaraguá (southwestern Española). Bartolomé's attitude to the rebellion had been vigorous; Christopher's was temporizing. He wrote a subservient letter to Roldán; concluded an agreement with him; allowed the rebels to break it. He finally concluded a second agreement that amounted to complete surrender to Roldán, for Roldán's past was condoned and he was appointed *alcalde* (mayor). This done, Columbus sent a messenger to Spain reporting the incident and explaining why the agreement was void since, among other reasons, he had signed it as viceroy but on board a caravel where he was not viceroy but admiral. He felt depressed, and his troubles increased with the sudden arrival on September 5, 1499, of a small fleet he was not expecting. One of his youngest and most enterprising lieutenants, Alonso de Ojeda, while in Seville had heard of Columbus' discovery of pearls in Paria. He thereupon obtained from Fonseca permission to go "discovering," sailed about with varied fortunes, and, when short of supplies, calmly turned up at Española. Columbus had every right to consider him a poacher. He sent Roldán with two caravels to meet Ojeda, who listened to Roldán but went his own way to Xaraguá (February 1500). Roldán followed him there. Ojeda left again, but Xaraguá's settlers, who had already been in a state of near rebellion before Ojeda's visit, had to be terrorized into obedience by a capital execution ordered by Columbus himself. The rule of Columbus then came to a sudden end.

Columbus removed from command

Reluctantly and slowly the King and Queen had come to the conclusion that Columbus was a good admiral but a bad governor. He himself had written to them in October 1498 asking for a learned man to administer justice in the New World. They chose an old member of their household with an excellent reputation, Francisco de Bobadilla, to whom, in March 1499, they gave powers strictly limited to an enquiry into the rebellion and the punishment of the rebels. A fresh batch of bad news made them go one step further, and on May 21, 1499, Bobadilla was appointed governor and chief magistrate of Española by letters patent in which no mention was made of Columbus. Bobadilla was empowered to send back to Spain any person regardless of rank. A special letter addressed to Columbus and his brothers enjoined them to obey and deliver to Bobadilla all royal property, castles, arms, and stores. But the rest of the year went by and Bobadilla did not sail. There are strong reasons for thinking that the final decision to let him go was prompted by information the sovereigns had obtained of Columbus' obscure dealings, bordering on treason, with Genoese agents.

Bobadilla arrived at Santo Domingo in Española on August 23, 1500. Before the wind allowed him to land he was informed by a messenger sent out by Diego that several Spaniards had been hanged that week and five more were waiting for the gallows. He was also told that Christopher Columbus was in Concepción and Bartolomé in Xaraguá, each with a confessor to dispatch guilty prisoners by the same method. On August 24, Bobadilla, after mass, had his first dispatch read aloud—that appointing him inquirer into the rebellion—and asked for all prisoners to be handed over. Diego refused. Next day, after mass, he had his second dispatch read—that appointing him governor. Obstructed again, he had all his guns unmasked and revealed the royal letter enjoining him to pay all arrears of salaries, if need be, on Columbus' account. On Diego's persistent refusal, Bobadilla forced his way into the citadel and took possession of the prisoners. On September 15, Bobadilla met Columbus, who refused to acknowledge the appointment on the ground that the King and Queen could not deprive him of his position. Bobadilla had both him and Diego arrested and put in chains. Bartolomé, who in Xaraguá "had 16 Spaniards in a ditch or well waiting to be hanged," was summoned to Santo Domingo and also shackled. When he was brought to embark for Spain, Columbus feared that he was being conveyed to the gallows. Though he was treated with respect throughout the voyage he refused to be unchained. Sad, and with his irons still on, he landed in Cádiz toward the end of November 1500. The King and Queen, who were at Granada, were shocked at the news, ordered him to be set free, and sent him 2,000 ducats to enable him and his brother to come to court. When Columbus saw his royal patrons, he stood dumb with emotion, then fell on his knees and burst into tears.

The King and Queen were sympathetic and generous but by no means disposed to alter their view as to his political capacity. They probably disapproved of the harsh ways of Bobadilla but not, in general, of his handling of a delicate and possibly dangerous situation. Though legally bound by their agreement with Columbus, they had to consider that the land already discovered far exceeded their expectations and that he was no man to be entrusted with a government. Generous with him they could afford to be, but not at the cost of the public good. On the other hand, Columbus, now frustrated in his dreams of greatness in the New World, cast about for another grand work to achieve and found it in his biblical, prophetic mind: he would liberate Jerusalem. He did not need to consult cosmographers, merchants, or ambassadors. He read the prophets, and having collected all the texts he could showing that Jerusalem would be liberated by Spain, he presented his *Book of Prophecies* to the King and Queen.

Meanwhile, a crowd of discoverers was sailing his seas, landing in his islands, and stealing his future discoveries; everybody seemed to be entering the business to a point that made Columbus write bitterly: "They all made fun of my plan then; now even tailors wish to discover." That was all very well, but how was he going to liberate Jerusalem if he were not allowed to find more gold in that New World where gold was born? The King and Queen were willing, provided he did not return to Española; and he suggested a search for a passage through the newly discovered lands to the other seas.

**Last voyage.** Nicolás de Ovando, appointed governor to succeed Bobadilla, sailed for Española from Sanlúcar on February 13, 1502, in command of 32 ships, 2,500 men, and 12 Franciscan friars. He carried orders to restore to Columbus his property confiscated by Bobadilla. A few months earlier, in October 1501, Columbus had felt confident enough to go to Seville to prepare his fourth expedition. He bought four ships of about 50–60 tons, which he thought the best size for discovery, and he selected 146 men. He left his eldest son, Diego (then 21), to represent him at court, and took with him his younger boy, Fernando (then 13). He asked to be allowed to call at Española and was expressly forbidden to do so, at any rate on the way out. He sailed from Cádiz on May 9, 1502.

On May 25 Columbus began his crossing from the Canary Islands and on June 15, 1502, he discovered an island he called Matininó (Martinique), one of the Lesser Antilles. A few days later he sailed straight for Santo Domingo (now the capital city of the Dominican Republic). Ovando refused him admittance, and he left for Puerto Hermoso, 16 leagues (55 miles) to the west, where he sheltered from a storm. He produced pretexts for his disobedience to royal orders, but the true reason was his stiff-necked reluctance to bow before authority; it was but natural that he should regard as his those lands he had brought to light. He sailed for Jamaica on July 14, 1502, overcame a mutiny of his men, sailed along the coast of Honduras, and was at Cape Gracias a Dios on September 14. His men were mutinous again, but he sailed on, wandering in lands of his imagination. He wrote to the King and Queen that "from there to the river Ganges there are ten days," and he referred again to the earthly paradise. Formidable storms all but smashed his ships. On Epiphany Sunday, 1503, he was pushed by the wind into a kind of estuary he named Belén (Bethlehem). As there seemed to be gold about, he resolved to leave his brother Bartolomé there with 80 men; but because of the hostility of the natives he had to give up the plan and sail away,

> in the name of the Holy Trinity, on Easter night, with the ships rotten, worm eaten, all in holes. Two only remained in the same state, boatless, empty of supplies, to cross 7,000 miles [11,263 kilometres] of sea with a son, a brother and so many men.

On May 13, 1503, he was "in Mango province (Panama), which is next to that of Cathay, and from there I sailed towards Española." His geography remained that mixture of truth and fantasy that his whole life incarnated. After more storms, mutinies, imprecations to God, and narrow escapes, he found himself derelict in a small cove in Jamaica (June 23, 1503), whence he wrote to the King and Queen a truly magnificent letter telling of his perils, wanderings, dreams, and voices he heard from on high.

Return to Spain and death

For a whole year, Columbus struggled against all of his usual troubles: indiscipline, native restlessness, ill health, lack of shelter and food, doubt and disillusionment. He had sent a messenger to ask Ovando for help. Two caravels arrived late in the spring, and at last he was able to sail for Española on June 28, 1504, and for Spain on September 12. Storms pursued him again in his last crossing, but he landed safely in Sanlúcar on November 7, 1504. Crippled by arthritis, he could hardly move and had to be conveyed to Seville; travelling north was out of the question. On November 26, 1504, Queen Isabella died. Columbus hoped that she had provided in her will for him to be "restored in the possession of the Indies." His hopes were disappointed, and he wrote to his son to seek the help of King Ferdinand's men. He wrote memoir upon memoir on the Indies and their gold. Sick in body but worse still in mind with humiliation, frustration, and impatience, he was not able to move northward until May 1505, when he was received by Ferdinand in Segovia, and he asked to be restored "the capital of my honour . . . the government and possession of the Indies." The King was sympathetic and willing to please Columbus and, on his suggestion, put Columbus' old friend and staunch supporter Diego de Deza in charge of his affairs. But as a responsible monarch Ferdinand could not reappoint the passionate admiral as governor of the Indies. Columbus' health went from bad to worse. He dictated his will at Valladolid on May 19, 1506, and died two days later.

After the funeral ceremonies at Valladolid, Columbus' remains were in 1513 transferred to the Carthusian monastery of Santa María de las Cuevas in Seville, where the bones of his son Diego were also laid. Exhumed in 1542, the bodies of both were taken to Hispaniola and interred in the cathedral of Santo Domingo.

**BIBLIOGRAPHY.** C. FERNANDEZ DURO, *Colón y Pinzón* (1883); SALVADOR DE MADARIAGA, *Christopher Columbus,* new ed. (1949); R.H. MAJOR (ed.), *Select Letters of Christopher Columbus* (Hakluyt Society, 2nd ed., 1870), standard collection of the most important chapters on the life and discoveries of Columbus; SAMUEL ELIOT MORISON, *Admiral of the Ocean Sea: A Life of Christopher Columbus* (1942), and *Christopher Columbus, Mariner* (1955), two of the finest biographies of Columbus in the English language; M. FERNANDEZ DE NAVARRETE (ed.), *Colección de los viajes y descubrimientos que hicieron por mar los españoles,* 5 vol. (1825–37); MILTON A. RUGOFF (ed.), *The Great Travellers,* 2 vol. (1960), general text on the most important explorers, including Columbus; EARL P. HANSON, *South from the Spanish Main: South America Seen Through the Eyes of Its Discoverers* (1967), an anthology.

(Ed.)

# Combinatorics and Combinatorial Geometry

Combinatorics and combinatorial geometry are concerned with arrangements of mathematical elements, problems of selection or choice, permutations and combinations, and certain aspects of the theory of probability. The fundamental concepts and methods of these two closely related fields of mathematics are treated in this article, as are the factors that led to their development. (Ed.)

The article is divided into the following sections:


Combinatorics 611
- History 611
  - Early developments
  - Combinatorics during the 20th century
- Problems of enumeration 612
  - Permutations and combinations
  - Recurrence relations and generating functions
  - Partitions
  - The principle of inclusion and exclusion: derangements
  - Polya's theorem
  - The Möbius inversion theorem
  - Special problems
- Problems of choice 614
  - Systems of distinct representatives
  - Ramsey's numbers
- Designs, Latin squares, arrays, and coding 614
  - BIB (balanced incomplete block) designs
  - PBIB (partially balanced incomplete block) designs
  - Orthogonal Latin squares
  - Orthogonal arrays and the packing problem
- Graph theory 616
  - Definitions
  - Enumeration of graphs
  - Characterization problems of graph theory
  - Planar graphs
  - The four-colour map problem
  - Eulerian cycles and the Königsberg bridge problem
  - Directed graphs

Combinatorial geometry 618
- Some historically important topics of combinatorial geometry 619
  - Packing and covering
  - Polytopes
  - Incidence problems
  - Helly's theorem
- Methods of combinatorial geometry 620
  - Exhausting the possibilities
  - Use of extremal properties
  - Use of figures with special properties
  - Use of transformations between different spaces and applications of Helly's theorem

Bibliography 622


## Combinatorics

The scope of combinatorics is hard to define with any exactitude. In general, however, it may be said that it is concerned with arrangements, operations, and selections within a finite or a discrete system.

The problems studied with combinatorics

One of the basic problems is to determine the number of possible configurations (*e.g.*, graphs, designs, arrays) of a given type. Even when the rules specifying the configuration are relatively simple, enumeration may sometimes present formidable difficulties. The mathematician may have to content himself with finding an approximate answer or at least a good lower and upper bound.

In mathematics, generally, an entity is said to "exist" if a mathematical example satisfies the abstract properties that define the entity. In this sense it may not be apparent that even a single configuration with certain specified properties exists. This situation gives rise to problems of existence and construction. There is again an important class of theorems that guarantee the existence of certain choices under appropriate hypotheses. Besides their intrinsic interest these theorems may be used as existence theorems in various combinatorial problems.

Finally, there are problems of optimization. As an example, a function $f$, the economic function, assigns the numerical value $f(x)$ to any configuration $x$ with certain specified properties. In this case the problem is to choose a configuration $x_0$ that minimizes $f(x)$, or makes it $\varepsilon =$ minimal—that is, for any number $\varepsilon > 0$, $f(x_0) \leq f(x) + \varepsilon$, for all configurations $x$, with the specified properties.

### HISTORY

**Early developments.** Certain types of combinatorial problems have attracted the attention of mathematicians since early times. Magic squares, for example, which are square arrays of numbers with the property that the rows, columns, and diagonals add up to the same number, occur in the *I Ching*, an old Chinese book dating back to 2200 BC. The binomial coefficients, or integer coefficients in the expansion of $(a + b)^n$, were known to the 12th-century Indian mathematician Bhāskara, who in his *Līlāvatī* ("The Graceful"), dedicated to a beautiful woman, gave the rules for calculating them together with illustrative examples. "Pascal's triangle," a triangular array of binomial coefficients, had been taught by the 13th-century Persian philosopher Naṣīr ad-Dīn aḷ-Ṭūsī.

In the West, combinatorics may be considered to begin in the 17th century with Blaise Pascal and Pierre de Fermat, both of France, who discovered many classical combinatorial results in connection with the development of the theory of probability. The term combinatorial was first used in the modern mathematical sense by the German philosopher and mathematician Gottfried Wilhelm Leibniz in his *Dissertatio de Arte Combinatoria* ("Dissertation Concerning the Combinational Arts"). He foresaw the applications of this new discipline to the whole range of the sciences. The Swiss mathematician Leonhard Euler was finally responsible for the development of a school of authentic combinatorial mathematics beginning in the 18th century. He became the father of graph theory when he settled the Königsberg bridge problem, and his famous conjecture on Latin squares was not resolved until 1959.

Combinatorics before 1920

In England, Arthur Cayley, near the end of the 19th century, made important contributions to enumerative graph theory, and James Joseph Sylvester discovered many combinatorial results. The British mathematician George Boole at about the same time used combinatorial methods in connection with the development of symbolic logic, and the combinatorial ideas and methods of Henri Poincaré, which developed in the early part of the 20th century in connection with the problem of $n$ bodies, have led to the discipline of topology, which occupies the centre of the stage of mathematics. Many combinatorial problems were posed during the 19th century as purely recreational problems and are identified by such names as "the problem of eight queens" and "the Kirkman school girl problem." On the other hand, the study of triple systems begun by Thomas P. Kirkman in 1847 and pursued by Jakob Steiner, a Swiss-born German mathematician, in the 1850s was the beginning of the theory of design. Among the earliest books devoted exclusively to combinatorics are the German mathematician Eugen Netto's *Lehrbuch der Combinatorik* (1901; "Textbook of Combinatorics") and the British mathematician Percy Alexander MacMahon's *Combinatory Analysis* (1915–16), which provide a view of combinatorial theory as it existed before 1920.

**Combinatorics during the 20th century.** Many factors have contributed to the quickening pace of development of combinatorial theory since 1920. One of these was the development of the statistical theory of the design

of experiments by the English statisticians Ronald Fisher and Frank Yates, which has given rise to many problems of combinatorial interest; the methods initially developed to solve them have found applications in such fields as coding theory. Information theory, which arose around midcentury, has also become a rich source of combinatorial problems of a quite new type.

Another source of the revival of interest in combinatorics is graph theory, the importance of which lies in the fact that graphs can serve as abstract models for many different kinds of schemes of relations among sets of objects. Its applications extend to operations research, chemistry, statistical mechanics, theoretical physics, and socioeconomic problems. The theory of transportation networks can be regarded as a chapter of the theory of directed graphs. One of the most challenging theoretical problems, the four-colour problem (see below) belongs to the domain of graph theory. It has also applications to such other branches of mathematics as group theory.

The development of computer technology in the second half of the 20th century is a main cause of the interest in finite mathematics in general and combinatorial theory in particular. Combinatorial problems arise not only in numerical analysis but also in the design of computer systems and in the application of computers to such problems as those of information storage and retrieval.

Statistical mechanics is one of the oldest and most productive sources of combinatorial problems. Much important combinatorial work has been done by applied mathematicians and physicists since the mid-20th century—for example, the work on Ising models (see below *The Ising problem*).

Use of combinatorics in pure mathematics

In pure mathematics, combinatorial methods have been used with advantage in such diverse fields as probability, algebra (finite groups and fields, matrix and lattice theory), number theory (difference sets), set theory (Sperner's theorem), and mathematical logic (Ramsey's theorem).

In contrast to the wide range of combinatorial problems and the multiplicity of methods that have been devised to deal with them stands the lack of a central unifying theory. Unifying principles and cross connections, however, have begun to appear in various areas of combinatorial theory. The search for an underlying pattern that may indicate in some way how the diverse parts of combinatorics are interwoven is a challenge that faces mathematicians in the last quarter of the 20th century.

## PROBLEMS OF ENUMERATION

**Permutations and combinations.** *Binomial coefficients.* An ordered set $a_1, a_2, \ldots, a_r$ of $r$ distinct objects selected from a set of $n$ objects is called a permutation of $n$ things taken $r$ at a time. The number of permutations is given by ${}_nP_n = n(n-1)(n-2)\cdots(n-r+1)$. When $r=n$, the number ${}_nP_r = n(n-1)(n-2)\cdots$ is simply the number of ways of arranging $n$ distinct things in a row. This expression is called factorial $n$ and is denoted by $n!$. It follows that ${}_nP_r = n!/(n-r)!$. By convention $0! = 1$.

A set of $r$ objects selected from a set of $n$ objects without regard to order is called a combination of $n$ things taken $r$ at a time. Because each combination gives rise to $r!$ permutations, the number of combinations, which is written $\binom{n}{r}$, can be expressed in terms of factorials (see Box, formula 1).

The number $\binom{n}{r}$ is called a binomial coefficient because it occurs as the coefficient of $p^r q^{n-r}$ in the binomial expansion—that is, the re-expression of $(q+p)^n$ in a linear combination of products of $p$ and $q$ (see 2).

If $0 \leqslant p \leqslant 1$, and $q = 1-p$, then the term $\binom{n}{r} p^r q^{n-r}$ in the binomial expansion is the probability that an event the chance of occurrence of which is $p$ occurs exactly $r$ times in $n$ independent trials (see PROBABILITY THEORY).

The answer to many different kinds of enumeration problems can be expressed in terms of binomial coefficients. The number of distinct solutions of the equation $x_1 + x_2 + \cdots + x_n = m$, for example, in which $m$ is a non-negative integer $m \geqslant n$ and in which only non-negative integral values of $x_i$ are allowed is expressible this way, as was found by the 17th–18th-century French-born British mathematician Abraham De Moivre (see 3).

*Multinomial coefficients.* If $S$ is a set of $n$ objects, and $n_1, n_2, \cdots, n_k$ are non-negative integers satisfying $n_1 + n_2 + \cdots + n_k = n$, then the number of ways in which the objects can be distributed into $k$ boxes, $X_1, X_2, \cdots, X_k$, such that the box $X_i$ contains exactly $n_i$ objects is given in terms of a ratio constructed of factorials (see 4). This number, called a multinomial coefficient, is the coefficient in the multinomial expansion of the $n$th power of the sum of the $\{p_i\}$ (see 5). If all of the $\{p_i\}$ are non-negative and sum to 1 and if there are $k$ possible outcomes in a trial in which the chance of the $i$th outcome is $p_i$, then the $i$th summand in the multinomial expansion is the probability that in $n$ independent trials the $i$th outcome will occur exactly $n_i$ times, for each $i$, $1 \leqslant i \leqslant k$.

**Recurrence relations and generating functions.** If $f_n$ is a function defined on the positive integers, then a relation that expresses $f_{n+k}$ as a linear combination of function values of integer index less than $n+k$, in which a fixed constant in the linear combination is written $a_i$, is called a recurrence relation (see 6). The relation together with the initial values $f_0, f_1, \cdots, f_{k-1}$ determines $f_n$ for all $n$. The function $F(x)$ constructed of a sum of products of the type $f_n x^n$, the convergence of which is assumed in the neighbourhood of the origin, is called the generating function of $f_n$ (see 7).

Generating functions

The set of the first $n$ positive integers will be written $X_n$. It is possible to find the number of subsets of $X_n$ containing no two consecutive integers, with the convention that the null set counts as one set. The required number will be written $f_n$. A subset of the required type is either a subset of $X_{n-1}$ or is obtained by adjoining $n$ to a subset of $X_{n-2}$. Therefore $f_n$ is determined by the recurrence relation $f_n = f_{n-1} + f_{n-2}$ with the initial values $f_0 = 1$, $f_1 = 2$. Thus $f_2 = 3$, $f_3 = 5$, $f_4 = 8$, and so on. The generating function $F(x)$ of $f_n$ can be calculated (see 8), and from this a formula for the desired function $f_n$ can be obtained (see 9). That $f_n = f_{n-1} + f_{n-2}$ can now be directly checked.

**Partitions.** A partition of a positive integer $n$ is a representation of $n$ as a sum of positive integers $n = x_1 + x_2 + \cdots + x_k$, $x_i \geqslant 1$, $i = 1, 2, \cdots, k$. The numbers $x_i$ are called the parts of the partition. The number of ordered partitions into $k$ parts is $\binom{n-1}{k-1}$, for this is the number of ways of putting $k-1$ separating marks in the $n-1$ spaces between $n$ dots in a row. The theory of unordered partitions is much more difficult and has many interesting features. An unordered partition can be standardized by listing the parts in a decreasing order. Thus $n = x_1 + x_2 + \cdots + x_k$, $x_1 \geqslant x_2 \geqslant \cdots \geqslant x_k \geqslant 1$. In what follows partition will mean an unordered partition.

The number of partitions of $n$ into $k$ parts will be denoted by $P_k(n)$, and a recurrence formula for it can be obtained from the definition (see 10). This recurrence formula, together with the initial conditions $P_k(n) = 0$ if $n < k$, and $P_k(k) = 1$ determines $P_k(n)$. It can be shown that $P_k(n)$ depends on the value of $n$ (mod $k!$), in which the notation $x \equiv a \pmod{b}$ means that $x$ is any number that, if divided by $b$, leaves the same remainder as $a$ does. For example, $P_3(n) = n^2 + c_n$, in which $c_n = 0, -1/12, -1/3, +1/4, -1/3$, or $-1/12$, according as $n$ is congruent to 0, 1, 2, 3, 4, or 5 (mod 6). $P(n)$, which is a sum over all values of $k$ from 1 to $n$ of $P_k(n)$, denotes the number of partitions of $n$ into $n$ or fewer parts.

Many results on partitions can be obtained by the use of Ferrers' diagram. The diagram of a partition is obtained by putting down a row of squares equal in number to the largest part, then immediately below it a row of squares equal in number to the next part, and so on. Such a diagram for $14 = 5+3+3+2+1$ is shown in Figure 1.

Ferrers' partitioning diagram

By rotating the Ferrers' diagram of the partition about the diagonal, it is possible to obtain from the partition $n = x_1 + x_2 + \cdots + x_k$ the conjugate partition $n = x_1^* + x_2^* + \cdots + x_n^*$, in which $x_i^*$ is the number of parts in the original partition of cardinality $i$ or more.

(1) $\binom{n}{r} = \frac{{}_nP_r}{r!} = \frac{n!}{r!(n-r)!}$

(2) $(q+p)^n = q^n + \binom{n}{1}pq^{n-1} + \cdots + \binom{n}{r}p^r q^{n-r} + \cdots + p^n$

(3) $N = \binom{m+n-1}{n-1} = \frac{(m+n-1)!}{(n-1)!m!}$

(4) $\binom{n}{n_1, n_2, \cdots, n_k} = \frac{n!}{n_1!n_2! \cdots n_k!}$

(5) $(p_1 + p_2 + \cdots + p_k)^n = \sum \binom{n}{n_1, n_2, \cdots, n_k} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$
The summation is over all non-negative $n_1, n_2, \cdots, n_k$ for which $n_1 + n_2 + \cdots + n_k = n$.

(6) $f_{n+k} = a_1 f_{n+k-1} + a_2 f_{n+k-2} + \cdots + a_k f_n$

(7) $F(x) = \sum_{n=0}^{\infty} f_n x^n$

(8) $F(x) = \frac{1+x}{1-x(1+x)}$

(9) $f_n = \sum_k \binom{n+1-k}{k}$
The summation extends over all values of $k$ from 0 to the largest integer not exceeding $(n+1)/2$.

(10) $P_k(n) = P_k(n-k) + P_{k-1}(n-k) + \cdots + P_1(n-k)$

Thus the conjugate of the partition of 14 already given is $14 = 5+4+3+1+1$. Hence, the following result is obtained:

($F_1$) The number of partitions of $n$ into $k$ parts is equal to the number of partitions of $n$ with $k$ as the largest part.

Other results obtainable by using Ferrers' diagrams are:

($F_2$) The number of self-conjugate partitions of $n$ equals the number of partitions of $n$ with all parts unequal and odd.

($F_3$) the number of partitions of $n$ into unequal parts is equal to the number of partitions of $n$ into odd parts.

Generating functions can be used with advantage to study partitions. For example, it can be proved that:

($G_1$) The generating function $F_1(x)$ of $P(n)$, the number of partitions of the integer $n$, is a product of reciprocals of terms of the type $(1-x^k)$, for all positive integers $k$, with the convention that $P(0) = 1$ (see 11).

($G_2$) The generating function $F_2(x)$ of the number of partitions of $n$ into unequal parts is a product of terms like $(1+x^k)$, for all positive integers $k$ (see 12).

($G_3$) The generating function $F_3(x)$ of the number of partitions of $x$ consisting only of odd parts is a product of reciprocals of terms of the type $(1-x^k)$, for all positive odd integers $k$ (see 13).

Thus to prove ($F_3$) it is necessary only to show that the generating functions described in ($G_2$) and ($G_3$) are equal. This method was used by Euler.

**The principle of inclusion and exclusion: derangements.** For a case in which there are $N$ objects and $n$ properties $A_1, A_2, \cdots A_n$, the number $N(A_1, A_2)$, for example, will be the number of objects that possess the properties $A_1, A_2$. If $N(\bar{A}_1, \bar{A}_2, \cdots, \bar{A}_n)$ is the number of objects possessing none of the properties $A_1, A_2, \cdots, A_n$, then this number can be computed as an alternating sum of sums involving the numbers of objects that possess the properties (see 14).

Sylvester principle

This is the principle of inclusion and exclusion expressed by Sylvester.

The permutation of $n$ elements that displaces each object is called a derangement. The permutations themselves may be the objects and the property $i$ may be the property that a permutation does not displace the $i$th element. In such a case $N = n!$ and $N(A_1, A_2) = (n-2)!$, for example. Hence the number $D_n$ of derangements can be shown to be approximated by $n!/e$ (see 15). This number was first obtained by Euler. If $n$ persons check their hats in a restaurant, and the waiter loses the checks and returns the hats at random, the chance that no one receives his own hat is $D_n/n! = e^{-1}$ approximately. It is surprising that the approximate answer is independent of $n$. To six places of decimals $e^{-1} = 0.367879$. When $n = 6$ the error of approximation is less than 0.0002.

If $n$ is expressed as the product of powers of its prime factors $p_1, p_2, \cdots p_k$, and if the objects are the integers less than or equal to $n$, and if $A_i$ is the property of being divisible by $p_i$, then Sylvester's formula gives, as the number of integers less than $n$ and prime to it, a function of $n$, written $\varphi(n)$, composed of a product of $n$ and $k$ factors of the type $(1 - 1/p_i)$ (see 16). The function $\varphi(n)$ is the Euler function.

**Polya's theorem.** It is required to make a necklace of $n$ beads out of an infinite supply of beads of $k$ different colours. The number of different necklaces, $c(n, k)$, that can be made is given by the reciprocal of $n$ times a sum of terms of the type $\varphi(n)\, k^{n/d}$, in which the summation is over all divisors $d$ of $n$ and $\varphi$ is the Euler function (see 17).

Though the problem of the necklaces appears to be frivolous, the formula given above can be used to solve a difficult problem in the theory of Lie algebras, of some importance in modern physics.

The general problem of which the necklace problem is a special case was solved by the Hungarian-born U.S. mathematician George Polya in a famous 1937 memoir in which he established connections between groups, graphs, and chemical bonds. It has been applied to enumeration problems in physics, chemistry, and mathematics.

**The Möbius inversion theorem.** In 1832 the German astronomer and mathematician August Ferdinand Möbius proved that, if $f$ and $g$ are functions defined on the set of positive integers, such that $f$ evaluated at $x$ is a sum of values of $g$ evaluated at divisors of $x$, then inversely $g$ at $x$ can be evaluated as a sum involving $f$ evaluated at divisors of $x$ (see 18).

In 1964 the U.S. mathematician Gian-Carlo Rota obtained a powerful generalization of this theorem, providing a fundamental unifying principle of enumeration. One consequence of Rota's theorem is the following:

Rota's theorem

If $f$ and $g$ are functions defined on subsets of a finite set $A$, such that $f(A)$ is a sum of terms $g(S)$, in which $S$ is a subset of $A$, then $g(A)$ can be expressed in terms of $f$ (see 19).

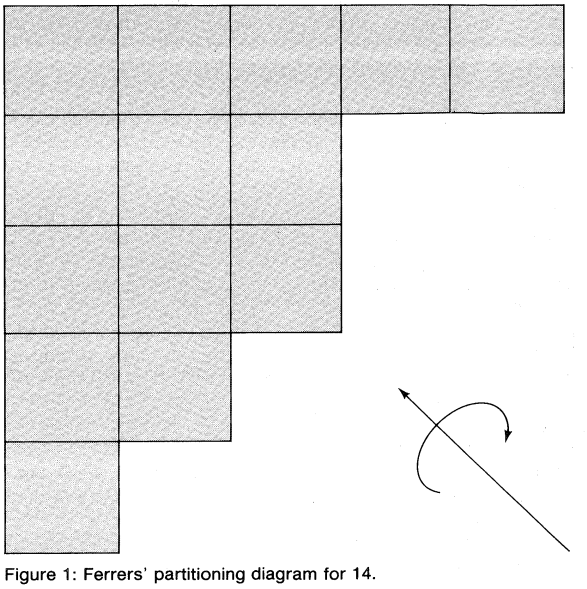


Figure 1: Ferrers' partitioning diagram for 14.

$$(11)\quad F_1(x)=(1-x)^{-1}(1-x^2)^{-1}(1-x^3)^{-1}\cdots$$

$$(12)\quad F_2(x)=(1+x)(1+x^2)(1+x^3)\cdots$$

$$(13)\quad F_3(x)=(1-x)^{-1}(1-x^3)^{-1}(1-x^5)^{-1}\cdots$$

$$(14)\quad \begin{cases} N(\overline{A}_1,\overline{A}_2,\cdots,\overline{A}_n) \\ \quad =N-\sum N(A_{i_1})+\sum N(A_{i_1},A_{i_2})+\cdots+ \\ \quad +(-1)^k\sum N(A_{i_1},A_{i_2},\cdots,A_{i_k})+\cdots+ \\ \quad +(-1)^n N(A_1,A_2,\cdots,A_n) \\ \text{In the general term the summation is over all combinations of } k \text{ properties from the set of } n \\ \text{properties} \quad A_1,A_2,\cdots,A_n. \end{cases}$$

$$(15)\quad \begin{cases} D_n=n!-\binom{n}{1}(n-1)!+\cdots+(-1)^k\binom{n}{k}(n-k)!+ \\ \quad +\cdots+(-1)^n\binom{n}{n} \\ \quad =n!\left(1-\frac{1}{1!}+\frac{1}{2!}+\cdots+(-1)^k\frac{1}{k!}+\right. \\ \quad \left.+\cdots+(-1)^n\frac{1}{n!}\right) \\ \quad =n!/e \quad \text{approximately.} \end{cases}$$

$$(16)\quad \phi(n)=n\left(1-\frac{1}{p_1}\right)\left(1-\frac{1}{p_2}\right)\cdots\left(1-\frac{1}{p_k}\right)$$

$$(17)\quad c(n,k)=\frac{1}{n}\sum_{d|n}\phi(d)k^{n/d}$$

$$(18)\quad \begin{cases} \text{If} \quad f(x)=\sum_{d|x} g(d), \\ \text{in which } d|x \text{ means that } d \text{ is a divisor of } x, \text{ then} \\ g(x)=\sum_{d|x}\mu(d,x)f(d), \\ \text{in which} \\ \mu(d,n)=\begin{cases} 1 & \text{if } n=d \\ (-1)^k & \text{if } n=p_1p_2\cdots p_kd \\ 0 & \text{otherwise.} \end{cases} \end{cases}$$

**Special problems.** Despite the general methods of enumeration already described, there are many problems in which they do not apply and which therefore require special treatment. Two of these are described below, and others will be met further in this article.

*The Ising problem.* A rectangular $m \times n$ grid is made up of unit squares, each coloured either red or green. How many different colour patterns are there if the number of boundary edges between red squares and green squares is prescribed?

This problem, though easy to state, proved very difficult to solve. A complete and rigorous solution was not achieved until the early 1960s. The importance of the problem lies in the fact that it is the simplest model that exhibits the macroscopic behaviour expected from certain natural assumptions made at the microscopic level. Historically, the problem arose from an early attempt, made in 1925, to formulate the statistical mechanics of ferromagnetism.

The three-dimensional analogue of the Ising problem remains unsolved in spite of persistent attacks.

*Non-self-intersecting random walk.* A random walk consists of a sequence of $n$ steps of unit length on a flat rectangular grid, taken at random either in the $x$- or the $y$-direction, with equal probability in each of the four directions. What is the number $R_n$ of random walks that do not touch the same vertex twice? This problem has defied solution, except for small values of $n$, though a large amount of numerical data has been amassed.

## PROBLEMS OF CHOICE

**Systems of distinct representatives.** Subsets $S_1, S_2, \cdots, S_n$ of a finite set $S$ are said to possess a set of distinct representatives if $x_1, x_2, \cdots, x_n$ can be found, such that $x_i \in S_i$, $i=1, 2, \cdots, n$, $x_i \neq x_j$ for $i \neq j$. It is possible that $S_i$ and $S_j$, $i \neq j$, may have exactly the same elements and are distinguished only by the indices $i, j$. In 1935 a mathematician, M. Hall, Jr., of the United States, proved that a necessary and sufficient condition for $S_1, S_2, \cdots, S_n$ to possess a system of distinct representatives is that, for every $k \leq n$, any $k$ of the $n$ subsets contain between them at least $k$ distinct elements. Hall's and König's theorems

For example, the sets $S_1=(1, 2, 2)$, $S_2=(1, 2, 4)$, $S_3=(1, 2, 5)$, $S_4=(3, 4, 5, 6)$, $S_5=(3, 4, 5, 6)$ satisfy the conditions of the theorem, and a set of distinct representatives is $x_1=1$, $x_2=2$, $x_3=5$, $x_4=3$, $x_5=4$. On the other hand, the sets $T_1=(1, 2)$, $T_2=(1, 3)$, $T_3=(1, 4)$, $T_4=(2, 3)$, $T_5=(2, 4)$, $T_6=(1, 2, 5)$ do not possess a system of distinct representatives because $T_1, T_2, T_3, T_4, T_5$ possess between them only four elements.

The following theorem due to König is closely related to Hall's theorem and can be easily deduced from it. Conversely, Hall's theorem can be deduced from König's: If the elements of rectangular matrix are 0s and 1s, the minimum number of lines that contain all of the 1s is equal to the maximum number of 1s that can be chosen with no two on a line.

**Ramsey's numbers.** If $X=\{1, 2, \ldots, n\}$, and if $T$, the family of all subsets of $X$ containing exactly $r$ distinct elements, is divided into two mutually exclusive families $\alpha$ and $\beta$, the following conclusion that was originally obtained by the British mathematician Frank Plumpton Ramsey follows. He proved that for $r \geq 1$, $p \geq r$, $q \geq r$, there exists a number $N_r(p, q)$ depending solely on $p, q, r$, such that if $n > N_r(p, q)$, there is either a subset $A$ of $p$ elements all of the $r$ subsets of which are in the family $\alpha$ or there is a subset $B$ of $q$ elements all of the $r$ subsets of which are in the family $\beta$.

The set $X$ can be a set of $n$ persons. For $r=2$, $T$ is the family of all pairs. If two persons have met each other the pair can belong to the family $\alpha$. If two persons have not met, the pair can belong to the family $\beta$. If these things are assumed, then by Ramsey's theorem, for any given $p \geq 2$, $q \geq 2$ there exists a number $N_2(p, q)$, such that if $n > N_2(p, q)$, then among $n$ persons invited to a party there will be either a set of $p$ persons all of whom have met each other or a set of $q$ persons no two of whom have met.

Although the existence of $N_r(p, q)$ is known, actual values are known only for a few cases. Because $N_r(p, q)=N_r(q, p)$, it is possible to take $p \leq q$. It is known that $N_2(3, 3)=6$, $N_2(3, 4)=9$, $N_2(3, 5)=14$, $N_2(3, 6)=18$, $N_2(4, 4)=18$. Some bounds are also known, for example, $25 \leq N_2(4, 5) \leq 28$.

A consequence of Ramsey's theorem is the following result obtained in 1935 by the Hungarian mathematicians Paul Erdös and George Szekeres. For a given integer $n$ there exists an integer $N=N(n)$, such that a set of any $N$ points on a plane, no three on a line, contains $n$ points forming a convex $n$-gon.

## DESIGNS, LATIN SQUARES, ARRAYS, AND CODING

**BIB (balanced incomplete block) designs.** A design is a set of $T=\{1, 2, \ldots, \upsilon\}$ objects called treatments and a family of subsets $B_1, B_2, \ldots, B_b$ of $T$, called blocks, such that the block $B_i$ contains exactly $k_i$ treatments, all distinct. The number $k_i$ is called the size of the block $B_i$, and the $i$th treatment is said to be replicated $r_i$ times if it occurs in exactly $r_i$ blocks. Specific designs are subject to further constraints. The name design comes from statistical theory in which designs are used to estimate effects of treatments applied to experimental units.

A BIB design is a design with $\upsilon$ treatments and $b$ blocks in which each block is of size $k$, each treatment is replicated $r$ times, and every pair of distinct treatments occurs together in $\lambda$ blocks. The design is said to have the parameters

$(\upsilon, b, r, k, \lambda)$. Some basic relations are easy to establish (see 20). These conditions are necessary but not sufficient for the existence of the design. The design is said to be proper if $k < \upsilon$—that is, the blocks are incomplete. For a proper BIB design Fisher's inequality $b \geqslant \upsilon$, or equivalently $r \geqslant k$, holds.

BIB design symmetry

A BIB design is said to be symmetric if $\upsilon = b$, and consequently $r = k$. Such a design is called a symmetric $(\upsilon, k, \lambda)$ design, and $\lambda(\upsilon - 1) = k(k-1)$. A necessary condition for the existence of a symmetric $(\upsilon, k, \lambda)$ design is given by the following:

A. If $\upsilon$ is even, $k - \lambda$ is a perfect square.

B. If $\upsilon$ is odd, a certain Diophantine equation (see 21) has a solution in integers not all zero.

For example, the designs $(\upsilon, k, \lambda) = (22, 7, 2)$ and $(46, 10, 2)$ are ruled out by (A) and the design $(29, 8, 2)$ by (B).

Because necessary and sufficient conditions for the existence of a BIB design with given parameters are not known, it is often a very difficult problem to decide whether a design with given parameters (satisfying the known necessary conditions) really exists. By 1972 there were only two unsettled cases with $r \leqslant 10$. These are $(\upsilon, b, r, k, \lambda) = (46, 69, 9, 6, 1)$ and $(51, 85, 10, 6, 1)$.

Methods of constructing BIB designs depend on the use of finite fields, finite geometries, and number theory. Some general methods were given in 1939 by the Indian mathematician Raj Chandra Bose, who has since emigrated to the United States.

A finite field is a finite set of marks with two operations, addition and multiplication, subject to the usual nine laws of addition and multiplication obeyed by rational numbers. In particular the marks may be taken to be the set $X$ of non-negative integers less than a prime $p$. If this is so, then addition and multiplication are defined by modified addition and multiplication laws (see 22) in which $a, b, r,$ and $p$ belong to $X$. For example, if $p = 7$, then $5 + 4 = 2$, $5 \cdot 4 = 6$. There exist more general finite fields in which the number of elements is $p^n$, $p$ a prime. There is essentially one field with $p^n$ elements, with given $p$ and $n$. It is denoted by $GF(p^n)$.

Finite geometries can be obtained from finite fields in which the coordinates of points are now elements of a finite field.

A set of $k+1$ non-negative integers $d_0, d_1, \cdots, d_k$, is said to form a perfect difference set mod $\upsilon$, if among the $k(k-1)$ differences $d_i - d_j$, $i \neq j$, $i, j = 0, 1, \cdots, k$, reduced mod $\upsilon$, each nonzero positive integer less than $\upsilon$ occurs exactly the same number of times $\lambda$. For example, 1, 4, 5, 9, 3 is a difference set mod 11, with $\lambda = 2$. From a perfect difference set can be obtained the symmetric $(\upsilon, k, \lambda)$ design using the integers $0, 1, 2, \cdots, \upsilon - 1$. The $j$th block contains the treatments obtained by reducing mod $\upsilon$ the numbers $d_0 + j, d_1 + j, \cdots, d_i + j, j = 0, 1, \cdots, \upsilon - 1$.

It can be shown that any two blocks of a symmetric $(\upsilon, k, \lambda)$ design intersect in exactly $k$ treatments. By deleting one block and all the treatments contained in it, it is possible to obtain from the symmetric design its residual, which is a BIB design (unsymmetric) with parameters $\upsilon^* = \upsilon - k$, $b^* = \upsilon - 1$, $r^* = k$, $k^* = k - \lambda$, $\lambda^* = \lambda$. One may ask whether it is true that a BIB design with the parameters of a residual can be embedded in a symmetric BIB design. The truth of this is rather easy to demonstrate when $\lambda = 1$. Hall and W.S. Connor in 1953 showed that it is also true for $\lambda = 2$. The Indian mathematician K.N. Bhattacharya in 1944, however, gave a counterexample for $\lambda = 3$ by exhibiting a BIB design with parameters $\upsilon = 16$, $b = 24$, $r = 9$, $k = 6$, $\lambda = 3$ for which two particular blocks intersect in four treatments and which for that reason cannot be embedded in a symmetric BIB design.

A BIB design is said to be resolvable if the set of blocks can be partitioned into subsets, such that the blocks in any subset contain every treatment exactly once. For the case $k = 3$ this problem was first posed during the 19th century by the British mathematician T.P. Kirkman as a recreational problem. There are $\upsilon$ girls in a class. Their teacher wants to take the class out for a walk for a number of days, the girls marching abreast in triplets. It is required to arrange the walk so that any two girls march abreast in the same triplet exactly once. It is easily shown that this is equivalent to the construction of a resolvable BIB design with $\upsilon = 6t + 3$, $b = (2t+1)(3t+1)$, $r = 3t + 1$, $k = 3$, $\lambda = 1$. Solutions were known for only a large number of special values of $t$ until a completely general solution was finally given by the Indian and U.S. mathematicians Dwijendra K. Ray-Chaudhuri and R.M. Wilson in 1970.

**PBIB (partially balanced incomplete block) designs.** Given $\upsilon$ objects $1, 2, \cdots, \upsilon$, a relation satisfying the following conditions is said to be an $m$-class partially balanced association scheme:

Conditions for PBIB designs

A. Any two objects are either 1st, or 2nd, $\cdots$, or $m$th associates, the relation of association being symmetrical.

B. Each object $\alpha$ has $n_i$ $i$th associates, the number $n_i$ being independent of $\alpha$.

C. If any two objects $\alpha$ and $\beta$ are $i$th associates, then the number of objects that are $j$th associates of $\alpha$ and $k$th associates of $\beta$ is $p_{jk}^{\,i}$ and is independent of the pair of $i$th associates $\alpha$ and $\beta$.

The constants $\upsilon$, $n_i$, $p_{jk}^{\,i}$ are the parameters of the association scheme. A number of identities connecting these parameters were given by the Indian mathematicians Bose and K.R. Nair in 1939, but Bose and the U.S. mathematician D.M. Mesner in 1959 discovered new identities when $m > 2$.

A PBIB design is obtained by identifying the $\upsilon$ treatments with the $\upsilon$ objects of an association scheme and arranging them into $b$ blocks satisfying the following conditions:

A. Each contains $k$ treatments.

B. Each treatment occurs in $r$ blocks.

C. If two treatments are $i$th associates, they occur together in $\lambda_i$ blocks.

Two-class association schemes and the corresponding designs are especially important both from the mathematical point of view and because of statistical applications. For a two-class association scheme the constancy of $\upsilon$, $n_i$, $p_{11}^{\,1}$, and $p_{11}^{\,2}$ ensures the constancy of the other parameters. Seven relations hold (see 23). Sufficient conditions for the existence of association schemes with given parameters are not known, but for a two-class association scheme Connor and the U.S. mathematician Willard H. Clatworthy in 1954 obtained some necessary conditions (see 24).

$$(19)\quad \begin{cases} \text{If} & f(A) = \sum_{S \subset A} g(S), \\ \text{then} & g(A) = \sum_{S \subset A} (-1)^{|A|-|S|} f(S). \end{cases}$$

$$(20)\quad bk = \upsilon r, \qquad \lambda(\upsilon - 1) = r(k-1)$$

$$(21)\quad x^2 = (k-\lambda)y^2 + (-1)^{(\upsilon-1)/2}\lambda z^2$$

$$(22)\quad a + b = s \pmod p, \qquad ab = c \pmod p$$

$$(23)\quad \begin{cases} p_{12}^1 = n_1 - p_{11}^1 - 1 = p_{21}^1, & p_{22}^1 = n_2 - n_1 + p_{11}^1 + 1 \\ p_{12}^2 = n_1 - p_{11}^2 = p_{21}^2, & p_{22}^2 = n_2 - n_1 + p_{11}^2 \\ \upsilon = n_1 + n_2 + 1, & n_1 p_{12}^1 = n_2 p_{11}^2, \qquad n_1 p_{22}^1 = n_2 p_{12}^2 \end{cases}$$

$$(24)\quad \begin{cases} \alpha_1, \alpha_2 = \dfrac{n_1 + n_2}{2} \pm \dfrac{(n_1 - n_2) + \gamma(n_1 + n_2)}{2\sqrt{\Delta}} \\ \text{These numbers must be non-negative integers, in which} \\ \gamma = p_{12}^2 - p_{12}^1, \qquad \Delta = (p_{12}^2 - p_{12}^1)^2 + 2(p_{12}^2 + p_{12}^1) + 1. \end{cases}$$

$$(25)\quad \text{If } k = p_1^{n_1} p_2^{n_2} \cdots p_u^{n_u},\ n(k) = \min(p_1^{n_1}, p_2^{n_2}, \cdots, p_u^{n_u})$$

**Orthogonal Latin squares.** A Latin square of order $k$ is defined as a $k \times k$ square grid, the $k^2$ cells of which are occupied by $k$ distinct symbols of a set $X = 1, 2, \ldots, k$, such that each symbol occurs once in each row and each column. Two Latin squares are said to be orthogonal if, when superposed, any symbol of the first square occurs exactly once with each symbol of the second square.

Two orthogonal Latin squares of order 4 are exhibited in Figure 2.

A set of mutually orthogonal Latin squares is a set of Latin squares any two of which are orthogonal. It is easily shown that there cannot exist more than $k-1$ mutually orthogonal Latin squares of a given order $k$. When $k-1$ mutually orthogonal Latin squares of order $k$ exist, the set is complete. A complete set always exists if $k$ is the power of a prime. An unsolved question is whether there can exist a complete set of mutually orthogonal Latin squares of order $k$ if $k$ is not a prime power.

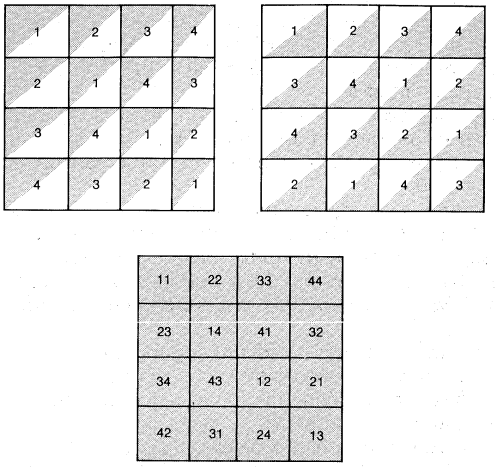



Figure 2: Two orthogonal Latin squares of order 4 and their superposition.

Many types of experimental designs are based on Latin squares. Hence, the construction of mutually orthogonal Latin squares is an important combinatorial problem. Letting the prime power decomposition of an integer $k$ be given, the arithmetic function $n(k)$ is defined by taking the minimum of the factors in such a decomposition (see 25).

Letting $N(k)$ denote the maximum number of mutually orthogonal Latin squares of order $k$, the U.S. mathematician H.F. MacNeish in 1922 showed that there always exist $n(k)$ mutually orthogonal Latin squares of order $k$ and conjectured that this is the maximum number of such squares—that is, $N(k)=n(k)$. There was also the long-standing conjecture of Euler, formulated in 1782, that there cannot exist mutually orthogonal Latin squares of order $4t+2$, for any integer $t$. MacNeish's conjecture, if true, would imply the truth of Euler's but not conversely. The U.S. mathematician E.T. Parker in 1958 disproved the conjecture of MacNeish. This left open the question of Euler's conjecture. Bose and the Indian mathematician S.S. Shrikhande in 1959–60 obtained the first counterexample to Euler's conjecture by obtaining two mutually orthogonal Latin squares of order 22 and then generalized their method to disprove Euler's conjecture for an infinity of values of $k=2(\text{mod } 4)$. Parker in 1959 used the method of differences to show the falsity of Euler's conjecture for all $k=(3q+1)/2$, in which $q$ is a prime power, $q\equiv 3(\text{mod } 4)$. Finally these three mathematicians in 1960 showed that $N(k)\geq 2$ whenever $k>6$. It is pertinent to inquire about the behaviour of $N(k)$ for large $k$. The best result in this direction is due to R.M. Wilson in 1971. He shows that $N(k)\geq k^{1/17}-2$ for large $k$.

Euler's conjecture

**Orthogonal arrays and the packing problem.** A $k\times N$ matrix $A$ with entries from a set $X$ of $s\geq 2$ symbols is called an orthogonal array of strength $t$, size $N$, $k$ constraints, and $s$ levels if each $t\times N$ submatrix of $A$ contains all possible $t\times 1$ column vectors with the same frequency $\lambda$. The array may be denoted by $(N, k, s, t)$. The number $\lambda$ is called the index of the array, and $N=\lambda s^t$. This concept is due to the Indian mathematician C.R. Rao and was obtained in 1947.

Orthogonal arrays are a generalization of orthogonal Latin squares. Indeed, the existence of an orthogonal array of $k$ constraints, $s$ levels, strength 2, and index unity is combinatorially equivalent to the existence of a set of $k-2$ mutually orthogonal Latin squares of order $s$. For a given $\lambda$, $s$, and $t$ it is an important combinatorial problem to obtain an orthogonal array $(N, k, s, t)$, $N=s^t$, for which the number of constraints $k$ is maximal.

Orthogonal arrays play an important part in the theory of factorial designs in which each treatment is a combination of factors at different levels. For an orthogonal array $(\lambda s^t, k, s, t)$, $t\geq 2$, the number of constraints $k$ satisfies an inequality (see 26) in which $\lambda s^t$ is greater than or equal to a linear expression in powers of $(s-1)$, with binomial coefficients giving the number of combinations of $k-1$ or $k$ things taken $i$ at a time $(i\leq u)$.

Letting $GF(q)$ be a finite field with $q=p^n$ elements, an $n\times r$ matrix with elements from the field is said to have the property $P_t$ if any $t$ rows are independent. The problem is to construct for any given $r$ a matrix $H$ with the maximum number of rows possessing the property $P_t$. The maximal number of rows is denoted by $n_t(r, q)$. This packing problem is of great importance in the theory of factorial designs and also in communication theory, because the existence of an $n\times r$ matrix with the property $P_t$ leads to the construction of an orthogonal array $(q^r, n, q, t)$ of index unity.

Again $n\times r$ matrices $H$ with the property $P_t$ may be used in the construction of error-correcting codes. A row vector $c'$ is taken as a code word if and only if $c'H=0$. The code words then are of length $n$ and differ in at least $t+1$ places. If $t=2u$, then $u$ or fewer errors of transmission can be corrected if such a code is used. If $t=2u+1$, then an additional error can be detected.

A general solution of the packing problem is known only for the case $t=2$, the corresponding codes being the one-error-correcting codes of the U.S. mathematician Richard W. Hamming. When $t=3$ the solution is known for general $r$ when $q=2$ and for general $q$ when $r\leq 4$. Thus, $n_2(r, 2)=(q^r-1)/(q-1)$, $n_3(r, 2)=2^{r-1}$, $n_3(3, q)=q+1$ or $q+2$, according as $q$ is odd or even. If $q>2$, then $n_3(4, q)=q^2+1$. The case $q=2$ is especially important because in practice most codes use only two symbols, 0 or 1. Only fairly large values of $r$ are useful, say, $r\geq 25$. The optimum value of $n_t(r, 2)$ is not known. The BCH codes obtained by Bose and Ray-Chaudhuri and independently by the French mathematician Alexis Hocquenghem in 1959 and 1960 are based on a construction that yields an $n\times r$ matrix $H$ with the property $P_{2u}$ in which $r\leq mu$, $n=2^m-1$, $q=2$. They can correct up to $u$ errors.

BCH codes

## GRAPH THEORY

**Definitions.** A graph $G$ consists of a non-empty set of elements $V(G)$ and a subset $E(G)$ of the set of unordered pairs of distinct elements of $V(G)$. The elements of $V(G)$, called vertices of $G$, may be represented by points. If $(x, y)\in E(G)$, then the edge $(x, y)$ may be represented by an arc joining $x$ and $y$. Then $x$ and $y$ are said to be adjacent, and the edge $(x, y)$ is incident with $x$ and $y$. If $(x, y)$ is not an edge, then the vertices $x$ and $y$ are said to be nonadjacent. $G$ is a finite graph if $V(G)$ is finite. A graph $H$ is a subgraph of $G$ if $V(H)\subset V(G)$ and $E(H)\subset E(G)$.

A chain of a graph $G$ is an alternating sequence of vertices and edges $x_0, e_1, x_1, e_2, \cdots e_n, x_n$, beginning and ending with vertices in which each edge is incident with the two vertices immediately preceding and following it. This chain joins $x_0$ and $x_n$ and may also be denoted by $x_0, x_1, \cdots, x_n$, the edges being evident by context. The chain is closed if $x_0=x_n$ and open otherwise. If the chain is closed, it is called a cycle, provided its vertices (other than $x_0$ and $x_n$) are distinct and $n\geq 3$. The length of a chain is the number of edges in it.

A graph $G$ is labelled when the various $\upsilon$ vertices are distinguished by such names as $x_1, x_2, \cdots x_\upsilon$. Two graphs $G$ and $H$ are said to be isomorphic (written $G\simeq H$) if there exists a one-one correspondence between their vertex sets that preserves adjacency. For example, $G_1$ and $G_2$, shown in Figure 3, are isomorphic under the correspondence $x_i\leftrightarrow y_i$.

Two isomorphic graphs count as the same (unlabelled) graph. A graph is said to be a tree if it contains no cycle—for example, the graph $G_3$ of Figure 3.

**Enumeration of graphs.** The number of labelled graphs with $\upsilon$ vertices is $2^{\upsilon(\upsilon-1)/2}$ because $\upsilon(\upsilon-1)/2$ is the number

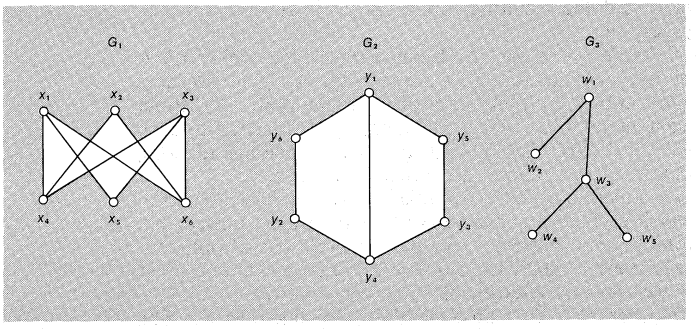



Figure 3: Two isomorphic graphs and a tree.

of pairs of vertices, and each pair is either an edge or not an edge. Cayley in 1889 showed that the number of labelled trees with $\upsilon$ vertices is $\upsilon^{\upsilon-2}$.

The number of unlabelled graphs with $\upsilon$ vertices can be obtained by using Polya's theorem. The first few terms of the generating $F(x)$, in which the coefficient of $x^{\upsilon}$ gives the number of (unlabelled) graphs with $\upsilon$ vertices, can be given (see 27).

A rooted tree has one point, its root, distinguished from others. If $T_{\upsilon}$ is the number of rooted trees with $\upsilon$ vertices, the generating function for $T_{\upsilon}$ can also be given (see 28).

Polya in 1937 showed in his memoir already referred to that the generating function for rooted trees satisfies a functional equation (see 29). Letting $t_{\upsilon}$ be the number of (unlabelled) trees with $\upsilon$ vertices, the generating function $t(x)$ for $t_{\upsilon}$ can be obtained in terms of $T(x)$ (see 30). This result was obtained in 1948 by the U.S. mathematician Richard R. Otter.

Many enumeration problems on graphs with specified properties can be solved by the application of Polya's theorem and a generalization of it made by a Dutch mathematician, N.G. de Bruijn, in 1959.

**Characterization problems of graph theory.** If there is a class $C$ of graphs each of which possesses a certain set of properties $P$, then the set of properties $P$ is said to characterize the class $C$, provided every graph $G$ possessing the properties $P$ belongs to the class $C$. Sometimes it happens that there are some exceptional graphs that possess the properties $P$. Many such characterizations are known. Here is presented a typical example.

A complete graph $K_m$ is a graph with $m$ vertices, any two of which are adjacent. The line graph $H$ of a graph $G$ is a graph the vertices of which correspond to the edges of $G$, any two vertices of $H$ being adjacent if and only if the corresponding edges of $G$ are incident with the same vertex of $G$.

A graph $G$ is said to be regular of degree $n_1$ if each vertex is adjacent to exactly $n_1$ other vertices. A regular graph of degree $n_1$ with $\upsilon$ vertices is said to be strongly regular with parameters $(\upsilon, n_1, p_{11}^1, p_{11}^2)$ if any two adjacent vertices are both adjacent to exactly $p_{11}^1$ other vertices and any two nonadjacent vertices are both adjacent to exactly $p_{11}^2$ other vertices. A strongly regular graph and a two-class association are isomorphic concepts. The treatments of the scheme correspond to the vertices of the graph, two treatments being either first associates or second associates according as the corresponding vertices are either adjacent or nonadjacent.

It is easily proved that the line graph $T_2(m)$ of a complete graph $K_m$, $m \geq 4$ is strongly regular with parameters $\upsilon = m(m-1)/2$, $n_1 = 2(m-2)$, $p_{11}^1 = m-2$, $p_{11}^2 = 4$.

It is surprising that these properties characterize $T_2(m)$ except for $m = 8$, in which case there exist three other strongly regular graphs with the same parameters nonisomorphic to each other and to $T_2(m)$.

Partial geometries

A partial geometry $(r, k, t)$ is a system of two kinds of objects, points and lines, and a relation incidence obeying the following axioms.

1. Any two points are incident with not more than one line.
2. Each point is incident with $r$ lines.
3. Each line is incident with $k$ points.
4. Given a point $P$ not incident with a line $\ell$, there are exactly $t$ lines incident with $P$ and also with some point of $\ell$.

A graph $G$ in a partial geometry is obtained by taking the points of the geometry as vertices of $G$, two vertices of $G$ being adjacent if and only if the corresponding points are incident with the same line of the geometry. It is strongly regular with parameters

$$\upsilon = k[(r-1)(k-1)+t]/t,\ n_1 = r(k-1),$$
$$p_{11}^1 = (r-1)(t-1)+(k-2),\ p_{11}^2 = rt.$$

The question of whether a strongly regular graph with the above parameters is the graph of some partial geometry is of interest. It was shown by Bose in 1963 that the answer is in the affirmative if a certain condition holds (see 31). Not much is known about the case if this condition is not satisfied, except for certain values of $r$ and $t$. For example, $T_2(m)$ is isomorphic with the graph of a partial geometry $(2, m-1, 2)$. Hence, for $m > 8$ its characterization is a consequence of the above theorem. Another consequence is the following:

Given a set of $k$-1-$d$ mutually orthogonal set of Latin squares of order $k$, the set can be extended to a complete set of $k$-1 mutually orthogonal squares if a condition holds (see 32). The case $d = 2$ is due to Shrikhande in 1961 and the general result to the U.S. mathematician Richard H. Bruck in 1963.

**Planar graphs.** A graph $G$ is said to be planar if it can be represented on a plane in such a fashion that the vertices are all distinct points, the edges are simple curves, and no two edges meet one another except at their terminals. For example, $K_4$, the complete graph on four vertices, is planar, as Figure 4A shows.

Two graphs are said to be homeomorphic if both can be obtained from the same graph by subdivisions of edges. For example, the graphs in Figure 4A and Figure 4B are homeomorphic.

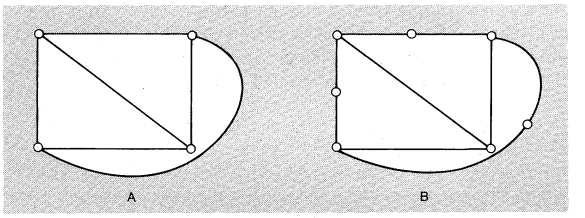



Figure 4: Two homeomorphic graphs A and B.

The $K_{m,n}$ graph is a graph for which the vertex set can be divided into two subsets, one with $m$ vertices and the other with $n$ vertices. Any two vertices of the same subset are nonadjacent, whereas any two vertices of different subsets are adjacent. The Polish mathematician Kazimierz Kuratowski in 1930 proved the following famous theorem:

A necessary and sufficient condition for a graph $G$ to be planar is that it does not contain a subgraph homeomorphic to either $K_5$ or $K_{3,3}$ shown in Figure 5.

An elementary contraction of a graph $G$ is a transformation of $G$ to a new graph $G_1$, such that two adjacent vertices $u$ and $\upsilon$ of $G$ are replaced by a new vertex $w$ in $G_1$ and $w$ is adjacent in $G_1$ to all vertices to which either $u$ or $\upsilon$ is adjacent in $G$. A graph $G^*$ is said to be a contraction of $G$ if $G^*$ can be obtained from $G$ by a sequence of elementary contractions. The following is another characterization of a planar graph due to the German mathematician K. Wagner in 1937.

A graph is planar if and only if it is not contractible to $K_5$ or $K_{3,3}$.

**The four-colour map problem.** For more than a century the solution of the four-colour map problem eluded every analyst who attempted it. The problem may have attracted the attention of Möbius, but the first written reference to it seems to be a letter from one Francis Guthrie to his brother, a student of Augustus De Morgan, in 1852.

The problem concerns planar maps—that is, subdivisions of the plane into nonoverlapping regions bounded by simple closed curves. In geographical maps it has been observed empirically, in as many special cases as have

$$(26)\quad \begin{cases} \lambda s^t \geq 1+\binom{k}{1}(s-1)+\cdots+\binom{k}{u}(s-1)^u, & \text{if } t=2u \\ \lambda s^t \geq 1+\binom{k}{1}(s-1)+\cdots+\binom{k}{u}(s-1)^u+ \\ \quad +\binom{k-1}{u}(s-1)^{u+1}, & \text{if } t=2u+1. \end{cases}$$

$$(27)\quad \begin{cases} F(x)=1+x+2x^2+4x^3+11x^4+34x^5+156x^6+ \\ \quad +1{,}044x^7+12{,}346x^8+308{,}708x^9+\cdots \end{cases}$$

$$(28)\quad T(x)=\sum_{v=1}^{\infty} T_v x^v = x\prod_{r=1}^{\infty}(1-x^r)^{-T_r}$$

$$(29)\quad T(x)=x\exp\sum_{r=1}^{\infty}\frac{1}{r}T(x^r)$$

$$(30)\quad t(x)=T(x)-\tfrac{1}{2}[T^2(x)-T(x^2)]$$

$$(31)\quad k>\tfrac{1}{2}[r(r-1)+t(r+1)(r^2-2r+2)]$$

$$(32)\quad k>\tfrac{1}{2}(d-1)(d^3-d^2+d+2)$$

been tried, that, at most, four colours are needed in order to colour the regions so that two regions that share a common boundary are always coloured differently, and in certain cases that at least four colours are necessary. (Regions that meet only at a point, such as the states of Colorado and Arizona in the United States, are not considered to have a common boundary). A formalization of this empirical observation constitutes what is called "the four-colour theorem." The problem is to prove or disprove the assertion that this is the case for every planar map. That three colours will not suffice is easily demonstrated, whereas the sufficiency of five colours was proved in 1890 by the British mathematician P.J. Heawood.

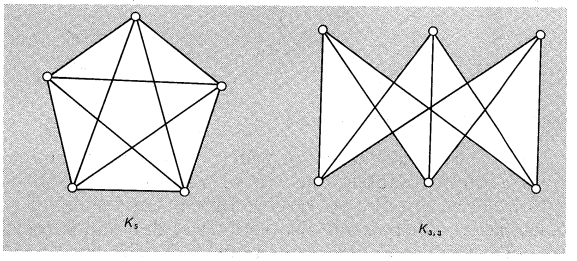



Figure 5: Two graphs important to planar properties.

In 1879 A.B. Kempe, an Englishman, proposed a solution of the four-colour problem. Although Heawood showed that Kempe's argument was flawed, two of its concepts proved fruitful in later investigation. One of these, called unavoidability, correctly states the impossibility of constructing a map in which every one of four configurations is absent (these configurations consist of a region with two neighbours, one with three, one with four, and one with five). The second concept, that of reducibility, takes its name from Kempe's valid proof that if there is a map that requires at least five colours and that contains a region with four (or three or two) neighbours, then there must be a map requiring five colours for a smaller number of regions. Kempe's attempt to prove the reducibility of a map containing a region with five neighbours was erroneous, but it was rectified in a proof published in 1976 by Kenneth Appel and Wolfgang Haken of the United States. Their proof attracted some criticism because it necessitated the evaluation of 1,936 distinct cases, each involving as many as 500,000 logical operations. Appel, Haken, and their collaborators devised programs that made it possible for a large digital computer to handle these details. The computer required more than 1,000 hours to perform the task, and the resulting formal proof is several hundred pages long.

**Eulerian cycles and the Königsberg bridge problem.** A multigraph $G$ consists of a non-empty set $V(G)$ of vertices and a subset $E(G)$ of the set of unordered pairs of distinct elements of $V(G)$ with a frequency $f \geq 1$ attached to each pair. If the pair $(x_1, x_2)$ with frequency $f$ belongs to $E(G)$, then vertices $x_1$ and $x_2$ are joined by $f$ edges.

An Eulerian cycle of a multigraph $G$ is a closed chain in which each edge appears exactly once. Euler showed that a multigraph possesses an Eulerian cycle if and only if it is connected (apart from isolated points) and the number of vertices of odd degree is either zero or two.

This problem first arose in the following manner. The Pregel River, formed by the confluence of its two branches, runs through the town of Königsberg and flows on either side of the island of Kneiphof. There were seven bridges, as shown in Figure 6A. The townspeople wondered whether it was possible to go for a walk and cross each bridge once and once only. This is equivalent to finding an Eulerian cycle for the multigraph in Figure 6B. Euler showed it to be impossible because there are four vertices of odd order.

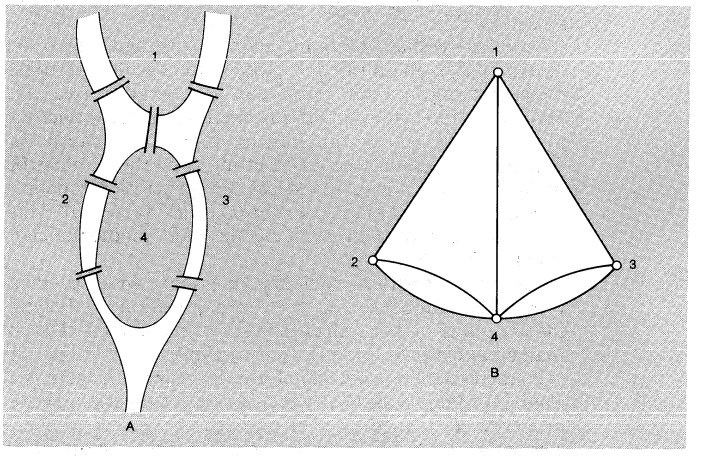



Figure 6: (A) Seven bridges of Königsberg and (B) multigraph.

**Directed graphs.** A directed graph $G$ consists of a non-empty set of elements $V(G)$, called vertices, and a subset $E(G)$ of ordered pairs of distinct elements of $V(G)$. Elements $(x, y)$ of $E(G)$ may be called edges, the direction of the edge being from $x$ to $y$. Both $(x, y)$ and $(y, x)$ may be edges.

A closed path in a directed graph is a sequence of vertices $x_0x_1x_2 \cdots x_n = x_0$, such that $(x_i, x_{i+1})$ is a directed edge for $i = 0, 1, \cdots, n-1$. To each edge $(x, y)$ of a directed graph $G$ there can be assigned a non-negative weight function $f(x, y)$. The problem then is to find a closed path in $G$ traversing all vertices so that the sum of the weights of all edges in the path is a minimum. This is a typical optimization problem. If the vertices are certain cities, the edges are routes joining cities, and the weights are the lengths of the routes, then this becomes the travelling-salesman problem—that is, can he visit each city without retracing his steps? This problem still remains unsolved except for certain special cases. (R.C.Bo./Ed.)

## Combinatorial geometry

The name combinatorial geometry, first used by Hadwiger, is not quite accurately descriptive of the nature of the subject. Combinatorial geometry does touch on those aspects of geometry that deal with arrangements, combinations, and enumerations of geometric objects; but it takes in much more. The field is so new that there has scarcely been time for it to acquire a well-defined position in the mathematical world. Rather it tends to overlap parts of topology (especially algebraic topology), number theory, analysis, and, of course, geometry. The subject concerns itself with relations among members of finite systems of geometric figures subject to various conditions and restrictions. More specifically, it includes problems of

Principal concerns

covering, packing, symmetry, extrema (maxima and minima), continuity, tangency, equalities, and inequalities, many of these with special emphasis on their application to the theory of convex bodies. A few of the fundamental problems of combinatorial geometry originated with Newton and Euler; the majority of the significant advances in the field, however, have been made since 1946.

The unifying aspect of these disparate topics is the quality or style or spirit of the questions and the methods of attacking these questions. Among those branches of mathematics that interest serious working mathematicians, combinatorial geometry is one of the few branches that can be presented on an intuitive basis, without recourse by the investigator to any advanced theoretical considerations or abstractions.

Yet the problems are far from trivial, and many remain unsolved. They can be handled only with the aid of the most careful and often delicate reasoning that displays the variety and vitality of geometric methods in a modern setting. A few of the answers are natural and are intuitively suggested by the questions. Many of the others, however, require proofs of unusual ingenuity and depth even in the two-dimensional case. Sometimes a plane solution may be readily extendible to higher dimensions, but sometimes just the opposite is true, and a three-dimensional or $n$-dimensional problem may be entirely different from its two-dimensional counterpart. Each new problem must be attacked individually. Attempts to create standard methods or theories capable of being applied to the solution of any significant group of the currently unsolved problems in the field had by the late 20th century met with no success. The continuing charm and challenge of the subject are at least in part due to the relative simplicity of the statements coupled with the elusive nature of their solutions.

## SOME HISTORICALLY IMPORTANT TOPICS OF COMBINATORIAL GEOMETRY

**Packing and covering.** It is easily seen that six equal circular disks may be placed around another disk of the same size so that the central one is touched by all the others but no two overlap (Figure 7) and that it is not possible to place seven disks in such a way. In the analogous three-dimensional situation, around a given ball (solid sphere) it is possible to place 12 balls of equal size, all touching the first one but not overlapping it or each other. One such arrangement may be obtained by placing the 12 surrounding balls at the midpoints of edges of a suitable cube that encloses the central ball; each of the 12 balls then touches four other balls in addition to the central one. But if the 12 balls are centred at the 12 vertices of a suitable regular icosahedron surrounding the given ball, there is an appreciable amount of free space between each of the surrounding balls and its neighbours. (If the spheres have radius 1, the distances between the centres of the surrounding spheres are at least $2/\cos 18° = 2.1029 \cdots$.) It appears, therefore, that by judicious positioning it might be possible to have 13 equal non-overlapping spheres touch another of the same size. This dilemma between 12 and 13, one of the first nontrivial problems of combinatorial geometry, was the object of discussion between Isaac Newton and David Gregory in 1694. Newton believed 12 to be the correct number, but this claim was not proved until 1874. The analogous problem in four-dimensional space is still open, the answer being one of the numbers 24, 25, or 26.

Problem of the 13 balls

The problem of the 13 balls is a typical example of the branch of combinatorial geometry that deals with packings and coverings. In packing problems the aim is to place figures of a given shape or size without overlap as economically as possibly, either inside another given figure or subject to some other restriction.

Problems of packing and covering have been the objects of much study, and some striking conclusions have been obtained. For each plane convex set $K$, for example, it is possible to arrange nonoverlapping translates of $K$ so as to cover at least two-thirds of the plane; if $K$ is a triangle (and only in that case), no arrangement of nonoverlapping translates covers more than two-thirds of the plane (Figure 8). On the other hand, many easily stated questions are still open. One of them concerns the densest packing of spheres. If the spheres are packed in cannonball fashion—that is, in the way cannonballs are stacked to form a triangular pyramid, indefinitely extended—then they fill $\pi/\sqrt{18}$, or about 0.74, of the space. Whether this is the greatest density possible is not known, but it was proved in 1958 by the British mathematician C. Ambrose Rogers that, if there exists a closer packing, its density cannot exceed 0.78.

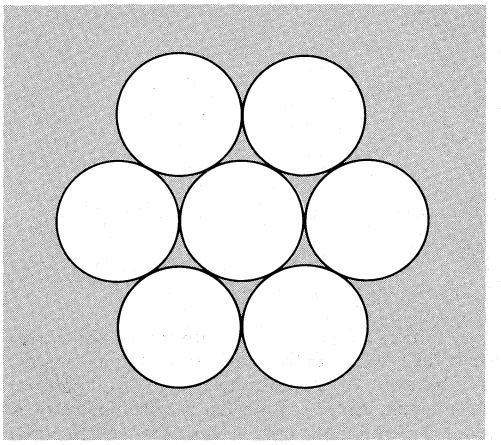


Figure 7: Packing of disks.

Covering problems deal in an analogous manner with economical ways of placing given figures so as to cover (that is, contain in their union) another given figure. One famous covering problem, posed by the French mathematician Henri Lebesgue in 1914, is still unsolved: What is the size and shape of the universal cover of least area? Here a convex set $C$ is called universal cover if for each set $A$ in the plane such that diam $A \leq 1$ it is possible to move $C$ to a suitable position in which it covers $A$. The diameter diam $A$ of a set $A$ is defined as the least upper bound of the mutual distances of points of the set $A$. If $A$ is a compact set, then diam $A$ is simply the greatest distance between any two points of $A$. Thus, if $A$ is an equilateral triangle of side 1, then diam $A = 1$; and if $B$ is a cube of edge length 1, then diam $B = \sqrt{3}$.

The Lebesgue problem

**Polytopes.** A (convex) polytope is the convex hull of some finite set of points (see GEOMETRY: *Euclidean geometry*). Each polytope of dimensions $d$ has as faces finitely many polytopes of dimensions 0 (vertices), 1 (edge), 2 (2-faces), $\cdots$, $d$-1 (facets). Two-dimensional polytopes are usually called polygons, three-dimensional ones polyhedra. Two polytopes are said to be isomorphic, or of the same combinatorial type, provided there exists a one-to-one correspondence between their faces, such that two faces of the first polytope meet if and only if the corresponding faces of the second meet. The prism and the

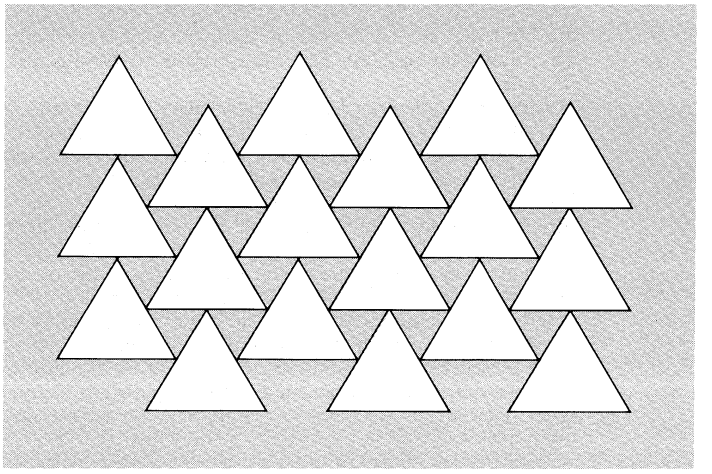


Figure 8: Covering of part of a plane with triangles.

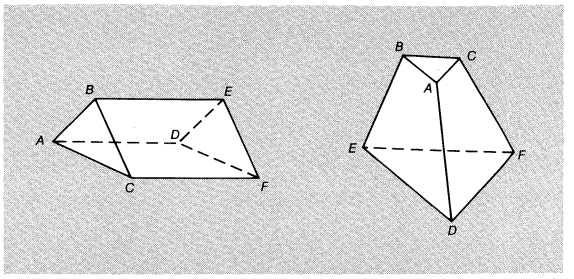



Figure 9: (Left) prism and (right) truncated pyramid.

truncated pyramid of Figure 9 are isomorphic, the correspondence being indicated by the letters at the vertices. To classify the convex polygons by their combinatorial types, it is sufficient to determine the number of vertices $v$; for each $v \geq 3$, all polygons with $v$ vertices ($v$-gons) are of the same combinatorial type, while a $v$-gon and a $v'$-gon are not isomorphic if $v \neq v'$. Euler was the first to investigate in 1752 the analogous question concerning polyhedra. He found that $v - e + f = 2$ for every convex polyhedron, where $v$, $e$, and $f$ are the numbers of vertices, edges, and faces of the polyhedron. Though this formula became one of the starting points of topology (see GEOMETRY: *Topology*), Euler was not successful in his attempts to find a classification scheme for convex polytopes or to determine the number of different types for each $v$. Despite efforts of many famous mathematicians since Euler (J. Steiner, Kirkman, Cayley, O. Hermes, M. Brückner, to mention only a few from the 19th century), the problem is still open. It was established by P.J. Federico in the U.S. that there are 2,606 different combinatorial types of convex polyhedra with nine vertices. The numbers of different types with four, five, six, seven, or eight vertices have been known for some time to be 1, 2, 7, 34, and 257, respectively.

Theory of convex polytopes

The theory of convex polytopes has been more successful in developments in other directions. The regular polytopes have been under investigation since 1880 in dimensions higher than three, together with extensions of Euler's relation to the higher dimensions. (The Swiss geometer Ludwig Schläfli made many of these discoveries some 30 years earlier, but his work was published only posthumously in 1901.) The interest in regular polyhedra and other special polyhedra goes back to ancient Greece, as indicated by the names Platonic solids and Archimedean solids.

Since 1950 there has been considerable interest, in part created by practical problems related to computer techniques such as linear programming, in questions of the following type: for polytopes of a given dimension $d$ and having a given number $v$ of vertices, how large and how small can the number of facets be? Such problems have provided great impetus to the development of the theory. The U.S. mathematician Victor L. Klee solved the maximum problem in 1963 in most cases (that is, for all but a finite number of $v$'s for each $d$), but the remaining cases were disposed of only in 1970 by P. McMullen, in the United States, who used a completely new method. The minimum problem and many related questions are still unsolved.

**Incidence problems.** In 1893 Sylvester posed the question: If a finite set $S$ of points in a plane has the property that each line determined by two points of $S$ meets at least one other point of $S$, must all points of $S$ be on one line? Sylvester never found a satisfactory solution to the problem, and the first (affirmative) solutions were published a half century later. Since then, Sylvester's problem has inspired many investigations and led to many other open questions, both in the plane and in higher dimensions.

**Helly's theorem.** In 1912 E. Helly proved the following theorem, which has since found applications in many areas of geometry and analysis and has led to numerous generalizations, extensions and analogues known as Helly-type theorems. If $K_1, K_2, \cdots, K_n$ are convex sets in $d$-dimensional Euclidean space $E^d$, in which $n \geq d+1$, and if for every choice of $d+1$ of the sets $K_i$ there exists a point that belongs to all the chosen sets, then there exists a point that belongs to all the sets $K_1, K_2, \cdots K_n$. The theorem stated in two dimensions is easier to visualize and yet is not shorn of its strength: If every three of a set of $n$ convex figures in the plane have a common point (not necessarily the same point for all trios), then all $n$ figures have a point in common. If, for example, convex sets $A$, $B$, and $C$ have the point $p$ in common, and convex sets $A$, $B$, and $D$ have the point $q$ in common, and sets $A$, $C$, and $D$ have the point $r$ in common, and sets $B$, $C$, and $D$ have the point $s$ in common, then some point $x$ is a member of $A$, $B$, $C$, and $D$.

Although the connection is often far from obvious, many consequences may be derived from Helly's theorem. Among them are the following, stated for $d = 2$ with some higher dimensional analogues indicated in square brackets:

A. Two finite subsets $X$ and $Y$ of the plane [$d$-space] may be strictly separated by a suitable straight line [hyperplane] if and only if, for every set $Z$ consisting of at most 4 [$d+2$] points taken from $X \cup Y$, the points of $X \cap Z$ may be strictly separated from those of $Y \cap Z$. (A line [hyperplane] $L$ strictly separates $X$ and $Y$ if $X$ is contained in one of the open half planes [half spaces] determined by $L$ and if $Y$ is contained in the other.)

B. Each compact convex set $K$ in the plane [$d$-space] contains a point $P$ with the following property: each chord of $K$ that contains $P$ is divided by $P$ into a number of segments so the ratio of their lengths is at most $2d$.

C. If $G$ is an open subset of the plane [$d$-space] with finite area [$d$-dimensional content], then there exists a point $P$, such that each open half plane [half space] that contains $P$ contains also at least $1/3$ [$1/(d+1)$] of the area [$d$-content] of $G$.

D. If $I_1, \cdots, I_n$ are segments parallel to the $y$-axis in a plane with a coordinate system $(x, y)$, and if for every choice of three of the segments there exists a straight line intersecting each of the three segments, then there exists a straight line that intersects all the segments $I_1, \cdots, I_n$.

Theorem D has generalizations in which $k$th degree polynomial curves $y = a_k x^k + \cdots + a_1 x + a_0$ take the place of the straight lines and $k+2$ replaces 3 in the assumptions. These are important in the theory of best approximation of functions by polynomials.

## METHODS OF COMBINATORIAL GEOMETRY

Many other branches of combinatorial geometry are as important and interesting as those mentioned above, but rather than list them here it is more instructive to provide a few typical examples of frequently used methods of reasoning. Because the emphasis is on illustrating the methods rather than on obtaining the most general results, the examples will deal with problems in two and three dimensions.

**Exhausting the possibilities.** Using the data available concerning the problem under investigation, it is often possible to obtain a list of all potential, a priori possible, solutions. The final step then consists in eliminating the possibilities that are not actual solutions or that duplicate previously found solutions. An example is the proof that there are only five regular convex polyhedra (the Platonic solids) and the determination of what these five are.

The five Platonic solids

From the definition of regularity it is easy to deduce that all the faces of a Platonic solid must be congruent regular $k$-gons for a suitable $k$, and that all the vertices must belong to the same number $j$ of $k$-gons. Because the sum of the face angles at a vertex of a convex polyhedron is less than $2\pi$, and because each angle of the $k$-gon is $(k-2)\pi/k$, it follows that $j(k-2)\pi/k < 2\pi$, or $(j-2)(k-2) < 4$. Therefore, the only possibilities for the pair $(j, k)$ are (3, 3), (3, 4), (3, 5), (4, 3), and (5, 3). It may be verified that each of these pairs actually corresponds to a Platonic solid, namely, to the tetrahedron, the cube, the dodecahedron, the octahedron, and the icosahedron, respectively. Very similar arguments may be used in the determination of Archimedean solids and in other instances.

The most serious drawback of the method is that in many instances the number of potential (and perhaps actual) solutions is so large as to render the method unfeasible. For example, it is known that the number of different combinatorial types of convex polyhedra with 10 vertices

exceeds 30,000, the number with 11 vertices exceeds 400,000, and the number with 12 vertices exceeds 5,000,000. Therefore, the exact determination of these numbers by the method just discussed is out of the question, certainly if attempted by hand and probably even with the aid of a computer.

**Use of extremal properties.** In many cases the existence of a figure or an arrangement with certain desired properties may be established by considering a more general problem (or a completely different problem) and by showing that a solution of the general problem that is extremal in some sense provides also a solution to the original problem. Frequently there seems to be very little connection between the initial question and the extremal problem. As an illustration the following theorem will be proved: If $K$ is a two-dimensional compact convex set with a centre of symmetry, there exists a parallelogram $P$ containing $K$, such that the midpoints of the sides of $P$ belong to $K$. The proof proceeds as follows: Of all the parallelograms that contain $K$, the one with least possible area is labeled $P_0$. The existence of such a $P_0$ is a consequence of the compactness of $K$ and may be established by standard arguments. It is also easily seen that the centres of $K$ and $P_0$ coincide. The interesting aspect of the situation is that $P_0$ may be taken as the $P$ required for the theorem. In fact (Figure 10), if the midpoints $A'$ and $A''$ of a pair of sides of $P_0$ do not belong to $K$, it is possible to

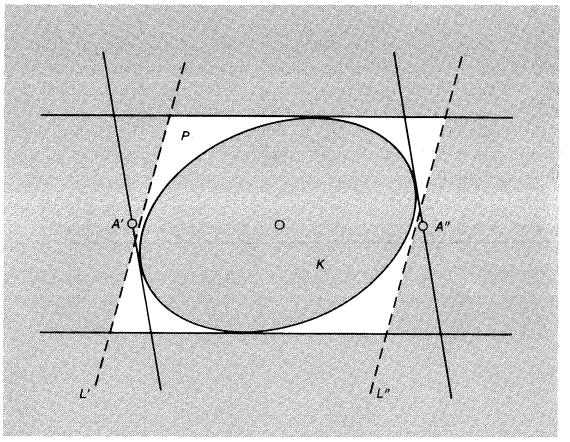



Figure 10: Example of theorem on extremal properties (see text).

strictly separate them from $K$ by parallel lines $L'$ and $L''$ that, together with the other pair of sides of $P_0$, determine a new parallelogram containing $K$ but with area smaller than that of $P_0$. The above theorem and its proof generalize immediately to higher dimensions and lead to results that are important in functional analysis (see ANALYSIS [IN MATHEMATICS]: *Functional analysis*).

Disproof of existence of extremal figures

Sometimes this type of argument is used in reverse to establish the existence of certain objects by disproving the possibility of existence of some extremal figures. As an example the following solution of the problem of Sylvester discussed above can be mentioned. By a standard argument of projective geometry (duality), it is evident that Sylvester's problem is equivalent to the question: If through the point of intersection of any two of $n$ coplanar lines, no two of which are parallel, there passes a third, are the $n$ lines necessarily concurrent? To show that they must be concurrent, contradiction can be derived from the assumption that they are not concurrent. If $L$ is one of the lines, then not all the intersection points lie on $L$. Among the intersection points not on $L$, there must be one nearest to $L$, which can be called $A$. Through $A$ pass at least three lines, which meet $L$ in points $B$, $C$, $D$, so that $C$ is between $B$ and $D$. Through $C$ passes a line $L^*$ different from $L$ and from the line through $A$. Since $L^*$ enters the triangle $ABD$, it intersects either the segment $AB$ or the segment $AD$, yielding an intersection point nearer to $L$ than the supposedly nearest intersection point $A$, thus providing the contradiction.

The difficulties in applying this method are caused in part by the absence of any systematic procedure for devising an extremal problem that leads to the solution of the original question.

**Use of figures with special properties.** Sometimes a general theorem may be established by the use of appropriate special figures, even if they are not of the kind that the theorem is concerned with. This method is used in considering the question known as Borsuk's problem.

The Polish mathematician K. Borsuk proved in 1933 that in any decomposition of the $d$-dimensional ball $B^d$ into $d$ subsets, at least one of the subsets has a diameter equal to diam $B^d$; and he asked whether it is possible to decompose every subset $A$ of the $d$-dimensional space into $d+1$ subsets, each of which has a diameter smaller than diam $A$. (Such a decomposition is easily found if $A$ is the ball $B^d$.) In case $d=2$ Borsuk's problem reduces to the question of whether each plane set $A$ may be decomposed into three parts, each of diameter less than diam $A$. An affirmative answer follows in this case from the fact (which is not hard to prove) that each planar set $A$ with diam $A=1$ may be covered by a regular hexagon $H$ of edge length $1\sqrt{3}=0.577\cdots$ (the diameter of $H$ is diam $H=2\sqrt{3}=1.155\cdots>1$, and the distance between the pairs of parallel sides is 1; see Figure 11). Such a hexagon $H$ may be cut into three pentagons (indicated in Figure 11 by dotted lines), each of which has a diameter of only $\sqrt{3}/2=0.866\cdots<1$. This partition of $H$ may clearly be used to partition each planar set of diameter 1, thus establishing the following stronger variant of Borsuk's problem in the plane: each planar set $A$ may be decomposed into three subsets, each of diameter at most $0.866\cdots\times$(diam $A$). An affirmative solution of Borsuk's problem in the three-dimensional case may be proved by a similar method, in which the hexagon $H$ is replaced by a polyhedron obtained by appropriate triple truncation of the regular octahedron.

Borsuk's problem on decomposition

**Use of transformations between different spaces and applications of Helly's theorem.** Although those two methods do not necessarily go together, both may be illustrated in one example—the proof of theorem D concerning parallel segments. Let the segment $I_i$ have end-points $(x_i, y_i)$ and $(x_i, y'_i)$, where $y_i \leq y'_i$ and $i=1, 2, \cdots, n$. The case that two of the segments are on one line is easily disposed of; so it may be assumed that $x_1, x_2, \cdots, x_n$ are all different. With each straight line $y=ax+b$ in the $(x, y)$-plane can be associated a point $(a, b)$ in another plane, the $(a, b)$-plane. Now, for $i=1, 2, \cdots, n$, the set consisting of all those points $(a, b)$ for which the corresponding line $y=ax+b$ in the $(x, y)$ plane meets the

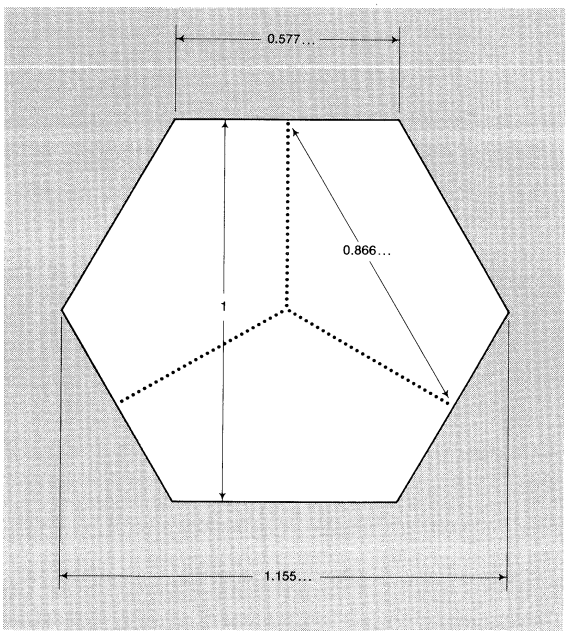



Figure 11: Illustration of Borsuk's problem.

segment $I_i$ can be denoted by $K_i$. This condition means that $y_i \leq ax_i + b \leq y'_i$ so that each set $K_i$ is convex. The existence of a line intersecting three of the segments $I_i$ means that the corresponding sets $K_i$ have a common point. Then Helly's theorem for the $(a, b)$-plane implies the existence of a point $(a^*, b^*)$ common to all sets $K_i$. This in turn means that the line $y = a^*x + b^*$ meets all the segments $I_i$, $I_2$, $\cdots$, $I_n$, and the proof of theorem D is complete.

In addition to the methods illustrated above, many other techniques of proof are used in combinatorial geometry, ranging from simple mathematical induction to sophisticated decidability theorems of formal logic. The variety of methods available and the likelihood that there are many more not yet invented continue to stimulate research in this rapidly developing branch of mathematics. (B.G.)

**BIBLIOGRAPHY.** JAMES LEGGE (trans.), *The Yî-King,* vol. 16 of the *Sacred Books of the East* (1882, reprinted 1962); *Algebra, with Arithmetic and Mensuration from the Sanscrit of Brahmagupta and Bhâskara,* trans. by H.T. COLEBROOKE (1817); and NASIR AD-DIN AL-TUSI, "Handbook of Arithmetic Using Board and Dust," *Math. Rev.,* 31:5776 (1966), complete Russian trans. by S.A. AHMEDOV and B.A. ROZENFELD in *Istor.-Mat. Issled.,* 15:431–444 (1963), give glimpses of some of the early beginnings of the subject in the Orient. The term combinatorial was first used in GOTTFRIED WILHELM LEIBNIZ, *Dissertatio de arte combinatoria* (1666). W.W. ROUSE BALL, *Mathematical Recreations and Essays,* rev. by H.S. MACDONALD COXETER (1942), contains an account of some of the famous recreational combinatorial problems of the 19th century, such as the problem of eight queens, Hamiltonian circuits, and the Kirkman schoolgirl problem. EUGEN NETTO, *Lehrbuch der Combinatorik,* 2nd ed. (1927, reprinted 1958); and PERCY A. MACMAHON, *Combinatory Analysis,* 2 vol. (1915–16, reprinted 1960), show the state of the subject in the early part of the 20th century. HERBERT J. RYSER, *Combinatorial Mathematics* (1963); MARSHALL HALL, JR., *Combinatorial Theory* (1967); C.L. LIU, *Introduction to Combinatorial Mathematics* (1968), all deal with combinatorics in general. JOHN RIORDAN, *An Introduction to Combinatorial Analysis* (1958); and CLAUDE BERGE, *Principes de combinatoire* (1968; Eng. trans., *Principles of Combinatorics,* 1971), deal with problems of enumeration. CLAUDE BERGE, *Théorie des graphes et ses applications* (1957; Eng. trans., *The Theory of Graphs and Its Applications,* 1962); CLAUDE BERGE and A. GHOUILAHOURI, *Programmes, jeux et réseaux de transport* (1962; Eng. trans., *Programming, Games and Transportation Networks,* 1965); FRANK HARARY, *Graph Theory* (1969), deal with graph theoretic problems. OYSTEIN ORE, *The Four-Color Problem* (1967), gives an introduction to this problem. PETER DEMBOWSKI, *Finite Geometries* (1968), contains most of the important developments on designs, including partially balanced and group divisible designs. ELWYN R. BERLEKAMP, *Algebraic Coding Theory* (1968), may be consulted for combinatorial aspects of coding theory. MARSHALL HALL, JR., "A Survey of Combinatorial Analysis," in *Surveys in Applied Mathematics,* vol. 4 (1958), gives a very good survey of combinatorial developments up to 1958. E.F. BECKENBACH (ed.), *Applied Combinatorial Mathematics* (1964), gives a good idea of the wide range of applications of modern combinatorics. G.C. ROTA, "Combinatorial Analysis," in G.A.W. BOEHM (ed.), *The Mathematical Sciences: A Collection of Essays* (1969), in addition to surveying some of the famous combinatorial problems brings out modern trends and indicates where combinatorics is headed. HUGO HADWIGER and HANS DEBRUNNER, *Combinatorial Geometry in the Plane* (1964); I.M. YAGLOM and V.G. BOLTYANSKY, *Convex Figures* (1961; orig. pub. in Russia, 1951); V.G. BOLTYANSKY, *Equivalent and Equidecomposable Figures* (1963; orig. pub. in Russian, 1956); and L.A. LYUSTERNIK, *Convex Figures and Polyhedra* (1966; orig. pub. in Russian, 1956), deal with aspects of combinatorial geometry on an elementary level. On an advanced level, see H.S. MACDONALD COXETER, *Regular Polytopes,* 2nd ed. (1963); L. FEJES TOTH, *Lagerungen in der Ebene, auf der Kugel und im Raum* (1953) and *Regular Figures* (1964); BRANKO GRUNBAUM, *Convex Polytopes* (1967) and *Arrangements and Spreads* (1972); and C.A. ROGERS, *Packing and Covering* (1964). See also KENNETH P. BOGART, *Introductory Combinatorics* (1983); and R.J. WILSON (ed.), *Applications of Combinatorics* (1982), both accessible to laymen.

(R.C.Bo./B.G.)

# Communication

Communication, the exchange of meanings between individuals through a common system of symbols, concerned scholars since the time of ancient Greece. Until modern times, however, the topic was usually subsumed under other disciplines and taken for granted as a natural process inherent to each. In 1928 the English literary critic and author I.A. Richards offered one of the first—and in some ways still the best—definitions of communication as a discrete aspect of human enterprise:

> Communication takes place when one mind so acts upon its environment that another mind is influenced, and in that other mind an experience occurs which is like the experience in the first mind, and is caused in part by that experience.

Richards' definition is both general and rough, but its application to nearly all kinds of communication—including those between humans and animals (but excluding machines)—separated the contents of messages from the processes in human affairs by which these messages are transmitted. More recently, questions have been raised concerning the adequacy of any single definition of the term communication as it is currently employed. The American psychiatrist and scholar Jurgen Ruesch has identified 40 varieties of disciplinary approaches to the subject, including architectural, anthropological, psychological, political, and many other interpretations of the apparently simple interaction described by Richards. In total, if such informal communications as sexual attraction and play behaviour are included, there exist at least 50 modes of interpersonal communication that draw upon dozens of discrete intellectual disciplines and analytic approaches. Communication may therefore be analyzed in at least 50 different ways.

Interest in communication has been stimulated by advances in science and technology, which, by their nature, have called attention to man as a communicating creature. Among the first and most dramatic examples of the inventions resulting from technological ingenuity were the telegraph and telephone, followed by others like wireless radio and telephoto devices. The development of popular newspapers and periodicals, broadcasting, motion pictures, and television led to institutional and cultural innovations that permitted efficient and rapid communication between a few individuals and large populations; these media have been responsible for the rise and social power of the new phenomenon of mass communication. (See also INFORMATION PROCESSING AND INFORMATION SYSTEMS; TELECOMMUNICATIONS SYSTEMS.)

This article is divided into the following sections:

Models of communication 623
  Linear models 623
  Dynamic models 624
  Applications of formal logic and mathematics 624
Types of communication 625
  Nonvocal communication 625
  Vocal communication 626
  Mass and public communication 627
The psychology of communication 627
Bibliography 628

Communication and other disciplines

Since about 1920 the growth and apparent influence of communications technology have attracted the attention of many specialists who have attempted to isolate communication as a specific facet of their particular interest. Psychologists, in their studies of behaviour and mind, have evolved concepts of communication useful to their investigations as well as to certain forms of therapy. Social scientists have identified various forms of communication by which myths, styles of living, mores, and traditions are passed either from generation to generation or from one segment of society to another. Political scientists and economists have recognized that communication of many types lies at the heart of the regularities in the social order. Under the impetus of new technology—particularly high-speed computers—mathematicians and engineers have tried to quantify and measure components of communicated information and to develop methods for translating various types of messages into quantities or amounts amenable to both their procedures and instruments. Numerous and differently phrased questions have been posed by artists, architects, artisans, writers, and others concerning the overall influences of various types of communication. Many researchers, working within the relevant concerns of their disciplines, have also sought possible theories or laws of cause and effect to explain the ways in which human dispositions are affected by certain kinds of communication under certain circumstances, and the reasons for the change.

In the 1960s a Canadian educator, Marshall McLuhan, drew the threads of interest in the field of communication into a view that associated many contemporary psychological and sociological phenomena with the media employed in modern culture. McLuhan's often repeated idea, "the medium is the message," stimulated numerous filmmakers, photographers, artists, and others, who adopted McLuhan's view that contemporary society had moved (or was moving) from a "print" culture to a "visual" one. The particular forms of greatest interest to McLuhan and his followers were those associated with the sophisticated technological instruments for which young people in particular display enthusiasm, namely motion pictures, television, and sound recordings.

Modern interests in communication

By the late 20th century the main focus of interest in communication seemed to be drifting away from McLuhanism and to be centring upon: (1) the mass communication industries, the people who run them, and the effects they have upon their audiences; (2) persuasive communication and the use of technology to influence dispositions; (3) processes of interpersonal communication as mediators of information; (4) dynamics of verbal and nonverbal (and perhaps extrasensory) communication between individuals; (5) perception of different kinds of communications; (6) uses of communication technology for social and artistic purposes, including education in and out of school; and (7) development of relevant criticism for artistic endeavours employing modern communications technology.

In short, a communication expert may be oriented to any of a number of disciplines in a field of inquiry that has, as yet, neither drawn for itself a conclusive roster of subject matter nor agreed upon specific methodologies of analysis.

## MODELS OF COMMUNICATION

Fragmentation and problems of interdisciplinary outlook have generated a wide range of discussion concerning the ways in which communication occurs and the processes it entails. Most speculation on these matters admits, in one way or another, that the communication theorist's task is to answer as clearly as possible the question, "*Who* says *what* to *whom* with *what effect?*" (This query was originally posed by the U.S. political scientist Harold D. Lasswell.) Obviously, all of the critical elements in this question may be interpreted differently by scholars and writers in different disciplines.

**Linear models.** One of the most productive schematic models of a communications system that has been proposed as an answer to Lasswell's question emerged in the late 1940s, largely from the speculations of two U.S. mathematicians, Claude Shannon and Warren Weaver. The simplicity of their model, its clarity, and its surface generality proved attractive to many students of commu-

Shannon's and Weaver's schematic model

nication in a number of disciplines, although it is neither the only model of the communication process extant nor is it universally accepted. As originally conceived, the model contained five elements—an information source, a transmitter, a channel of transmission, a receiver, and a destination—all arranged in linear order. Messages (electronic messages, initially) were supposed to travel along this path, to be changed into electric energy by the transmitter, and to be reconstituted into intelligible language by the receiver. In time, the five elements of the model were renamed so as to specify components for other types of communication transmitted in various manners. The information source was split into its components (both source and message) to provide a wider range of applicability. The six constituents of the revised model are: (1) a source, (2) an encoder, (3) a message, (4) a channel, (5) a decoder, and (6) a receiver. For some communication systems, the components are as simple to specify as, for instance, (1) a man on the telephone, (2) the mouthpiece of the telephone, (3) the words the man speaks, (4) the electrical wires along which the words (now electrical impulses) travel, (5) the earpiece of another telephone, and (6) the mind of the listener. In other communication systems, the components are more difficult to isolate; *e.g.*, the communication of the emotions of a fine artist by means of a painting to people who may respond to the message long after the artist's death.

Begging a multitude of psychological, aesthetic, and sociological questions concerning the exact nature of each component, the linear model appeared, from the commonsense perspective, at least, to explain in general terms the ways in which certain classes of communication occurred. It did not indicate the reason for the inability of certain communications—obvious in daily life—to fit its neat paradigm.

*Entropy, negative entropy, and redundancy.* Another concept, first called by Shannon a "noise source" but later associated with the notion of entropy (a principle derived from physics), was imposed upon the communication model. Entropy is analogous in most communication to audio or visual static—that is, to outside influences that diminish the integrity of the communication and, possibly, distort the message for the receiver. Negative entropy may also occur in instances in which incomplete or blurred messages are nevertheless received intact, either because of the ability of the receiver to fill in missing details or to recognize, despite distortion or a paucity of information, both the intent and content of the communication.

Although rarely shown on diagrammatic models of this version of the communication process, redundancy—the repetition of elements within a message that prevents the failure of communication of information—is the greatest antidote to entropy. Most written and spoken languages, for example, are roughly half-redundant. If 50 percent of the words of this article were taken away at random, there would still remain an intelligible—although somewhat peculiar—essay. Similarly, if one-half of the words of a radio news commentator are heard, the broadcast can usually be understood. Redundancy is apparently involved in most human activities, and, because it helps to overcome the various forms of entropy that tend to turn intelligible messages into unintelligible ones (including psychological entropy on the part of the receiver), it is an indispensable element for effective communication.

Messages are therefore susceptible to considerable modification and mediation. Entropy distorts, while negative entropy and redundancy clarify; as each occurs differentially in the communication process, the chances of the message being received and correctly understood vary. Still, the process (and the model of it) remains conceptually static, because it is fundamentally concerned with messages sent from point to point, and not with their results or possible influences upon sender and receiver.

*Feedback.* To correct this flaw, the principle of feedback was added to the model and provided a closer approximation of interpersonal human interaction than was known theretofore. This construct was derived from the studies of Norbert Wiener, the so-called father of the science of cybernetics. Wiener's cybernetic models, some of which provide the basis for current computer technology, were designed to be responsive to their own behaviour; that is, they audited their own performances mathematically or electronically in order to avoid errors of entropy, unnecessary redundancy, or other simple hazards.

Aspects of feedback

Certain types of common communications—Christmas cards, for instance—usually require little feedback. Others, particularly interactions between human beings in conversation, cannot function without the ability of the message sender to weigh and calculate the apparent effect of his words on his listener. It is largely the aspect of feedback that provides for this model the qualities of a process, because each instance of feedback conditions or alters the subsequent messages.

**Dynamic models.** Other models of communication processes have been constructed to meet the needs of students of communication whose interests differ from those of quantitatively oriented theorists like Shannon, Weaver, and Wiener. While the model described above displays some generality and shows simplicity, it lacks some of the predictive, descriptive, and analytic powers found in other approaches. A psychologist, Theodore M. Newcomb, for example, has articulated a more fluid system of dimensions to represent the individual interacting in his environment. Newcomb's model and others similar to it are not as precisely mathematical (quantitative) as Shannon's and thus permit more flexible accounts of human behaviour and its variable relationships. They do not deny the relevance of linear models to Shannon and Weaver's main concerns—quanta of information and the delivery of messages under controlled conditions—but they question their completeness and utility in describing cognitive, emotional, and artistic aspects of communication as they occur in sociocultural matrices.

Students concerned mainly with persuasive and artistic communication often centre attention upon different kinds, or modes, of communication (*i.e.*, narrative, pictorial, and dramatic) and theorize that the messages they contain, including messages of emotional quality and artistic content, are communicated in various manners to and from different sorts of people. For them, the stability and function of channel or medium are more variable and less mechanistically related to the process than they are for followers of Shannon and Weaver and psychologists like Newcomb. (McLuhan, indeed, asserts that the channel actually dictates, or severely influences, the message—both as sent and received.) Many analysts of communication, linguistic philosophers, and others are concerned with the nature of messages, particularly their compatibility with sense and emotion, their style, and the intentions behind them. They find both linear and geometric models of process of little interest to their concerns, although considerations related to these models, particularly those of entropy, redundancy, and feedback, have provided significant and productive concepts for most students of communication.

**Applications of formal logic and mathematics.** Despite the numerous types of communication or information theory extant today—and those likely to be formulated tomorrow—the most rationally and experimentally consistent approaches to communication theory so far developed follow the constructions of Shannon and others described above. Such approaches tend to employ the structural rigours of logic rather than the looser syntaxes, grammars, and vocabularies of common languages, with their symbolic, poetic, and inferential aspects of meaning.

Cybernetic theory and computer technology require rigorous but straightforward languages to permit translation into nonambiguous, special symbols that can be stored and utilized for statistical manipulations. The closed system of formal logic proved ideal for this need. Premises and conclusions drawn from syllogisms according to logical rules may be easily tested in a consistent, scientific manner, as long as all parties communicating share the rational premises employed by the particular system.

Formal logic, syllogisms, and communication

That this logical mode of communication drew its frame of discourse from the logic of the ancient Greeks was inevitable. Translated into an Aristotelian manner of discourse, meaningful interactions between individuals could be transferred to an equally rational closed system of

mathematics: an arithmetic for simple transactions, an algebra for solving certain well-delimited puzzles, a calculus to simulate changes, rates and flows, and a geometry for purposes of illustration and model construction. This progression has proved quite useful for handling those limited classes of communications that arise out of certain structured, rational operations, like those in economics, inductively oriented sociology, experimental psychology, and other behavioral and social sciences, as well as in most of the natural sciences.

The basic theorem of information theory rests, first, upon the assumption that the message transmitted is well organized, consistent, and characterized by relatively low and determinable degrees of entropy and redundancy. (Otherwise, the mathematical structure might yield only probability statements approaching random scatters, of little use to anyone.) Under these circumstances, by devising proper coding procedures for the transmitter, it becomes possible to transmit symbols over a channel at an average rate that is nearly the capacity of units per second of the channel (symbolized by C) as a function of the units per second from an information source (H)—but never at rates in excess of capacity divided by units per second (C/H), no matter how expertly the symbols are coded. As simple as this notion seems, upon determining the capacity of the channel and by cleverly coding the information involved, precise mathematical models of information transactions (similar to electronic frequencies of energy transmissions) may be evolved and employed for complex analyses within the strictures of formal logic. They must, of course, take into account as precisely as possible levels of entropy and redundancy as well as other known variables.

The internal capacities of the channel studied and the sophistication of the coding procedures that handle the information limit the usefulness of the theorem presented above. At present such procedures, while they may theoretically offer broad prospects, are restricted by formal encoding procedures that depend upon the capacities of the instruments in which they are stored (nowadays, mostly on magnetic tape and disk-packs in computers). Although such devices can handle quickly the logic of vast amounts of relatively simple information, they cannot match the flexibility and complexity of the human brain, still man's prime instrument for managing the subtleties of most communication.

## TYPES OF COMMUNICATION

Signals, signs, and symbols

**Nonvocal communication.** Signals, signs, and symbols, three related components of communication processes found in all known cultures, have attracted considerable scholarly attention because they do not relate primarily to the usual conception of words or language. Each is apparently an increasingly more complex modification of the former, and each was probably developed in the depths of prehistory before, or at the start of, man's early experiments with vocal language.

*Signals.* A signal may be considered as an interruption in a field of constant energy transfer. An example is the dots and dashes that open and close the electromagnetic field of a telegraph circuit. Such interruptions do not require the construction of a man-made field; interruptions in nature (*e.g.*, the tapping of a pencil in a silent room, or puffs of smoke rising from a mountain top) may produce the same result. The basic function of such signals is to provide the change of a single environmental factor in order to attract attention and to transfer meaning. A code system that refers interruptions to some form of meaningful language may easily be developed with a crude vocabulary of dots, dashes, or other elemental audio and visual articulations. Taken by themselves, the interruptions have a potential breadth of meaning that seems extremely small; they may indicate the presence of an individual in a room, his impatience, agreement, or disagreement with some aspect of his environment or, in the case of a scream for help, a critical situation demanding attention. Coded to refer to spoken or written language, their potential to communicate language is extremely great.

*Signs.* While signs are usually less germane to the development of words than signals, most of them contain greater amounts of meaning of and by themselves. Ashley Montagu, an anthropologist, has defined a sign as a "concrete denoter" possessing an inherent specific meaning, roughly analogous to the sentence "This is it; do something about it!" The most common signs encountered in daily life are pictures or drawings, although a human posture like a clenched fist, an outstretched arm, or a hand posed in a "Stop" gesture may also serve as signs. The main difference between a sign and a signal is that a sign (like a policeman's badge) contains meanings of an intrinsic nature; a signal (like a scream for help) is merely a device by which one is able to formulate extrinsic meanings. Their difference is illustrated by the observation that many types of animals respond to signals, while only a few intelligent and trained animals (usually dogs and apes) are competent to respond even to simple signs.

All known cultures utilize signs to convey relatively simple messages swiftly and conveniently. Signs may depend for their meaning upon their form, setting, colour, or location. In the United States, traffic signs, uniforms, badges, and barber poles are frequently encountered signs. Taken en masse, any society's lexicon of signs makes up a rich vocabulary of colourful communications.

*Symbols.* Symbols are more difficult than signs to understand and to define because, unlike signs and signals, they are intricately woven into an individual's ongoing perceptions of the world. They appear to contain a dimly understood capacity that (as one of their functions), in fact, defines the very reality of that world. The symbol has been defined as any device with which an abstraction can be made. Although far from being a precise construction, it leads in a profitable direction. The abstractions of the values that people imbue in other people and in things they own and use lie at the heart of symbolism. Here is a process, according to the British philosopher Alfred North Whitehead, whereby

> some components of [the mind's] experience elicit consciousness, beliefs, emotions, and usages respecting other components of experience.

In Whitehead's opinion, symbols are analogues or metaphors (that may include written and spoken language as well as visual objects) standing for some quality of reality that is enhanced in importance or value by the process of symbolization itself.

Symbol systems in all societies

Almost every society has evolved a symbol system whereby, at first glance, strange objects and odd types of behaviour appear to the outside observer to have irrational meanings and seem to evoke odd, unwarranted cognitions and emotions. Upon examination each symbol system reflects a specific cultural logic, and every symbol functions to communicate information between members of the culture in much the same way as, but in a more subtle manner than, conventional language. Although a symbol may take the form of as discrete an object as a wedding ring or a totem pole, symbols tend to appear in clusters and depend upon one another for their accretion of meaning and value. They are not a language of and by themselves; rather they are devices by which ideas too difficult, dangerous, or inconvenient to articulate in common language are transmitted between people who have acculturated in common ways. It does not appear possible to compile discrete vocabularies of symbols, because they lack the precision and regularities present in natural language that are necessary for explicit definitions.

*Icons.* Rich clusters of related and unrelated symbols are usually regarded as icons. They are actually groups of interactive symbols, like the White House in Washington, D.C., a funeral ceremony, or an Impressionist painting. Although in examples such as these, there is a tendency to isolate icons and individual symbols for examination, symbolic communication is so closely allied to all forms of human activity that it is generally and nonconsciously used and treated by most people as the most important aspect of communication in society. With the recognition that spoken and written words and numbers themselves constitute symbolic metaphors, their critical roles in the worlds of science, mathematics, literature, and art can be understood. In addition, with these symbols, an individual is able to define his own identity.

*Gestures.* Professional actors and dancers have known since antiquity that body gestures may also generate a vocabulary of communication more or less unique to each culture. Some U.S. scholars have tried to develop a vocabulary of body language, called kinesics. The results of their investigations, both amusing and potentially practical, may eventually produce a genuine lexicon of American gestures similar to one prepared in detail by François Delsarte, a 19th-century French teacher of pantomime and gymnastics who described the ingenious and complex language of contemporary face and body positions for theatrical purposes.

*Proxemics.* Of more general, cross-cultural significance are the theories involved in the study of "proxemics" developed by a U.S. anthropologist, Edward Hall. Proxemics involves the ways in which people in various cultures utilize both time and space as well as body positions and other factors for purposes of communication. Hall's "silent language" of nonverbal communications consists of such culturally determined interactions as the physical distance or closeness maintained between individuals, the body heat they give off, odours they perceive in social situations, angles of vision they maintain while talking, the pace of their behaviour, and the sense of time appropriate for communicating under differing conditions. By comparing matters like these in the behaviour of different social classes (and in varying relationships), Hall elaborated and codified a number of sophisticated general principles that demonstrate how certain kinds of nonverbal communication occur. Although Hall's most impressive arguments are almost entirely empirical, and many of them are open to question, the study of proxemics does succeed in calling attention to major features of communication dynamics rarely considered by linguists and symbologists. Students of words have been more interested in objective formal vocabularies than in the more subtle means of discourse unknowingly acquired by the members of a culture.

**Vocal communication.** Significant differences between nonvocal and vocal communication are matters more of degree than of kind. Signs, signals, symbols, and possibly icons may, at times, be easily verbalized, although most people tend to think of them as visual means of expression. Kinesics and proxemics may also, in certain instances, involve vocalizations as accompaniments to nonverbal phenomena or as somehow integral to them. Be they grunts, words, or sentences, their function is to help in forwarding a communication that is fundamentally nonverbal.

Origins of speech

Although there is no shortage of speculation on the issue, the origins of human speech remain obscure at present. It is plausible that man is born with an instinct for speech. A phenomenon supporting this belief is the presence of unlearned cries and gurgles of infants operating as crude, vocal signs directed to others the baby cannot possibly be aware of. Some anthropologists claim that within the vocabularies of kinesics and proxemics are the virtual building blocks of spoken language; they postulate that primitive men made various and ingenious inventions (including speech) as a result of their need to communicate with others in order to pool their intellectual and physical resources. Other observers suggest similar origins of speech, including the vocalization of physical activity, imitation of the sounds of nature, and sheer serendipity. Scientific proof of any of these speculations is at present impossible.

Not only is the origin of speech disputed among experts but the precise reasons for the existence of the numerous languages of the world are also far from clear. In the 1920s, an American linguistic anthropologist, Edward Sapir, and, later Benjamin Lee Whorf, centred attention upon the various methods of expression found in different cultures. Drawing their evidence primarily from the languages of primitive societies, they made some very significant observations concerning spoken (and probably written) language. First, man's language reflects in subtle ways those matters of greatest relevance and importance to the value system of each particular culture. Thus, language may be said to reflect culture, or, in other words, people seem to find ways of saying what they need to say. A familiar illustration is the many words (or variations of words) that Eskimos use to describe whale blubber in its various states; *e.g.*, on the whale, ready to eat, raw, cooked, rancid. Another example is the observation that "drunk" possesses more synonyms than any other term in the English language. Apparently, this is the result of a psychological necessity to euphemize a somewhat nasty, uncomfortable, or taboo matter, a device also employed for other words that describe seemingly important, but improper, behaviour or facets of culture.

Need for euphemisms

*Adaptability of language.* Other observations involve the discovery that any known language may be employed, without major modification, to say almost anything that may be said in any other language. A high degree of circumlocution and some nonverbal vocalization may be required to accomplish this end, but, no matter how alien the concept to the original language, it may be expressed clearly in the language of another culture. Students of linguistic anthropology have been able to describe adequately in English esoteric linguistic propositions of primitive societies, just as it has been possible for anthropologists to describe details of Western technology to natives in remote cultures. Understood as an artifact of culture, spoken language may therefore be considered as a universal channel of communication into which various societies dip differentially in order to expedite and specify the numerous points of contact between individuals.

Language remains, however, a still partially understood phenomenon used to transact several types of discourse. Language has been classified on the basis of several criteria. One scheme established four categories on the basis of informative, dynamic, emotive, and aesthetic functions. Informative communication deals largely with narrative aspects of meaning; dynamic discourse concerns the transaction of dispositions such as opinions and attitudes; the emotive employment of language involves the evocation of feeling states in others in order to impel them to action; and aesthetic discourse, usually regarded as a poetic quality in speech, conveys stylistic aspects of expression.

*Laughter.* Although most vocal sounds other than words are usually considered prelinguistic language, the phenomenon of laughter as a form of communication is in a category by itself, with its closest relative being its apparent opposite, crying. Contemporary ethologists, like Konrad Lorenz, have attempted to associate laughter with group behaviour among animals in instances in which aggression is thwarted and laughlike phenomena seem to result among herds. Lorenz's metaphors, while apparently reasonable, cannot be verified inductively. They seem less reasonable to many than the more common notions of Freud and others that laughter either results from, or is related to, the nonconscious reduction of tensions or inhibitions. Developed as a form of self-generated pleasure in the infant and rewarded both physically and psychologically by feelings of gratification, laughter provides a highly effective, useful, and contagious means of vocal communication. It deals with a wide range of cultural problems, often more effectively than speech, in much the same manner that crying, an infantile, probably instinctive reaction to discomfort, communicates an unmistakable emotional state to others.

The reasons for laughter in complex social situations is another question and is answered differently by philosophers and psychologists. The English novelist George Meredith proposed a theory, resulting from his analysis of 18th-century French court comedies, that laughter serves as an enjoyable social corrective. The two best known modern theories of the social wellsprings of laughter are the philosopher Henri Bergson's hypothesis that laughter is a form of rebellion against the mechanization of human behaviour and nature, and Freud's concept of laughter as repressed sexual feeling. The writer Arthur Koestler regarded laughter as a means of individual enlightenment, revelation, and subsequent freedom from confusion or misunderstanding concerning some part of the environment.

Man's vocal instrument as a device of communication represents an apex of physical and intellectual evolution. It can express the most basic instinctual demands as well as a range of highly intellectual processes, including the possible mastery of numerous complex languages, each

Capacities of the vocal organs

with an enormous vocabulary. Because of the imitative capacity of the vocal mechanism (including its cortical directors), suitably talented individuals can simulate the sounds of nature in song, can communicate in simple ways with animals, and can indulge in such tricks as ventriloquism and the mimicry of other voices. Recent tape recording techniques have even extended this flexibility into new domains, allowing singers to accompany their own voices in different keys to produce effects of duets or choruses composed electronically from one person's voice.

**Mass and public communication.** *Prerequisites for mass communication.* The technology of modern mass communication results from the confluence of many types of inventions and discoveries, some of which (the printing press, for instance) actually preceded the Industrial Revolution. Technological ingenuity of the 19th and 20th centuries has developed the newer means of mass communication, particularly broadcasting, without which the present near-global diffusion of printed words, pictures, and sounds would have been impossible. The steam printing press, radio, motion pictures, television, and sound recording—as well as systems of mass production and distribution—were necessary before public communication in its present form might occur.

Technology was not, however, the only prerequisite for the development of mass communication in the West. A large public of literate citizens was necessary before giant publishing and newspaper empires might employ extant communications technology to satisfy widespread desires or needs for popular reading materials. Affluence and interest were (and are) prerequisites for the maintenance of the radio, television, cinema, and recording industries, institutions that are presently most highly developed in wealthy, industrial nations. Even in countries in which public communication is employed largely for government propaganda, certain minimal economic and educational standards must be achieved before this persuasion is accepted by the general public.

*Control of mass communication.* Over the years, control of the instruments of mass communication has fallen into the hands of relatively small (some claim diminishing) numbers of professional communicators who seem, as populations expand and interest widens, to reach ever increasing numbers of people. In the United States, for example, far fewer newspapers currently serve more readers than ever before, three television networks are predominant, and a handful of book publishers produce the majority of the best-sellers.

Public communicators are not entirely free to follow their own whims in serving the masses, however. As is the case of any market, consumer satisfaction (or the lack of it) limits the nature and quantity of the material produced and circulated. Mass communicators are also restricted in some measure by laws governing libel, slander, and invasion of privacy and, in most countries, by traditions of professionalism that entail obligations of those who maintain access to the public's eyes and ears. In almost every modern nation, privileges to use broadcasting frequencies are circumscribed either loosely or rigidly by government regulations. In some countries, national agencies exercise absolute control of all broadcasting, and in certain areas print and film media operate under strict government control. Written and film communications may be subject to local legal restraints in regard to censorship and have restrictions similar to those of other private businesses. Traditions of decorum and self-censorship, however, apply variably to publishers and filmmakers, depending usually upon the particular markets to which their fare is directed.

Opinions on the effect of public communication

*Effects of mass communication.* Lively controversy centres on the effect of public communication upon audiences, not only in matters concerning public opinion on political issues but in matters of personal life-styles and tastes, consumer behaviour, the sensibilities and dispositions of children, and possible inducements to violence. Feelings regarding these matters vary greatly. Some people construe the overall effects of mass communication as generally harmless to both young and old. Many sociologists follow the theory that mass communication seems to influence attitudes and behaviour only insofar as it confirms the status quo—*i.e.,* it influences values already accepted and operating in the culture. Numerous other analysts, usually oriented to psychological or psychiatric disciplines, believe that mass communications provide potent sources of informal education and persuasion. Their conclusions are drawn largely from observations that many, or most, people in technological societies form their personal views of the social realities beyond their immediate experience from messages presented to them through public communication.

To assume that public communication is predominantly reflective of current values, morals, and attitudes denies much common experience. Fashions, fads, and small talk are too obviously and directly influenced by material in the press, in films, and in television to support this view. The success of public communication as an instrument of commercial advertising has also been constant and noticeable. Present evidence indicates that various instruments of mass communication produce varying effects upon different segments of the audience. These effects seem too numerous and short-lived to be measured effectively with currently available instruments. Much of the enormous output on television and radio and in print is probably simply regarded as "play" and of little consequence in affecting adult dispositions, although many psychologists believe that the nature of children's play experiences is critical to their maturation.

The role of newspapers, periodicals, and television in influencing political opinion is fairly well established in the voting behaviour of the so-called undecided voters. Numerous studies have shown that while the majority of citizens in the United States cast their votes along party lines and according to social, educational, and economic determinants, middle-of-the-road voters often hold the balance of power that determines the outcomes of elections. Politicians have become sensitive to their television images and have devised much of their campaign strategy with the television audience in mind. Advertising agencies familiar with television techniques have been brought into the political arena to plan campaigns and develop their clients' images. The effectiveness of television campaigning cannot yet be determined reliably.

Public communication is a near-ubiquitous condition of modernity. Most reliable surveys show that the majority of the people of the world (including those of totalitarian countries) are usually satisfied with the kind of mass communication available to them. Lacking alternatives to the communication that they easily and conveniently receive, most people seem to accept what they are given without complaint. Mass communication is but one facet of life for most individuals, whose main preoccupations centre on the home and on daily employment. Public communication is an inexpensive addendum to living, usually directed to low common denominators of taste, interest, and refinement of perception. Although mass communication places enormous potential power in the hands of relatively few people, traditional requirements for popular approval and assent have prevented its use for overt subversion of culturally sanctioned institutions. Fear of such subversion is sometimes expressed by critics.

Role of popular approval

## THE PSYCHOLOGY OF COMMUNICATION

Contemporary psychologists have, since World War II, shown considerable interest in the ways in which communications occur. Behaviourists have been prone to view communication in terms of stimulus–response relationships between sources of communications and individuals or groups that receive them. Those who subscribe to Freud's analysis of group psychology and ego theory tend to regard interactions in communication as reverberations of family group dynamics experienced early in life.

By the middle 1950s, psychological interest settled largely on the persuasive aspects of various types of messages. Psychologists have attempted to discover whether a general factor of personality called "persuasibility" might be identified in people at large. It would appear, though with qualifications, that individuals are indeed variably persuasible and that, at times, factors of personality are related to this quality.

Selective perception, attention, retention

Other psychologists have studied the recipients of communication, evolving concepts of "selective perception," "selective attention," and "selective retention" in order to explain not only the ways in which communication changed attitudes but also the reasons for resistance to change. Among their interests were the dynamics of the communication of rumours, the effects of "scare messages," the degree of credulity that sources of prestige value provide, and the pressure of group consensus upon individual perceptions of communications.

Some of the suggestions that emerged from the work of certain modern psychologists may be subsumed under a theory of what is called "cognitive dissonance," which is based upon the observation that most people cannot tolerate more than a specific degree of inconsistency in the environments they perceive. An example of cognitive dissonance may involve a person who considers himself a superb bowler but who on one occasion earns an extremely low score. The dissonant or inconsistent elements include the bowler's knowledge of his skill and the fact of his poor score. This produces tension. To reduce this tension—dissonance—the bowler may change his behaviour or misinterpret or reinterpret the dissonant elements in order to lessen the difference between the facts. For example, he may blame his performance on the bowling ball, the alley, or the temperature of the room. Thus he seeks a psychological equilibrium.

This modification of an individual's perception of reality is of fundamental interest to the psychologist of communications. Because the agreement or disagreement of a communication with an individual's cognitive structure not only affects his behaviour but his perception as well, the major criterion for the psychological analysis of communication is neither the message nor the medium but the expectation of the person receiving the message.

It must not be assumed that any of the theories of audience psychology offered to date (including those of Gestaltists, Freudians, Behaviourists, and others) lack relevance to an understanding of communication processes. None, however, seems to account fully for all of the effects of communications upon people. The many facets of communication offer substantial problems for future psychological experimentation and theorizing.

**BIBLIOGRAPHY.** COLIN CHERRY, *On Human Communication,* 2nd ed. (1966), a scientifically and mathematically oriented analysis of information theory and cybernetics as they relate to meaning, including a treatment of the philosophical and practical implications of the logical analysis of communication; EDWIN EMERY, P.H. AULT, and W.K. AGEE, *Introduction to Mass Communications,* 3rd ed. (1970), a comprehensive history and survey of the press, radio, television, and films in modern society, especially in the United States; G.N. GORDON, *The Languages of Communication* (1969), a detailed examination of the major elements in mass and interpersonal communication, including consideration of their current cultural roles; G.N. GORDON, *Persuasion: The Theory and Practice of Manipulative Communication* (1971), a study of the development and modification of beliefs, attitudes, and opinions by propaganda, education, and instruments of mass communication; E.T. HALL, *The Hidden Dimension* (1966), an unconventional anthropological study of "proxemics" and allied concerns, illustrated by empirical evidence, drawings, and photographs (written with style and humour for the general reader); F.W. MATSON and ASHLEY MONTAGU (eds.), *The Human Dialogue* (1967), a broad collection of articles that touch almost every phase of communication in contemporary society, including education, religion, and social problems, with contributors ranging from popes to professors; J.R. PIERCE, *Symbols, Signals and Noise* (1961), a lucid guide for the nonspecialist in the study of communication processes, information theory, cybernetics, mathematical models of communications, entropy, coding methods, and allied matters; A.G. SMITH (ed.), *Communication and Culture* (1966), a comprehensive anthology of specialized approaches to communication theory and practice, including numerous articles on language, nonverbal communication, mass communication, and other aspects of the subject, written by authorities in various fields; A.N. WHITEHEAD, *Symbolism* (1927, reissued 1958), a short, definitive study of the origins, uses, and ramifications of symbolism, both from the perspective of the individual and from the historical-social viewpoint.

(G.N.G.)

# Computer Science

Computer science is the body of knowledge pertaining to the automatic processing of symbolic information. The most important device for such processing is the computer. Computer scientists investigate the capabilities of automatic computation and develop effective algorithms, symbolic representations, and processors for such computation.

The origins of computer science lie in the quest of mathematicians and logicians to construct mechanical devices for numeric computation and formal reasoning. The theoretical advances in mathematics in the 1930s and the creation of the general-purpose digital computer in the 1940s led to the emergence of computer science, which became formally established as a scholarly discipline in the early 1960s.

Computer science encompasses not only general theoretical fields but also empirical areas that have yielded significant bodies of theory. Its most important theoretical fields have to do with the theory of computation and the development and analysis of algorithms. Empirical areas of the science include the physical organization of the components of computer systems (from the design of computer chips to that of computer networks), the development of expressive programming languages and techniques for translating them automatically, and the investigation of principal software aggregates (computer programs): operating systems and data-base systems. Two other fields of inquiry complete the spectrum of computer science: one focuses on numeric computation and studies the computer as a means of calculating values through mathematical operations; the other concentrates on artificial intelligence and pursues the elusive and promising goal of effecting cognition through automatic computation.

Because computers have pervaded so many areas of human endeavour, the work of computer scientists has had a profound effect on the way people live. It also has influenced the entire family of scientific disciplines by furnishing them with a powerful intellectual tool that hastens and facilitates scientific discovery.

For coverage of the technology of computers themselves, see COMPUTERS. For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 735 and 10/23.

This article is divided into the following sections:


The nature and scope of computer science 629
- The nature of computer science
- The scope of computer science

Development of computer science 630
- Origins of the discipline
- Landmarks in the fields of computer science

Fields of computer science 632
- Synthesis and analysis of algorithms
- Theory of computation
- Computer architecture
- Programming languages and methodology
- Operating systems
- Data-base systems
- Numeric computation
- Artificial intelligence

Role of computer science 636
- Effects of computers on society
- Computer science in the family of disciplines

Bibliography 637


## THE NATURE AND SCOPE OF COMPUTER SCIENCE

**The nature of computer science.** For the automatic manipulation of information, which can take the form of numbers, text, or other symbolic representations of concepts or data, the digital computer can be programmed to perform many diverse operations. With the aid of a computer, a business planner can find the solution to a large system of linear equations; a molecular biologist can compare the chemical structures of the genes of different organisms; an engineer can prepare the design of a new device. The computer can also process concepts, if given their encoded, symbolic representations and a set of rules for dealing with them. Consequently, computer scientists have three broad areas of concern: the methods (algorithms) for deriving information with synthetic processors (computers) and the limitations of those methods, the physical representation of symbolic information for computation, and the design of effective processing systems.

Computer science focuses on the computer as a universal information-processing machine with well-defined capabilities rather than on a specific implementation of these capabilities. Since the mid-1940s, digital computers have relied on electronic components. This does not preclude, however, such technological developments as optical computing in the future. Optical computing involves the use of light-guide circuits that employ microscopic optical fibres instead of the semiconductor-based integrated circuits of today. Such light-guide circuits would process data (as pulses of light) at much higher speeds than the very large-scale integrated circuits used in present-day computers.

The complexity of computational processes, reflected in the number of potential internal states of a computer during program execution, is far greater than the complexity of other automated processes. A principal objective of computer science is the development of techniques for handling this complexity in several levels of abstraction. For example, computer users interact with the machine through a rather simple interface that furnishes only a highly simplified view of the information-processing task: the underlying complexity is hidden. The task that the user presents to the computer undergoes several stages of translation and interpretation, first by programs resident in the computer (software) and then by the circuitry of the computer (hardware), which ultimately carries out the task. Another concern of computer science can thus be seen as the development of languages for representing, manipulating, and reasoning about symbols. Ultimately, the task specification has to be carried out by the hardware, in which the states "on" and "off" (or the symbols 1 and 0) are the only entities that have meaning.

Principal aim

As a science of an artificial symbolic world rather than of the natural world, computer science does not have naturally predefined boundaries. Many of the problems worthy of intellectual pursuit evolve with technological breakthroughs and with new areas of computer application. The interaction between the science and computer engineering is vital, and in many instances it is the rigorous pursuit of knowledge rather than the goal of constructing an economically viable artifact that classifies a project as a scholarly enterprise.

**The scope of computer science.** Although seminal work outlining the capabilities of automatic computation dates to the 1930s, computer science took shape in the early 1960s. It was only with the rapid development of the electronic digital computer in the 1950s and the resulting stimulation of research that computer studies coalesced into a scientific discipline.

Computer science is subdivided into several specialized areas. They are outlined here and are treated in detail in the section *Fields of computer science.*

The construction of effective and efficient procedures for solving distinct classes of problems is the subject of work on the synthesis and analysis of algorithms. The theory of computation is concerned with the limits and complexity of automatic computation, as well as the formalisms that describe languages and automata of varying powers.

The field of computer architecture deals with the organization of multiple levels of computer hardware, from elementary switching circuits to multiple computers in complex computational systems. The fields of programming languages and methodology, operating systems, and data-base systems investigate the general aspects of computer software.

The field of numeric computation pursues the original application of computers—namely, the manipulation of numeric values. Much of the future promise and present use of computers is determined by the field of artificial intelligence, the objective of which is the discovery of how computers might perform tasks considered to require intelligence when done by humans.

## DEVELOPMENT OF COMPUTER SCIENCE

The roots of computer science

The crucial developments that led to the establishment of computer science were fundamental theoretical discoveries concerning automatic computation that were made in the 1930s and the development of the general-purpose stored-program electronic digital computer in the 1940s. The deeper roots of computer science, however, lie in the efforts, undertaken throughout the centuries, to design mechanical devices for numeric computation and for reasoning. The interplay of theoretical and practical discoveries continues to influence the development of the science: its progress alternates between the invention of novel system organizations or new programming languages and the analysis of the results of those developments.

**Origins of the discipline.** The history of the tools of computation begins with the invention of the abacus several thousand years ago, perhaps in Babylonia. The first calculating devices to have a major influence on the modern world, however, were the mechanical calculators built in the 17th century by Blaise Pascal of France and Gottfried Wilhelm Leibniz of Germany.

During the first half of the 19th century the English mathematician Charles Babbage prepared designs for a mechanical computer that he called the Analytical Engine. Although Babbage never completed his device, he identified the principal components of a digital computer. He described how programs could be stored on punched cards similar to those used with the Jacquard loom, which had been introduced in France in 1805.

The abacus, mechanical calculators, and Babbage's Analytical Engine were digital machines in which numbers and arithmetic operations were represented symbolically. A different approach to computation was developed by the inventors of analog devices, in which data are converted into measurable physical properties of the device; these properties can be altered so as to imitate computation. In the early 19th century, mechanical planimeters were used to compute the area bounded by a continuous curve; in later analog devices electric voltages were manipulated to simulate calculations. The work on this type of computer culminated with the differential analyzers constructed by the American electrical engineer Vannevar Bush in 1930 and 1942. Analog computers have been used mainly as adjuncts to digital computers in so-called hybrid computer systems and have not significantly influenced the progress of computer science.

The English mathematician George Boole extended the scope of computing beyond its previous limitations to numerical quantities. Boole continued and expanded Leibniz' work in developing a method "in which all truths of the reason would be reduced to a kind of calculation." In 1847 and 1854 Boole published demonstrations that the propositions of logic can be expressed by a set of mathematical rules known today as Boolean algebra. Boolean algebra is a cornerstone of computer science: its direct influence on the design of digital computer circuitry dates from the work of the American mathematician Claude Shannon, who in 1938 pointed out that the the behaviour of switching circuits, which have only two signal states ("on" and "off"), could be expressed in a two-value algebra.

Boole and two-value algebra

Boole's work led to epoch-making results in the field of logic. In 1931 the Austrian-born mathematician Kurt Gödel proved that all sufficiently powerful mathematical systems are inherently incomplete because in such a system theorems exist that can be neither proved nor disproved using the system alone. Gödel also demonstrated that all symbols can be encoded as nonnegative integers; thus encoded, a mathematical proof may be treated like any other operation involving integers. The British mathematician Alan Turing continued the work of Gödel while focusing his efforts on creating an automaton that could determine whether a mathematical statement was provable. In 1936 he conceived an abstract computing device now referred to as the universal Turing machine. Because it was capable of accepting a "program" (instructions)—encoded according to Gödel's numbering system—along with data input, the machine represented a general-purpose computing device. Turing went on to prove the existence of problems that cannot be solved by any automatic computation. The best known of these is the halting problem, which shows that there can be no general algorithm for deciding whether a machine will stop (end computation) when carrying out an arbitrary program. Since Turing's time, many other problems have been proven to be unsolvable because they ultimately can be reduced to the halting problem.

The Turing machine

Also in 1936 the American mathematicians Alonzo Church, Stephen Kleene, and Emil Post independently introduced symbolic computation devices, all equivalent in computational power to the Turing machine. All these devices were able to compute the same class of functions, understood broadly as the class of solvable problems. This lends credibility to the claim known as Church's thesis that all these models indeed determine the broad limits of automatic computation.

The computers of today, however, originated in the 1940s. In 1946 the first general-purpose electronic computer was completed. Although this computer—called ENIAC, an acronym of Electronic Numerical Integrator and Calculator—could be programmed to do different tasks, this programming required a partial rewiring of the machine.

It was the concept of the stored-program computer, credited to the Hungarian-born American mathematician John von Neumann, that brought about the development of digital computers in their present form. In this computer the instructions (program) and data are kept in a common memory so that they can be processed in a uniform fashion, as appropriate to their respective roles in the computation. The first operational stored-program electronic digital computer (what is today called simply a computer) was completed at the University of Cambridge in 1949 under the direction of Maurice Wilkes, an English mathematician.

Since then, computer technology has undergone several generational changes. Changes in computer hardware have included shifts in electronic technology (from vacuum tubes to transistors to integrated circuits to very large-scale integration), the creation of single-chip processors (microprocessors), and the development of extensive networks of computers. Programming tools also have undergone generational changes: binary machine languages gave way to progressively higher-level programming languages capable of better supporting human thought processes about problem solution. Programmers no longer need concern themselves with every detail of program execution; they now can rely on the established base of support offered by complex operating, data-base, and telecommunication software systems.

Building on these theoretical and practical advances, computer science congealed into a scholarly discipline in the early 1960s. The first academic departments of computer science were established in 1962 at Stanford University, Stanford, Calif., and Purdue University, West Lafayette, Ind.

The first academic departments

**Landmarks in the fields of computer science.** The brief history of computer science is a complex chronicle of merging disciplines and concurrently developing related fields. This section discusses the history of computer sci-

ence by concentrating on the principal developments in the individual fields that constitute it. The intellectual content of these fields is covered in the following section of this article.

*Computer architecture.* The organization of the hardware components of computer systems has been based to a large degree on the so-called von Neumann machine. In this hardware design a single stream of machine-language instructions is applied to a single stream of data, both instructions and data residing in a common memory. The technique most often used to design the control unit (which translates and executes instructions) of computers, called microprogramming, was invented by Wilkes in 1951 and implemented in 1964. The organization of computer memory into a hierarchy of units with varying speeds and capacities was formulated in the late 1950s and early '60s. Multiprocessor designs, where two or more processors share a common memory, have been operational since the early 1960s.

In 1965 the first commercial minicomputer was installed. Although less powerful than mainframe (large-scale) computers, minicomputers found a wide market because they are cost-effective in many scientific and business applications. Continued efforts at size reduction resulted in the development of the microprocessor, a central processor that consists of a single silicon chip, in 1971. At about this time, it became apparent that further increases in processing speed would have to come from advances in architectural design rather than from increases in the speed of the electronic components. Innovative designs of supercomputers and of multiprocessors with large numbers of processors were produced, and some were implemented. While some of these designs are extensions of the von Neumann architecture, others constitute radical departures.

Development of computer networks

Rudimentary computer networks emerged in the late 1950s. Intensive research and development in this area, starting in 1968, has resulted in networks that permit a broad range of user access and worldwide resource sharing. Local-area network techniques were developed in the 1970s to connect computer systems within limited areas, such as a building or a campus.

*Programming languages and methodology.* The original method of programming, in which each instruction had to be expressed in the binary machine language of 1s and 0s, was tedious and impractical. Therefore, it became necessary to develop programming languages that would help humans to communicate more easily with computers. Mnemonic representations of instructions and data, which appeared in the late 1940s, soon evolved into assembly languages, each specific to a family of computers.

Assembly languages were still too inconvenient for large programming tasks. Higher-level languages (those closer to natural, or human, languages) were needed. Such machine-independent languages would facilitate algorithmic problem solving. The study of the syntax and semantics of higher-level programming languages and the research on their software translators have become a subfield of computer science. The first program for translating a higher-level language into machine language, called a compiler, was developed in 1952.

Study of higher-level programming languages

Although research on higher-level languages was begun in 1945 by the German engineer Konrad Zuse, the first reasonably complete such language, FORTRAN I, was developed during the years 1954 to 1956. The field reached maturity with the introduction, in 1963, of the language ALGOL 60, which offered a well-defined set of structures for controlling execution flow and the scope of data entities. This language soon became generally accepted for algorithmic notation.

The late 1950s saw the development of LISP, the fundamental language of artificial intelligence. In 1960 COBOL emerged as the language of data processing; it demonstrated the importance of separating the definitions of data from the procedural components of a program. In the mid-1960s BASIC was introduced as a simple language for interactive computing, a then-novel mode of user–system communication.

In the early 1970s Pascal (based on ALGOL) was developed as a simple and efficient language for methodical, structured programming, and C was invented as a language for programming operating systems.

The next generation of programming languages was influenced by the concept of data abstraction that was introduced by the language Simula 67 in 1967. It enabled programmers to isolate various aspects of problem solution into software modules called packages, each of which provided a set of operations on internally hidden (and potentially complex) data structures. Prominent among such languages is Ada, which was developed beginning in the mid-1970s. Ada offers extensive facilities for parallel programming, so that several tasks may be executed concurrently.

Modular languages

Studies gradually broadened to encompass programming methodology issues and to support software engineering, which was first recognized as a field of practice in 1968. Since then, researchers have addressed—with limited success—the problem of proving program correctness.

*Operating systems.* Originally, programmers had to operate computers themselves. In 1956 the first software monitor was created to give programmers easier access to the hardware. Operating systems thus emerged as a means of organizing the resources of a computer system for easy use and efficient management. The first operating systems of the 1950s increased the speed of computer processing by permitting batch processing, in which a series of different programs could be executed without interruption: as each one ended, the next was automatically initiated.

In the early 1960s multiprogramming operating systems were created to maintain several programs simultaneously in a state of execution in order to make more efficient use of the hardware resources. Virtual memory, a concept formulated in 1959, made it possible to apparently extend the main memory space to accommodate these multiple programs. Time-sharing operating systems, invented in the early 1960s, schedule the access of many users to a single processor. Because a computer can execute millions of operations per second and because human reaction time is slow, many users can interact with the system at one time. The late 1960s saw a consolidation of batch, time-sharing, and "real-time" (driven by external signals that require rapid response) modes in large operating systems.

The experience gained in the development of these systems led to the systematic study of operating systems. In 1965 the crucial problem in developing operating systems was recognized as the management of interacting parallel processes. In 1968 it was shown how to structure an operating system as layers of routines, each performing a limited and specific function.

*Data-base systems.* One important function of system software is the maintenance of large organized collections of data. In the late 1950s, file management systems were created for this purpose. They used simple relationships (*e.g.*, a sequence of identifying values) to store and retrieve records. In 1964–65 the first true data-base management systems emerged. These hierarchical systems provided a richer set of relationships among records. A more general approach to data storage and retrieval, the relational model was described in 1970. The theory of relational data bases and that of conceptual modeling of data relationships have since evolved into important topics of research.

Development of data-base management systems

*Theory of computation.* Continuing the mathematical work of Gödel and Turing, researchers have constructed the theory of computability, delimiting what can be automatically computed. In 1956 Noam Chomsky established a classification for formal grammars; computer scientists soon realized that programming languages largely conform to the context-free grammar described in this classification. Gradually, the attention of researchers in the field turned toward the problem of determining the computational complexity of broad classes of solvable problems. The main interest here is the determination of the theoretical minimum of the complexity of a problem, regardless of the complexity of the algorithm actually used to solve it. Complexity theory was outlined by 1960, but the mature framework of the field was not formulated until 1965. In the early 1970s a class of important problems—the so-called NP-complete—was identified as solvable, even though they are believed to be intractable in their general

form; that is, their solutions would require centuries of computing time.

Researchers are studying the semantics of programs in an attempt to produce a system that will automatically verify program correctness. In the late 1960s a language was developed in which "correctness" may be argued, but the goal of automatic verification appears to be remote.

*Artificial intelligence.* Artificial intelligence, or AI, is the field of computer science whose goal is to expand the qualitative capabilities of the computer by devising methods by which it can perform tasks that require cognitive abilities. The idea of cognition as a form of computation has its roots in the philosophical speculation of the 17th and 18th centuries. This ultimate potential of computers was foreseen by Alan Turing, who in 1950 formulated a test, named after him, whereby a machine can be deemed to perform intelligently if an interrogator using a remote terminal cannot distinguish its responses from those of a human.

In early research in the field, from the 1940s to the mid-1950s, the approach adopted by researchers was to use a computer to simulate the functioning of the neural nets of the brain. The field was named in 1956 by John McCarthy, who also equipped it with its lingua franca, a list-processing programming language called LISP.

Research from the 1950s to the mid-1960s concentrated on finding a general problem-solving technique. The method of heuristic search was invented: this trial-and-error search for a solution used problem-dependent rules of thumb to limit the boundaries of possible solutions. The emblematic product of this period of research was the General Problem Solver, created in the late 1950s.

Knowledge-based systems

During the mid-1960s further progress was achieved by creating knowledge-based systems. These systems contain interrelated information about a specific domain (a disease, for example) from which they can reason about that domain (render a diagnosis, for example). The reasoning process, called inference, can be based either on a form of heuristic search or, preferably, on a causal model of the domain. The knowledge-based approach has led to progress in the simulation of human cognitive faculties. By the late 1970s artificial intelligence had arrived as a discipline that was generating many practical results. In the early 1980s this work led to the establishment of the "fifth-generation research agenda" (originating in Japan), the goal of which was to organize computer systems that draw inferences rather than merely calculate.

## FIELDS OF COMPUTER SCIENCE

This section discusses the fields of computer science in a sequence that demonstrates the structure of the entire discipline. The demarcation between these areas, however, is neither rigid nor static: developments in one of the fields often cause structural changes in related fields.

**Synthesis and analysis of algorithms.** An algorithm is a finite set of unambiguous instructions that, when supplied with input values of a predefined nature, produce predefined outputs that result in the solution of a problem. Because algorithms are a central concern of all areas of computer science, many theoreticians concern themselves with the construction and analysis of efficient algorithms.

Classes of algorithms

Important classes of algorithms include those for sorting a list, searching for an item in a collection of data, pattern matching (looking for a specific word or phrase within a body of text), or generating random numbers so that sequences of apparently unrelated values can be used in computer simulation or in numeric techniques.

An algorithm is encoded in a programming language before it is presented to the computer. The efficiency of an algorithm is most often determined by measuring the time it requires to solve a problem of specified size. For example, the size of a problem that requires the counting of paths in a graph is the number of edges in the graph; the efficiency of a sorting algorithm is expressed as a function of the number of items to be sorted. The memory space that the computer requires to execute a program based on an algorithm is usually of secondary importance.

Sorting and searching operations on large collections of data consume much computer time. The following example uses these operations to demonstrate algorithms and make this primary concern of computer science tangible. In a table of data composed of individual entries, each entry is identified by a unique number (the "key"). For example, in a table of employee records, each entry might have the following components: identification code (the key), name, and salary.

If the records are placed in the table in a random order, to find the salary of an employee with a given identification code it is necessary to look at all of the records consecutively until the desired record is found. Looking through a table of $n$ records with such a sequential search algorithm requires that the computer access an average of $n/2$ records per search. Thus, the number of operations required in accessing a record in an unsorted table is proportional to $n$, or the computing time grows linearly with the problem size.

On the other hand, if the table is sorted into a sequence in which the key numbers are arranged in ascending order, the binary search algorithm could be used to find a specific record. This algorithm directs the program to look first in the middle of the table: if the key being sought is larger than that of the record in the middle, the program searches the "larger-key" half of the table by the same procedure, until the record with the desired key is found. In this case, because the search area of the look-up procedure is continually halved, the computing time is proportional to $\log_2 n$. The immense increase of the algorithmic efficiency is illustrated by the fact that, if $n$ increases from 1,000 to 1,000,000, the time required for the first algorithm increases by a factor of 1,000, but that required for the second merely doubles. Once sorted, a table may be searched efficiently any number of times. An efficient algorithm is often obtained through an insightful organization of data into data structures.

Recursion

A powerful technique used in the development of succinct algorithms is recursion, or expressing an algorithm partially in terms of itself. Note that the previous binary search algorithm is recursive: it looks at the middle item in the table and terminates only if this is the desired record; otherwise it performs the same binary search on the proper half of the table.

Computer scientists have devised many broadly applicable techniques of algorithm design. These include the "divide-and-conquer" method that splits a problem into subproblems. If the subproblems are of the same nature as the original problem, the algorithm can be expressed recursively. Other design methods include "backtracking" and "branch-and-bound," both techniques for systematic trial-and-error searches for solutions that satisfy certain criteria, with prompt rejection of unpromising alternatives.

**Theory of computation.** Computation theorists address the fundamental questions of computer science: the limits of automatic computation, the feasible efficiency of solutions to problems of general interest, and the formalization of the logic of programs so that they can be verified automatically.

The primary concern of studies in the theory of computability is the question of what can be computed. Because formal argument about the theoretical limitations of actual computers is not feasible, this theory is concerned with the analysis of automata and their corresponding formal languages. Automata are abstract models of computing devices that range in "power" from Turing machines (equivalent to general-purpose computers) to finite-state machines (restricted to a small number of internal states). A problem is proved unsolvable if no Turing machine can give a solution for all input values.

Each class of automata can recognize a certain class of symbolic expressions, such as text or real numbers. These expressions are generated by a grammar that specifies the syntax of the given formal language. As part of the formal study of computer systems, a hierarchy of increasingly powerful automata and correspondingly more expressive formal languages has been constructed. Some formal grammars play a significant role in the translation of programming languages.

The prominent research in this field concerns the theory of computational complexity. By studying the complexity

Computational complexity

of a problem, researchers try to ascertain the shortest computing time, least memory space, and fewest processors required for the solution of a problem, regardless of the algorithm. The complexity of the best available algorithm for the solution of a problem is compared with the complexity of the problem to decide whether a better algorithm can be devised. For example, an $n$th-degree polynomial cannot be evaluated with fewer than $n$ multiplications and $n$ additions. Because the long-existing algorithm for the solution of this problem (Horner's rule) matches this complexity, it is known that its performance cannot be bettered. In contrast, the known algorithms for matrix multiplication have a complexity of a higher order than the problem itself, making the development of a new algorithm for this frequently used operation a promising research subject. Recent research has concentrated on the question of whether the minimum computing time needed to solve certain crucial problems grows polynomially (as a power of the problem size, $n$) or exponentially (as the $n$th power of 2). The availability of a polynomial-time solution would mean tractability of the problem.

Another subfield of research is the process of formalizing and axiomatizing programs to facilitate reasoning about how they will behave when they are executed. The goals of these studies include automatic proofs of program correctness, automatic construction of programs from specifications of their intended function, and program optimization for the fastest execution.

**Computer architecture.** Computer architects study the design and organization of computer systems and their components on several levels, while considering such attributes as speed, cost, functionality, and reliability.

Theory of switching circuits

The theory of switching circuits is the foundation for the design and optimization of the electronic components of computers. Combinational switching circuits consist of so-called logic gates that perform Boolean operations so as to make the output values dependent only on the values of the current input. An "AND gate," for example, produces output only if it receives two inputs. Sequential switching circuits, which are more complex, include "flip-flops," (circuits that store single bits of data) and therefore can exist in distinct states that depend on previously supplied inputs. Very large-scale integration (VLSI) has made it possible to place millions of logic gates or flip-flops on a single semiconductor chip.

Designing the functional subassemblies of the computer (central processor, memories, and input–output processors) and melding them into a balanced system is the concern of computer organization. The central processing unit (CPU) performs the arithmetic-logic and control functions of the computer. The arithmetic-logic unit contains circuits that carry out arithmetic and Boolean operations. Separate algorithms have been devised for dealing with integers and real numbers. Limitations on the accuracy with which real numbers can be represented pose problems that continue to occupy computer scientists.

The control unit responds to each successive instruction in the machine-language program by issuing signals to other units and devices. In most control unit designs, the control signals are stored in the special control memory as a so-called microprogram. Some control units, however, are implemented as switching circuits because this method is faster.

Memory, in which instructions and data are stored, is organized into a hierarchy in which the units logically close to the CPU are faster and smaller than more remote memory units. Designers can use this cost-effective organization because, for efficient operation, the CPU needs very fast access to only a small fraction of the total instructions and data that are stored within a system.

Input–output processors control the devices that communicate with the outside world. Interconnections between units of the computer system generally consist of shared data buses (multiplexed paths that allow numerous devices to transmit data simultaneously).

Researchers in the field of computer architecture also study the organization of several processors into multiprocessor systems and multiple systems into computer networks. Multiprocessing, in which several processors share a common memory, is one means of increasing the degree of parallelism in computer systems; research on massive parallelism aims to organize thousands of microprocessors into a single system. Supercomputers are another area where two approaches have evolved: array processing and vector processing. Array processors contain many arithmetic-logic units, each of which executes the same instruction on different data as directed by a single control unit. Vector processors contain one or more "pipelined" arithmetic-logic units; such a unit can process several pairs of operands simultaneously.

Multiprocessing

In the subfield of computer networks, researchers study the organization of multiple computer systems into integrated networks linked by communication lines. Investigations include the study of network topology and protocols for establishing and maintaining reliable and secure communications between connected systems.

Among other subfields of computer architecture, research on system reliability bifurcates into work on fault elimination and fault tolerance. The development of automatic techniques of hardware synthesis is another area of study. The subfield called computer graphics combines hardware approaches with software algorithms to facilitate the manipulation of visual information.

**Programming languages and methodology.** The architecture of a computer provides a binary machine language that the processor interprets and executes directly but that is not used for programming. Programs are sometimes written in an assembly language—a simple low-level, mnemonic encoding of the computer's built-in machine language. Although a so-called macro facility is often incorporated to extend the power of assembly languages, most programming is done with far more expressive higher-level languages. Research in the field of programming languages and methodology is concerned with developing languages that further reliable programming as a methodical thinking process, with the automatic translation of these languages into machine language and with the development of methodologies for producing large and reliable programs.

Many higher-level programming languages have been developed since the mid-1950s, and a large body of knowledge about their symbols has been accumulated. Although the description of programming language syntax (which defines the symbols and sequences of symbols that are valid in each language) has to a large extent been formalized, work continues on the formalization of language semantics, or the effect of language symbols in a program.

The fundamental constructs of programming languages have been codified. These constructs include data types, which serve to define data according to potential values and operations. Among these, structured types, such as arrays or records, permit programmers to organize complex data and thus simplify the procedural part of programs. Other language constructs are control structures, which specify the order in which program statements must be carried out. An effective programming language includes the following control structures: sequence (permitting the consecutive execution of instructions as written), decision (permitting the selection of one of two or more possible execution paths depending on the existence of a specified condition), and loop (permitting the repeated execution of a group of statements while a certain condition exists). Other structures are included as a convenience to programmers. Because programs that address complex problems are commonly extensive, yet another group of constructs has been created to modularize programs. These constructs help manage complexity by separating the concerns and design of a program into a hierarchical structure of short modules, or groups of statements that are used as single units. Each module encapsulates a particular aspect of the overall program task, and some call upon other modules to perform subtasks. Units of modularization implemented in various languages include the following: procedures, which compute many values; functions, which compute a single value; packages, which provide operations on hidden data entities; and tasks, which execute in parallel with the modules that cause them to be executed. Thus, in procedure-oriented languages

Elements of programming languages

(such as FORTRAN, COBOL, or Ada), a procedure or similar type of module is the principal unit for segmenting various aspects of the program. Functional languages (pure LISP, for example) use function application as the only control structure; logic programming (PROLOG, for example) presents the logical constraints of the problem rather than a solution algorithm; and in object-oriented languages (Smalltalk, for example) data objects include sets of operations to manipulate the data, and programs are built by ordering these objects to apply the operations.

Application of metalanguages

Computer scientists use metalanguages, which contain their own symbols, to provide rigorous descriptions of programming languages. These descriptions are needed to use a language with confidence, to facilitate the translation of programs, and to give theoreticians a basis for discussion of the validity of a program. The most widely used language for syntax description is the Backus-Naur form. The formal semantic definition of a programming language is very difficult. An operational approach to programming semantics, which consists of defining an abstract machine that executes a program in a target language, led to the development of the Vienna Definition Language as the means of describing syntax and semantics.

A programming language is only one (albeit the essential one) of the tools used in program development. Research on programming methodology aims to establish the principles that govern the construction of reliable programs that can efficiently solve a problem yet can be easily modified to meet changing requirements. Although many techniques for producing reliable designs of large software products have been articulated and exemplified by application, only the initial elements of a theory of program construction and verification exist.

Software design

An understanding of good software design is gradually evolving. Research has shown that program design should be based on hierarchical structure and functional modularization, with each module independently working on only certain aspects of the overall task and communicating with other modules via minimal data interfaces. Proper development process for software products includes problem analysis (sometimes involving the development of a prototype system), logical and detailed program design, and program implementation, which entails encoding a program in a selected programming language and rigorously testing both the individual components and the program as a whole.

A major concern for researchers in this field is that software operate correctly. Formal techniques for testing programs, using the extensive support of specialized utility software, are being developed. Because validation by testing cannot give absolute proof of correctness (every situation can never be tested), research here overlaps with studies in the field of the theory of computation that aim at correctness proofs.

Languages that permit the formal specification of what a program will accomplish (as opposed to programming the detailed algorithm of how it is to be done) are being developed. These very high-level languages will permit programmers to state a problem in terms that will automatically produce the algorithm for its solution. This work is part of the research on automatic programming. Much of programming is a creative activity, however. Thus the development of increasingly sophisticated programming environments (or integrated software systems that support program development) offers more significant immediate promise for the field of software engineering.

Automatic programming research

**Operating systems.** An operating system, generally implemented entirely as software, facilitates the use of computer system resources. The operating system is the program that responds when a user initiates an interaction with a computer. This system keeps track of each user's permissions to access files and execute programs in a shared computer installation, maintains user programs and data files in directories, allocates the requisite hardware resources (such as main memory or disk space) as programs execute, and makes available the necessary system software (such as a language compiler). The general role of an operating system is, first, to guide the user's program during its execution by providing the needed resources (hardware, software, data files, and telecommunications). The operating system also ensures security of access in a shared computer system, provides a friendly interface to various classes of users, and allocates system resources efficiently.

Because most operating systems are dependent on specific families of processors (although broadly applicable systems, such as UNIX, have become popular), a systematic body of knowledge emerged slowly. A major accomplishment of researchers in the field has been the identification of a set of abstractions, mechanisms, and policies that are required to maintain the status of concurrently running programs (called processes) as they use the multiple resources of a computer system. Under the control of a multiprogrammed operating system, one of the processes can run on the central processor while several other processes can simultaneously perform input or output operations.

Operating systems include mechanisms such as queue data structures, for waiting processes, and scheduling policies (*e.g.*, first come, first serve) that allocate processors to processes and permit the parallel use of system resources. Families of scheduling algorithms have been studied, and general and efficient methods of synchronizing processes that cooperate in common computation have been identified. Procedures have been formulated to prevent deadlocks, making it impossible for two or more processes to hinder one another's operation.

The more programs that are running concurrently, the more memory a computer requires. A general technique, called virtual memory, organizes the main and secondary memories into an aggregate that combines the high speed of the former with the large capacity of the latter. Employing this technique, the operating system retains in the main memory only those parts of programs and data that will be needed in the immediate future and shunts the other parts of programs to secondary storage until they are needed.

Not only must coexisting processes be protected from each other's actions, but the entire computer system must also be protected from illegal access attempts. Safeguards have been developed to control the access of users and processes to system resources (*e.g.*, data files). The ultimate goal is to devise a set of controls whose presence can certify the security of a system.

Study of operating systems

The study of operating systems has led to the establishment of design principles for other complex time-dependent software systems. Thus, operating systems have become prototypes for all real-time programs that must be designed as a group of cooperating processes. The general organizing principle is one of stratifying the system into levels of abstraction. Each level is responsible for managing certain classes of objects (files, devices, and so on) and providing services (operations) for the higher organizational levels, from which the details of these operations are hidden.

**Data-base systems.** Data bases are collections of information that are maintained for an extended period of time and are stored together with certain relationships that reflect the meaning of the data. The study of data-base systems is concerned with the effective and efficient organization of data aggregates so that they can be accessed and modified by many categories of users.

Records of the same type, each consisting of data items and representing information about a certain entity, are organized into files, a practice that was established in the early business applications of computers. A collection of employee records, for example, can be stored permanently as a single file. The distinguishing feature of data bases is that such files are integrated with others by using semantic relationships that are stored along with the data. For example, if the record types in a company data base include "employee" and "department" records, relationships between the two record types might be established as "works for" and "manages."

As reservoirs of information, data bases provide a focus for many application programs; it is therefore crucial to ensure that even though the structure of the data base is modified, the programs using the data need not be

changed. This data independence is accomplished in two ways. First, the layout of data in secondary memories is such that it can be changed without affecting the overall logical structure of the data base. Second, portions of this logical structure are defined as "views" for individual application programs, so that a change in the overall structure does not affect the program in which the view is used.

Data models have been studied as the principles for representing data and relationships. One of these, the relational model, uses strictly defined operators for data extraction and offers the potential for automatic verification and optimization. Record types in the model are represented as tables in which a relationship between two types can be expressed implicitly by using a common data item in both record types. For example, the records of "employee" and "department" files can be related by the data item "department number" if it is included in both files. Alternative data models have been proposed.

To ensure data-base integrity, semantic constraints in the data base are employed to prevent invalid updates. Many users simultaneously access or modify the values in a data base. When one user transaction modifies a data entity, the entity has to be locked from all others until the transaction is complete. Access protocols that accomplish this with maximum overall concurrency are being developed, as are security rules that protect against the extraction of information from statistical data bases.

Data-base research

Because data bases are being used in networks, researchers study the principles of efficient distribution. Computers whose main purpose is to manipulate data bases are being designed to increase data-handling efficiency. The structure of multimedia data bases and user interfaces—which incorporate text, sound, and images—is also under investigation. Another direction of research is the merging of data bases and knowledge bases, primarily investigated by those in the field of artificial intelligence.

**Numeric computation.** Researchers in the field of numeric computation have significantly expanded classical numerical analysis by using computers to solve numeric problems. Because fully analytical solutions are impossible for the vast majority of mathematical problems occurring in practice, numeric methods must be used to obtain the desired results. The goals of numerical analysts include the development of efficient algorithms of predictable accuracy for various classes of numeric problems and the invention of methods for ensuring the stability of computation and error control.

The impetus for the development of this field has come from the simulation and optimization models used in activities such as weather forecasting, design optimization, and the control of spacecraft. The execution of models (simulation) has become a widely employed mode of experimentation in the natural sciences.

There are several sources of errors in computerized numeric computations, the most important being the limitations on representing numbers in computers. Because computer hardware limits a calculation to a finite number of significant places, round-off errors have a cumulative effect in a series of computations and make computer arithmetic inexact. Traditionally, forward error analysis has been performed to find the bounds to errors. This has recently been augmented for some classes of computations by a powerful backward error analysis, in which round-off errors are shown to be the equivalent of a perturbation of the data presented to the program. This permits the analysis of error sensitivity to be reduced to the analysis of the accuracy of the original data. Research has continued on the ability to estimate the convergence error in the iterations, which produce a sequence of successive approximations. Such iterations are terminated when a specified error criterion is satisfied.

Error analysis

Another general direction of work in this field is the analysis of problem stability and the devising of stable methods of computation. Problems are considered stable if small perturbations in data produce only commensurately small changes in the solution. The solution methods of stable problems also are studied because they themselves might be unstable, resulting in a progressively magnified round-off error as the computation continues.

Probably the most important class of problems researched in this field is the solution of differential equations, which specify relationships between quantities and the rates of their change and which are used in models to express most of the dependencies of natural phenomena. One powerful tool for the solution of partial differential equations is the finite element method of approximating a continuous system by an assembly of discrete elements.

Curve-fitting techniques such as spline functions, which approximate curves and surfaces by means of smoothly joined segments, are often employed in computer graphics. The fast Fourier transform (FFT), a group of algorithms particularly suitable for examining the behaviour of functions, has found broad application in signal processing, which is integral to such activities as speech recognition and radar observation and measurement.

Other areas of research in numeric computation include numeric differentiation and integration, the Monte Carlo method of solving very complex problems by probabilistic approximations, and algorithms for parallel processing. Notable, too, is the development of adaptive methods of computation, in which the analysis of the solution to a problem as it emerges during the execution of the program directs the selection of alternative algorithms for further execution.

**Artificial intelligence.** In the field of artificial intelligence, computer scientists study methods for developing computer programs (or software–hardware systems) that display aspects of intelligent behaviour. Intelligence in humans is characterized by goal-oriented adaptive behaviour; that is, a capacity to act effectively in novel situations through understanding and reasoning. Although the application of the notions of intelligence (or understanding) to machines has engendered controversy, achievements in the field are manifest and a research agenda has been promulgated. Paradoxically, more significant progress in this field has been made in areas that are deemed to require high intelligence from humans than in those areas in which a person requires only commonsense judgment.

AI systems

As a result of artificial intelligence research, there are now systems with which users can communicate in a restricted natural language. Other AI systems can offer a diagnosis within a medical specialty, analyze the structure of a chemical compound and suggest pathways for its synthesis, or manipulate manufacturing robots to perform a set of useful tasks. Indeed, research in this field progresses largely through the development and refinement of paradigmatic programs or systems.

The following discussion treats the fundamental approaches of solving artificial intelligence problems and then reviews the specific research that is directed at producing the equivalents of human cognitive faculties. It is the view of some that research in this field should attempt to understand these faculties; others believe that the goal should be merely to reproduce their effects.

Heuristic search, or the discovery of a solution to a problem (be it winning a chess game or proving a theorem in logic) through the disciplined generation and evaluation of a number of promising possibilities, has always been a fundamental technique in the field. This trial-and-error process of moving toward a solution is employed in common problems for which it is possible to establish the nature of a good solution ahead of time (a winning position in chess, for example) but where there is no algorithm that generates that solution directly. In nontrivial problems the possible solutions worthy of examination are so numerous that an exhaustive search is impractical. Heuristic search limits the solution paths that are considered at any given point to the promising ones. Heuristics do not guarantee the optimality of a solution or even that a solution will always be found; however, they generally produce good enough solutions and render problems tractable. Important classes of problems—such as theorem proving, action planning for robots, or playing master-level chess—have been programmed using heuristic search.

Knowledge representation

Although heuristic search remains an important method of "inferencing" (generating new facts from existing data), the emphasis on inference techniques has given way to a general approach based on knowledge representation.

This approach derives from the realization that much of what is called intelligent behaviour is based on accumulated knowledge, coupled with the techniques of inference. Thus, a knowledge base of interrelated concepts or facts is needed by an expert system to arrive at conclusions about the presence of an oil deposit at a given site, by a natural-language processing program to understand an utterance, or by a computer-vision system to interpret a scene. In effect, all such knowledge-based approaches to artificial intelligence problems rely on the explicit representation of knowledge in the limited domain addressed by the system, so that inferences may be drawn and new knowledge produced. This is the task of another component of these systems, the inference engine, whose structure is related to the knowledge representation method.

Several methods of knowledge representation that lend themselves to inference have been devised. Semantic networks are a form of knowledge representation in which nodes represent concepts and links between nodes represent relationships between the concepts. An example of this is the so-called IS-A link, which imposes a hierarchy on a network. Thus, if the system knows that a "dog" IS-A "mammal," the node "dog" inherits all of the properties of the "mammal" node, while only the specializing properties of the "dog" need to be stored in the "dog" node. An alternative representation is that of logical notation; here, using a variant of predicate calculus, programmers assert what is true about the domain of the system. Logical deduction is used as the rule of inference. Thus, a well-known proof method, called resolution, proceeds by refutation: it combines the negated statement to be proved with a series of statements in the knowledge base; if a contradiction results, the original goal statement is proved to be true.

Expert systems

Expert systems are most often constructed so that concepts and knowledge about the domain are stored as sets of facts and rules as follows: IF (condition) → THEN (action). For example, a fault-diagnosis system might contain the following rule: "IF the fuse is getting voltage and the dashboard lights do not glow, THEN there is a 70 percent probability that the fuse is bad." Inference is performed through either forward chaining (moving from the conditions to the conclusion) or backward chaining (hypothesizing a conclusion and moving toward the appropriate rules and presented data). The ability to produce fruitful hypotheses (abductive reasoning), much like the human thought process known as a "leap of imagination," might in the future determine the power of expert systems. Such systems are also equipped with knowledge-acquisition facilities and an explanation interface that presents the reasoning chain to the user.

The capability of learning is the element of intelligence that most expands the power of an artificial intelligence system. Humans learn by accumulating new knowledge within an existing framework of concepts and relations and by modifying their knowledge base (with refinements to concepts and generalizations about relationships). The best expert systems also learn from experience. One program that deduces the structure of organic molecules from mass spectra and other data can formulate general rules of mass spectrometry based on the observations of individual compounds. This generalization process is controlled by meta-heuristic knowledge (knowledge about the formation of new knowledge). Because humans also learn much by analogy, researchers are investigating methods that will permit programs to simulate this type of human learning.

Research on natural language processing

Another human faculty is the ability to produce and understand natural language. Natural language processing also has been a goal of artificial intelligence researchers. Although some aspects of language, such as syntax, can be processed with relative ease by computer programs, several other levels of language processing, notably semantics, are difficult to handle because they are ambiguous and depend on context. Restricted subsets of natural language within a limited domain of discourse can be processed, however.

Other perceptive faculties that AI researchers try to imitate are vision and speech understanding. Computer-vision studies deal with the processing of a two-dimensional image in such a way that a computer can recognize objects and infer their relative positions and various attributes, such as shape and colour. These types of AI programs construct an internal description of the image. This symbolic representation is further processed by the program through the use of its knowledge base to recognize the actual scene.

Perception of connected human speech requires several knowledge bases, focused on the characteristics of speech sounds (acoustics and phonetics), the variability of pronunciation (phonology), and the meaning of the utterances (semantics), which depends on context (pragmatics). The accurate recognition of the utterances of arbitrary speakers is limited, but systems have been "trained" to interpret the discrete (not connected) speech of a single user; their vocabularies can contain as many as a thousand words. Research in speech perception has led to the development of an important general approach to handling complex problems. Called the blackboard model, this method relies on several knowledge processes that run simultaneously and place their hypotheses into a common memory area (the "blackboard"), where they are either validated or rejected by other processes.

The blackboard model

Among the more prominent results of research in artificial intelligence are the expert systems that approximate the performance of human experts, robots that combine vision with other sensors and are able to generate an adaptable plan of action, machine-vision systems, and natural-language interfaces to various computer systems. The domain of operation (or understanding, as some refer to it) of all these systems remains restricted.

Artificial intelligence is richly articulated with other areas of computer science. Research on knowledge bases complements that in data bases: the first improves capabilities for storing semantic information, and the second offers better techniques for storing and retrieving large volumes of facts. Many problems (such as pattern recognition), once solved by the artificial intelligence community, have entered the knowledge base of other computer science fields. Artificial intelligence also has contributed significantly to the field of programming languages and methodology.

## ROLE OF COMPUTER SCIENCE

Since the computer has become an ever-present force in human affairs, computer science has played an increasingly important role in the way people work, live their private lives, organize their activities in society, and, to an extent, perceive themselves. The ubiquitous microprocessor and the evolving combination of computers, telecommunications, and large data bases are changing the complexion of life in industrialized societies.

Computer science also has influenced the entire family of scientific disciplines. Work in the field has led to the development of tools that have the potential of expanding the methodology of science.

**Effects of computers on society.** The computer, particularly when combined with communications technology, has made it possible to handle the vast amount of information that is needed to increase the efficiency of production, management, and government. What is sometimes called the "information revolution" is thus bringing to fruition the advantages of the mechanization that was fostered by the Industrial Revolution, which began in the late 1700s.

The so-called information revolution

Industry is being transformed because of the recognition that, from design to production, manufacturing processes are information-intensive activities. Making information flows explicit and organizing computer networks to handle them have increased productivity in offices as well as factories. Computer-aided design and engineering are integrated with the control of production, which relies on flexible automation (such as that provided by programmable robots) to offer production an adaptability that could never be attained with traditional assembly lines. Machine-vision equipment conducts automatic inspections, and expert systems aid in troubleshooting.

Management information systems are used to integrate organizational record-keeping and decision-making, as well as to process business transactions, and thus help plan and run the overall activities of an enterprise. Automation of office work has become extensive. The move to electronic

transactions, processed by computers and transmitted by electronic data interchange systems, has led to a progressive integration of the organizations in the value-added stream of the economy, from suppliers to customers.

Already extensively used at various levels of government and in grass-roots political activity, the computer offers the possibility of local autonomy, combined with the integrative powers of communications systems. The aggregate of computers and communications can have paradoxical consequences, however: it can be used to further participatory democracy or to strengthen centralized control of a society; it can help promote prosperity in peace or help wage a war more effectively. Although the widespread computerization of society creates its own demands for increased openness, these demands are not absolute. The only general statement that can be made in this regard is that the penetration of computer and communication technologies poses challenges—in government, in business, and in education—to which society must respond. Privacy is just one example of such a fundamental issue.

Real-time computer processing capabilities are critical in many important areas, including space exploration, weather forecasting, noninvasive techniques of medical diagnosis, modern financial markets, and battle-control systems. The benefits of universal connectivity, which can be provided by global computer-communications systems based on the digital transmission of data, voice, and visual imagery, are yet to come.

**Computer science in the family of disciplines.** Computer science is intimately related to many disciplines of knowledge. Although it is mostly empirical, its basic tools are mathematics and logic. Computer science furnishes other sciences and technologies with intellectual tools in addition to hardware and software—namely, techniques of process-oriented thinking. Modern particle physics owes many of its recent achievements to powerful computers that control experiments and analyze data. Molecular biologists use computerized machines for establishing the genetic structure of living organisms, knowledge that can be used to develop new drugs and improve breeds of livestock and strains of plants. Ecologists rely on computer models of the human body to estimate its response to chemical substances.

The manipulation of computer models furnishes a new mode of scientific understanding and blurs the distinction between theory and experiment. For example, assisted by the visualization of equations through techniques of computer graphics, mathematicians gain new insights into complex relationships. Even more striking methodologically is the emerging use of computer programs as parts of mathematical proofs. Thus, the conjecture that any map on a flat surface or a sphere can be coloured without using more than four colours was proved to be true—after more than a century of barren attempts at a proof—by using computer calculations of such volume that they would have required many human lifetimes.

Computers provide new tools for conducting scientific enterprises and have proved to be a true intellectual technology. Geographically distant scholars, working on common problems, now can be united on a computer network into an "invisible college." From information bases offering access to the exploding body of scientific knowledge, individual investigators can forge their own bases of information by means of the so-called hypertext system, which enables them to link related aspects of their respective studies.

Not the least of the contributions of computer science, artificial intelligence has promoted the emergence of cognitive science that deals with the problems of knowledge, understanding, inference, and learning.

**BIBLIOGRAPHY.** A general overview of the field is found in VLADIMIR ZWASS, *Introduction to Computer Science* (1981). The research agenda of computer science is presented in BRUCE W. ARDEN (ed.), *What Can Be Automated?: The Computer Science and Engineering Research Study (COSERS)* (1980). ANTHONY RALSTON and EDWIN D. REILLY, JR. (eds.), *Encyclopedia of Computer Science and Engineering,* 2nd ed. (1983), is a comprehensive reference work.

The emergence of the computer is traced in HERMAN H. GOLDSTINE, *The Computer from Pascal to Von Neumann* (1972, reprinted 1980); and the establishment of the field of artificial intelligence is covered in PAMELA MCCORDUCK, *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence* (1979).

Comprehensive information on the analysis and design of algorithms is presented in DONALD E. KNUTH, *The Art of Computer Programming,* 3 vol. (1968–73), the first two volumes of which, *Fundamental Algorithms* and *Semi-Numerical Algorithms,* came out in a second edition in 1973 and 1981, respectively. The theory of computation is explored in E.V. KRISHNAMURTHY, *Introductory Theory of Computer Science* (1983), a brief but accessible and inclusive text.

A layered view of the field of computer architecture is offered in ANDREW S. TANENBAUM, *Structured Computer Organization,* 2nd ed. (1984). Specific aspects of computer architecture are covered in ANDREW S. TANENBAUM, *Computer Networks* (1981); and JAMES D. FOLEY and ANDRIES VAN DAM, *Fundamentals of Interactive Computer Graphics* (1981, reprinted 1984).

ELLIS HOROWITZ, *Fundamentals of Programming Languages,* 2nd ed. (1984), studies the concepts of programming languages; and PETER CALINGAERT, *Assemblers, Compilers, and Program Translation* (1979), discusses automatic translation from these languages. The issues of programming methodology are reviewed in PETER WEGNER (ed.), *Research Directions in Software Technology* (1979). JAMES L. PETERSON and ABRAHAM SELBERSCHATZ, *Operating System Concepts,* 2nd ed. (1985); and C.J. DATE, *An Introduction to Database Systems,* vol. 1, 4th ed. (1986), introduce two distinct areas of software systems. IRVING ALLEN DODES, *Numerical Analysis for Computer Science* (1978), is a survey of numerical computation. EUGENE CHARNIAK and DREW MCDERMOTT, *Introduction to Artificial Intelligence* (1985), is a comprehensive source.

The role of the computer in society is studied in TOM FORESTER (ed.), *The Information Technology Revolution* (1984); and the place of computer science in the family of academic disciplines is analyzed in FRITZ MACHLUP and UNA MANSFIELD (eds.), *The Study of Information: Interdisciplinary Messages* (1983).

(V.Zw.)

# Computers

A computer is a device that solves problems by applying prescribed operations on data entered into it. The electronic computer can perform its data-processing operations accurately at high speeds without human intervention. There are two basic types of computers—analog and digital. The analog computer operates on data represented by variable physical quantities, such as voltages, and measures continuously. By contrast, the digital computer works with numbers, words, and symbols expressed as digits, which it manipulates and counts discretely. A third general class of computers, the hybrid computer, combines the features of the other two and utilizes both analog and discrete representations of data.

The majority of computers in use today are of the digital variety. Extremely versatile, digital computers can carry out a multitude of varied tasks, from routine accounting and bookkeeping to the control of spacecraft and analysis of scientific data. Because of its widespread use and the tremendous impact the digital computer has had on modern society, this article treats it in considerable detail, though attention also is given to the other varieties.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 735.

The article is divided into the following sections:


Types of computers 638
  Analog computers 638
    Early analog computers
    Basic operation and applications
  Digital computers 639
  Hybrid computers 639
History of digital computing 640
  Early developments 640
    Origins of computing instruments
    Key inventions and discoveries
  Modern digital systems 641
    First generation
    Second generation
    Third generation
    Fourth and future generations
Digital computer operation and components 643
  Basic functions and operational units 643
    Input
    Storage
    Control
    Processing
    Output
  Methods of operation 644
Digital computer programming 645
  Programming fundamentals 645
  Language categories 646
    Algorithmic and procedural languages
    Interactive languages
    Nonnumerical languages
    Simulation languages
    Process- and numerical-control languages
Applications of digital computers 649
  Control 649
    Process control
    Numerical control
    Transportation control applications
    Other control applications
  Communication 650
  Simulation 650
  Design 651
  Artificial intelligence 651
    Decision making
    Natural-language processing
    Computer vision
    Robotics
Bibliography 652


## Types of computers

As previously noted, modern computers are of three general types: analog, digital, and hybrid. They differ in operating principles, general design, and applications.

### ANALOG COMPUTERS

An analog computer solves problems by operating on continuous variables rather than on discontinuous or discrete units as do digital computers. Physical quantities, such as the position of a lever, voltage on a wire, or pressure in a pipe, stand for the values being computed, and the measurements of these quantities represent the values of the variables. An automobile speedometer can be regarded as a type of analog computer measuring the voltage output of a generator connected to the drive shaft. As the speed of rotation of the shaft changes, so the generator output changes—*i.e.,* the indicated voltage representing the generator output is used as an analogue of the road speed of the car.

To represent physical variables by means of analogues, an analog computer uses various types of amplifiers connected so as to perform such arithmetic operations as summation and multiplication or such mathematical operations as integrations and differentiations. Accuracy is somewhat limited by the precision available in some of the electrical components; it may vary slightly as a result of environmental or material conditions. Digital computers do not have this problem because they operate on the presence or absence of discrete signals.

imita-
ion to
ccuracy

Analog computers operate on data continuously, working in real time, as they solve problems. They constitute a physical system in which the analysis or solution of a problem is mirrored by the varying values or behaviour. A simple example of analog control is the industrial thermostat, or temperature controller. The measured temperature is converted into an electrical current that varies in magnitude as the temperature rises and falls. In analog computers the electrical circuitry of a controller is analogous to the problem to be solved, such as the maintenance of the measured temperature at some present value. Circuitry notes the measured value of the temperature and computes the amount for resetting of the final control element, which may be a valve, damper, or switch that is needed to return the temperature to the desired value. The corrective signal is then sent to the final control element.

**Early analog computers.** Among the early analog computers was a special-purpose computer developed by Lord Kelvin in 1872 as a tide predictor. Later devised along the same lines was a harmonic synthesizer with 80 components, each of which could generate a sinusoidal (wavelike) motion. The periods of the various units were simply related by gear trains; each component motion could be multiplied by constant factors by adjustment of the fulcrum on levers, and the components were added to produce a resultant. A pen attached to the sum lever plotted the output. From harmonic analyzers a machine called the isograph was developed in 1937 to solve polynomial equations with real coefficients. An electrical version of the equation solver was built in the same year, and both mechanical and electrical machines have since been constructed to solve systems of linear equations.

A significant development in analog computers was the invention of a device for amplifying torque by the U.S. electrical engineer Vannevar Bush. In 1931 Bush and his colleagues at the Massachusetts Institute of Technology

Differential analyzer

had fabricated a mechanical analog computer called the differential analyzer, which was to influence developments in various laboratories around the world. In a device of this kind, shaft motion represents variables, gears give multiplication or division, and differential gears provide addition and subtraction. Integration is accomplished by means of a knife-edged wheel rotating at a variable radius on a circular rotating table. A mechanical interconnection of such devices is the analogue of a system represented by a set of differential equations.

**Basic operation and applications.** Most analog computers operate by manipulating electrical potentials or, more usually, potential differences. The basic unit of an analog computer is the operational amplifier, a device whose output current is proportional to its input potential difference. By causing this output current to flow through suitable components, further potential differences are obtained and can be made to bear any one of a wide range of relationships to the input. Until the late 1950s all electrical analog computers used vacuum tubes; later designs, using solid-state devices, have been much smaller, and they consume less power and are less expensive.

Operational amplifier

The original electronic analog computers arose from the needs of antiaircraft artillery predictors. In these, the variables were latitude, longitude, and height of the target and projectile, all of which varied with time. Also input were fixed or slowly varying data, such as muzzle velocity and ballistic characteristics of the projectile and characteristics of the atmosphere; the computer had to solve two simultaneous equations so that the target and the projectile, each moving along its own course, would arrive at the point of intersection at the same time.

One interesting application of analog computers is in aerodynamics. Equations relating airflow and the attitude of and forces on lifting surfaces are complex, and their solutions by conventional methods were so difficult and tedious that in the past many designers preferred to ignore theory and rely on experiment. Analog computers became available at a time when the experimental approach was becoming impossibly expensive and dangerous; difficult equations involving such problems as flutter (a peculiar type of vibration) are now within their capability.

Applications of network analyzers

Network analyzers, devices or programs for analyzing special networks, are the bases for construction of special-purpose analog computers. Some areas of application of network analyzers include: (1) psychology–neurology, in which neuron networks based on simple models of brain behaviour are simulated; (2) management and control scheduling, in which elapsed time and costs of projects involving many special operations can be controlled; (3) electronics network analysis, in which the performance of passive and active component circuits, specifically frequency response and pulse response, for example, are used to compute the performance of long-distance telephone circuits under a wide variety of operating conditions; (4) hydrodynamics–hydraulics network analysis of the flow of liquids through networks of pipes, such as in a chemical plant or in sewer or water systems; and (5) electric power distribution network analogues in such areas as power transmission line distribution systems, power grids, and direct current circuits.

All of the above-mentioned analyzers have special conditions relating to responses to peak loads, overloads, transients, various types of fault defects (faults), and other related phenomena. An analysis of such networks calls for checkouts with which to determine methods of minimizing losses, rerouting, regulating power distribution, and fault elimination. Some networks are used to simulate mechanical or electronic systems for system behaviour studies.

Engineers can simulate individual processes on a general-purpose analog computer and thus develop data from which they can design a special-purpose machine that will handle exact requirements for a specific application. A custom-tailored single-purpose computer does one specific job rapidly and accurately and as many times as desired. Its single program is built-in, so that it is possible for plant personnel to learn quickly how to operate it.

It is a limitation intrinsic to analog computers that the accuracy of any calculation is restricted by the precision or fidelity with which signals can be combined and amplified. Even with high-grade components and a careful formulation of a problem, the result may be in error by a few parts per thousand. For high-accuracy work digital methods must therefore be used.

## DIGITAL COMPUTERS

A digital computer differs in principle from an analog computer rather in the way that counting differs from measuring. An analog signal has a continuously valid interpretation from the smallest to the largest value that it can achieve. A digital signal has only a discrete number of interpretations, usually (though not necessarily) two. Each signal thus falls in a number of bands within which its interpretation is clear, separated by regions of transition where the interpretation would be ambiguous. Signals being processed are arranged never to be in these ambiguous states, and in this way calculations can be made to deliver exact results. Suppose that a variable has any of $N$ different values. Inside the computer a pattern of signals chosen from a repertoire of $N$ possible patterns can represent the value of that variable exactly. There are many ways of doing this. The most commonly adopted is a binary positional notation in which each signal is either 0 or 1 and a sequence of $M$ such signals is interpreted as representing a value $V$ in the following manner:

Binary positional notation

$$V = 2^{M-1}d_{M-1} + 2^{M-2}d_{M-2} + \cdots 2^2 d_2 + 2^1 d_1 + d_0$$

where $d_i$ is the $i$th signal. Clearly there are $2^M$ different patterns and therefore $N = 2^M$ different values for $V$. Patterns can be chosen to code a range of integer values, the characters in an alphabet, the instruction set of the computer, and so on. The $M$ signals that it takes to represent each value may be carried in turn over a single wire (serial operation) or all at once over $M$ separate wires (parallel operation). Thus a parallel computer will have an $M$-fold replication of some of its circuits but may be expected to work faster than a serial machine. In most systems there will be examples of parts that work in serial and other parts that work in parallel, the intention being to retain a reasonable balance between speed, complexity, and cost.

Bits and bytes

In a binary computer each signal is called a bit (from *bi*nary digi*t*), a pattern of eight bits is known as a byte, and a number (varying from one machine to the next but typically two or four) of bytes is called a word. Many machine operations are on word-sized units of data. An instruction might specify, for example, that a certain word from the memory is to be transferred to a register within the arithmetic unit. Another might add the contents of two registers together and send the resulting word to a third register. In operation, a sequence of such instructions (a program) is held in memory and the instructions are performed in turn by the computer. A particular set of instructions that allows the computer to choose which of the instructions to execute next according to the results of some text is fundamental. Such a test permits the machine to alter its action at a stage in the program that may have been difficult, if not impossible, for the programmer to anticipate. In this manner an extraordinary diversity of behaviour can be achieved by computers solely by virtue of their speed and precision.

All digital systems, regardless of size, type, or application, have certain basic components and follow certain fundamental operational principles. For a discussion of these, see below *Digital computer operation and components.* See also *Applications of digital computers* for coverage of the common uses of these systems.

## HYBRID COMPUTERS

A hybrid computer system consists of a combination of analog and digital computers. In most of the early hybrid systems (introduced during the late 1950s), the digital machine was employed merely as a support device for the analog unit. Most recent hybrid computers, by contrast, are digitally based. The basic components of these systems include a digital processor and a main memory for the internal storage of a master digital program and data, a feature that allows for automated program setup and operation. Several analog units are used to provide con-

Data conversion

tinuous parallel computational capability; they operate directly under the control of the digital program. For proper interfacing, hybrid systems are equipped with special converters that translate data from the analog processors into the digits of the binary code and with devices that convert digitally processed information into analog representation.

Hybrid systems offer more control capability than analog machines do. Since the mid-1960s they have been widely used in simulation studies of nuclear power plant facilities, guided missiles, fighter aircraft, and spacecraft, where it is crucial to have a close representation of a dynamic system. Hybrid computers played a major role in the designing and evaluation of the U.S. Space Shuttle, the first reusable manned spacecraft. Real-time simulations of the Shuttle's rocket engine and hydraulic system were done with hybrid computers, as were those for determining crew responses.

(C.J.S./A.M.P./N.E.W./S.F.We.)

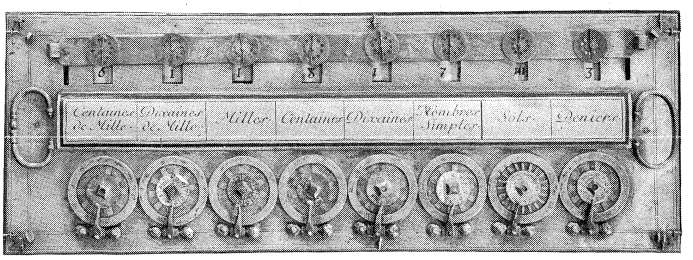
## History of digital computing

### EARLY DEVELOPMENTS

**Origins of computing instruments.** The history of computing instruments may be considered to begin with the abacus, which originated in the Orient more than 5,000 years ago. During the Middle Ages the abacus was used throughout the European and Arab worlds as well as in Asia. It is still employed in Japan and in parts of the Middle East. The abacus is so designed that calculations can be performed by sliding counters—usually beads or blocks—along parallel rods (in some cases, wires or grooves). The counters are divided into two sections by a bar perpendicular to the rods. One section has one or two counters, representing 0 and 5 depending on their position along the rod. The second section has four or five counters, representing units. Each bar represents a significant digit, with the least significant digit on the right.

The first adding machine

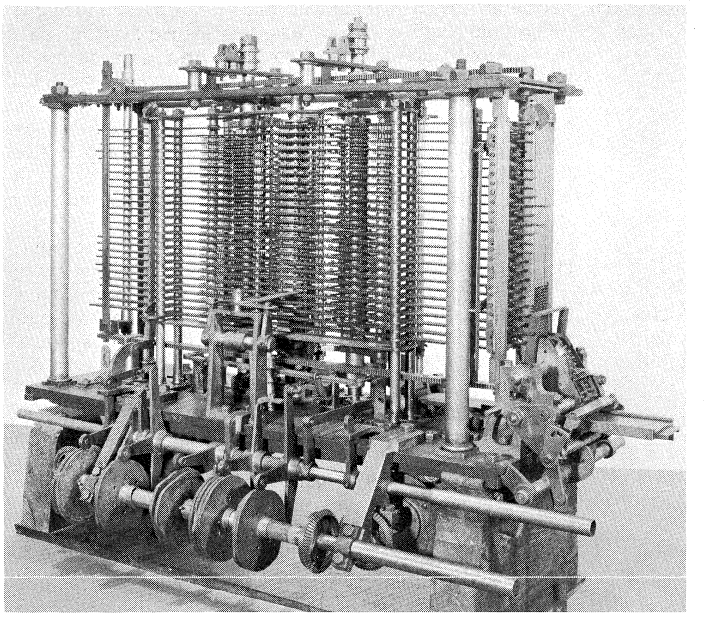
Mechanical calculating machines were invented in Europe during the 17th century. The first such device was an adding machine built in 1642 by the French scientist and philosopher Blaise Pascal to aid him in computations for his father's business accounts. This device consisted of a train of number wheels whose positions could be observed through windows in the cover of a box that enclosed the mechanism. Numbers were entered by means of dial wheels. Pascal's machine was a digital device, and it carried out its computations by a process of integer counting. It utilized a mechanical gear system to add and subtract, with as many as eight columns of digits.




Mechanical adding machine built by Blaise Pascal in 1642.

The German mathematician Gottfried Wilhelm Leibniz saw Pascal's machine in Paris and developed a more advanced version. Whereas Pascal's device was able only to count, Leibniz' so-called Stepped Reckoner could also multiply, divide, and extract square roots. A working model of the machine was completed in 1673 and exhibited at the Royal Society in London. The first commercially available calculator, Thomas de Colmar's arithmometer, was produced in 1820, but, like Leibniz' invention, it proved to be unreliable.

**Key inventions and discoveries.** The first automatic digital computer was conceived in the 1830s by the English inventor Charles Babbage. Called the Analytical Engine, this mechanical device was designed to combine arithmetic processes with decisions based on its own computations. Babbage's plans embodied most of the basic elements of the modern digital computer, as for example a punched-card input/output medium, arithmetic unit, memory in which to store numbers, and sequential control. The Analytical Engine was never completed, largely because precision techniques for fabricating metal parts to close tolerances had not yet been developed. Babbage's invention was forgotten until his writings were rediscovered in 1937.



Part of the Analytical Engine designed by Charles Babbage in the 1830s.

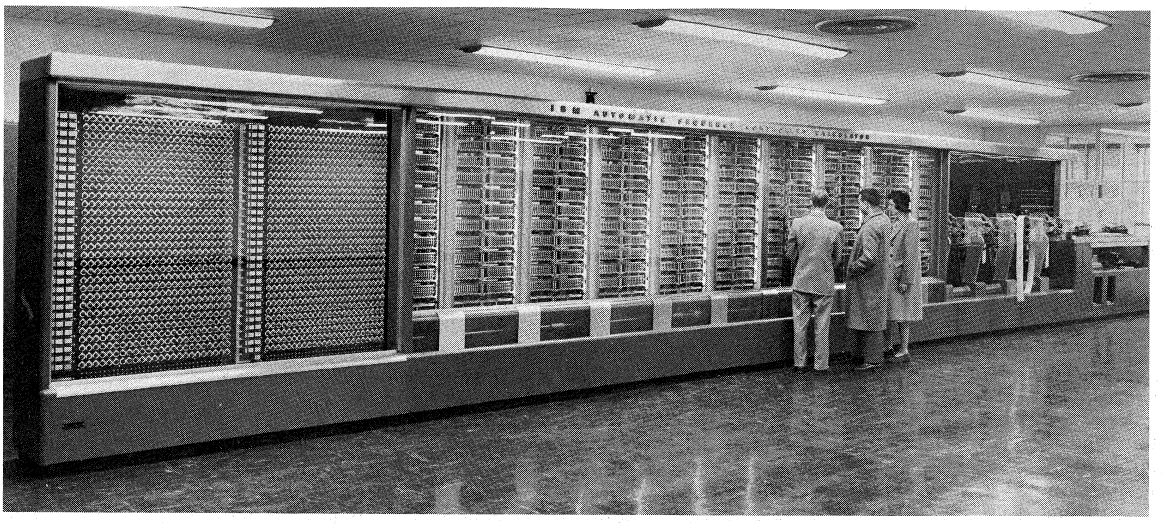
Reproduced by permission of the trustees of the Science Museum, London

Of greater significance in the evolution of the digital computer was the work of the English mathematician and logician George Boole. In various essays written during the mid-19th century, Boole called attention to the analogy between the symbols of algebra and those of logic as used to represent logical forms and syllogisms. Boole's system, with its binary logic operators (*e.g.,* AND, OR, and NOT), became the basis of what is now known as Boolean algebra, on which electronic computer switching theory and procedures are based.

Introduction of punched cards in data processing

Another major step in computer development was the introduction of punched cards in data processing by Herman Hollerith, a U.S. statistician. While working on the 1880 census returns, Hollerith recognized the possibility that a pattern of holes in perforated cards could be read electrically by a machine specially designed to sort and manipulate the numerical data represented by the holes. He may well have been inspired by the work of Joseph-Marie Jacquard of France. In the early 1800s Jacquard had devised a method of controlling the operation of a loom through the use of punched cards. By the U.S. census of 1890, Hollerith had invented a tabulating system that automated census counts. The electromechanical sensing and punching devices integral to Hollerith's system were forerunners of modern peripheral computer equipment.

In 1939 John V. Atanasoff, a U.S. mathematician and physicist, built a breadboard prototype of an electromechanical digital computer, a feature of which was the first known use of electronic vacuum tubes for computation. That same year Howard Aiken of Harvard University, in association with engineers of International Business Machines Corporation, began work on a fully automatic large-scale calculator using standard business machine components. In 1944 they completed the Automatic Sequence Controlled Calculator, commonly known as the Harvard Mark I. This was an enormous electromechanical machine, approximately 50 feet (15 metres) long and eight feet (2.4 metres) high, whose operations were controlled by a sequence of instructions coded on punched paper tape. Data were entered on punched cards, and the output was recorded either on such cards or by an electric typewriter. The input data were processed by means of electric relays (switching devices).



Automatic Sequence Controlled Calculator, also known as Harvard Mark I, completed in 1944.
By courtesy of the IBM Corporation

## MODERN DIGITAL SYSTEMS

Since the development of the Mark I, the digital computer has evolved at an extremely rapid pace. The succession of advances in computer hardware, most notably in logic circuitry and storage systems, is generally discussed in terms of the concept of generation. The machines in each generation of computers share a common technology.

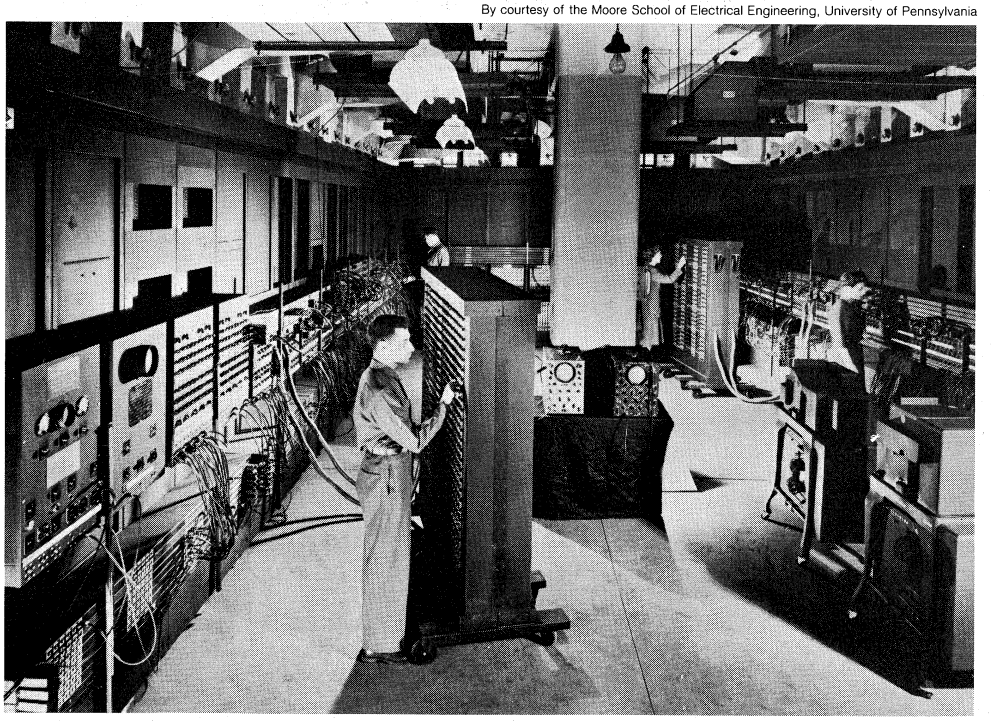
**First generation.** J. Presper Eckert and John W. Mauchly, both of the University of Pennsylvania, ushered in the first generation of modern computers with the ENIAC (Electronic Numerical Integrator and Calculator). Completed by 1946, ENIAC was the first all-purpose, all-electronic digital computer. A special-purpose, all-electronic computing machine called the Colossus had been developed earlier at Bletchley Park, north of London, and was in operation by December 1943. The Colossus was designed to decipher codes generated by the German electromechanical devices known as the Enigma and the Geheimschreiber ("secret writer"). Several different models were used before the end of World War II. Both the ENIAC and the Colossus series of machines used vacuum tubes instead of relays as their active logic elements. This resulted in a substantial increase in computational speed. The ENIAC was more than 1,000 times faster than its electromechanical predecessors and could execute up to 5,000 basic arithmetic operations per second.

The first all-electronic digital computer

The ENIAC was an externally programmed machine that utilized plugboards on which the sequence of operations required to solve any particular problem could be set up. Its successor at the University of Pennsylvania was EDVAC (Electronic Discrete Variable Automatic Computer), a machine of radically different design that was influenced by ideas germinating at the Institute for Advanced Study at Princeton University. The EDVAC was a stored-program computer in which instructions were treated in the same way as numerical data and stored electronically in the computer's memory. The stored-program concept led to the development of self-modifying computer programs since instructions could be manipulated just as readily as data by the arithmetic facilities of the ma-

The stored-program concept

By courtesy of the Moore School of Electrical Engineering, University of Pennsylvania



ENIAC (Electronic Numerical Integrator and Calculator) built by J. Presper Eckert and John W. Mauchly of the University of Pennsylvania, completed in the 1940s.

chine. The EDVAC was not completed until 1950, but a computer of rather similar design, the EDSAC (Electronic Delay Storage Automatic Calculator) at Cambridge University in England, began operating in early 1949 and may have been the first working version of a stored-program computer. Also, during the mid-1940s the Hungarian-born mathematician John von Neumann, together with Arthur W. Burks and Herman H. Goldstine, had produced a report describing the stored-program idea and its embodiment in the design for a computer at the Institute for Advanced Study at Princeton, the IAS computer. Several early computers, such as ORDVAC (Ordinance Variable Automatic Computer), Whirlwind I, and the IBM 701, were designed along the lines proposed in the report.

Other notable first-generation electronic digital computers included the UNIVAC I (Universal Automatic Computer), built in 1951 by Eckert and Mauchly for the U.S. Bureau of the Census. As with all computers of this generation, the UNIVAC I utilized vacuum tubes—primarily triodes and pentodes. Its main memory consisted of mercury delay lines. In each such delay line a tube of mercury served as an acoustic medium through which sonic pulses representing the binary digits 0 and 1 were transmitted over a specified distance between a pair of transducers. This memory system permitted an access time of approximately 500 microseconds. The UNIVAC I was the first digital computer to handle both numerical and alphabetical information with equal ease and utilized the principle of the separation of input/output from computation per se.

Although the logic technology of later first-generation computers remained much the same as that of UNIVAC I, they featured improved main memories. The Harvard Mark III, for example, was equipped with a magnetic drum memory, which provided relatively large storage capacity but slow access of approximately 17 milliseconds. By the mid-1950s reliable magnetic core memories had been developed and incorporated into machines such as UNIVAC II.

**Second generation.** The second computer generation began in 1959, when machines employing semiconductor devices known as transistors became commercially available. The transistor had been invented in 1947, but more than 10 years of developmental work was needed to render it a viable alternative to the vacuum tube. The small size of the transistor, its greater reliability, and its comparatively low power consumption made it far superior to the latter. By using transistors in control, arithmetic, and logic circuits, along with an improved magnetic core memory, computer manufacturers were able to produce more efficient, smaller, and faster (up to 100,000 instructions per second) digital systems.

**Third generation.** During the late 1960s and 1970s electronic components were further miniaturized, resulting in dramatic advances in computer hardware. The first significant breakthrough was the fabrication of the integrated circuit (IC), a solid-state device consisting of hundreds of transistors, diodes, and resistors on a tiny silicon chip.

The impact of integrated circuitry

The use of ICs permitted the construction of mainframe (large-scale) computers of higher operating speeds, capacity, and reliability at lower cost. Their implementation also enabled engineers to design minicomputers, machines compact enough to place on a desk yet powerful enough to control the operations of a small factory or monitor the instruments of a scientific laboratory.

The next major development was large-scale integration (LSI), which made it possible to pack thousands of transistors and related devices on a single integrated circuit. Such microcircuitry yielded two devices that revolutionized computer technology. The first of these devices was the microprocessor. Often referred to as "a computer on a chip," the microprocessor is an integrated circuit that contains all the arithmetic, logic, and control circuitry necessary to serve as a central processing unit (CPU)—*i.e.,* the part of the digital computer that interprets and executes instructions. Its development led to the production of microcomputers (systems no larger than portable television sets) and "intelligent" remote terminals.

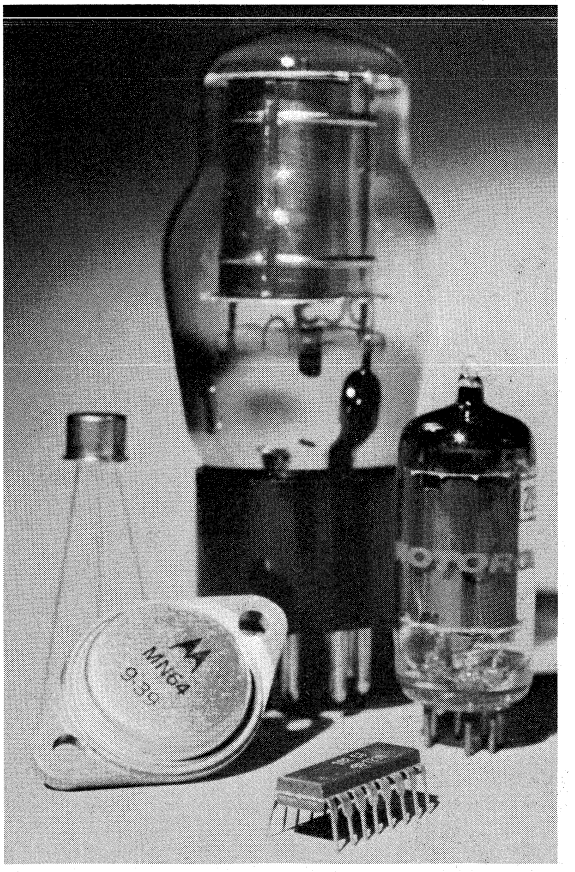
As the number of circuits that could be fabricated on a single chip increased, the need to adopt a different technology for logic and memory diminished, with the effect that semiconductor memories gradually replaced magnetic core memories in new computer systems. The random-access memory (RAM) chip used in constructing semiconductor memory units is the second important device that emerged from LSI technology.

**Fourth and future generations.** The computers of the 1980s have often been referred to as fourth generation, though the differences between them and third-generation systems are not always clear. Most of their distinguishing features are attributable to very large-scale integration (VLSI). This technology has vastly increased the circuit density of microprocessor, memory, and support chips (*i.e.,* those that serve to interface microprocessors with input/output devices). Whereas large-scale integrated circuits contain thousands of components on a silicon chip less than 0.2 inch (five millimetres) square, the very large-scale integrated circuit holds hundreds of thousands of parts within the same amount of space. The reduction in cost associated with these developments has been the main factor in making possible the production of simple computers for use in homes and schools.

The proliferation of reliable and inexpensive computer systems and concomitant advances in programming have resulted in the growth of computer networks. These networks are collections of individual computers and intelligent peripheral equipment (*e.g.,* automatic teller machines and point-of-sale terminals equipped with microprocessors) that are interconnected by telephone lines, microwave relays, and other high-speed communication links commonly established for the purpose of sharing and exchanging resources or data. Networking has been developed on all levels, from local to international, in diverse sectors of society. For example, government organizations, including law enforcement agencies, avail themselves of network facilities for rapid retrieval of information from a central

Growth of computer networks

By courtesy of Motorola, Inc.



The vacuum tube (centre and right) was replaced by the transistor (left). Further miniaturization made it possible to fit thousands of transistors on a single integrated circuit (foreground).

data bank. Commercial banks and retail merchants make extensive use of computer networks to facilitate the transfer of funds or credit verification. Computer-to-computer communication is now possible even for home computer owners; subscribers to special information networks are able to interact with a host computer to access various forms of useful data such as the latest stock-market reports, airline flight schedules, and entertainment listings.

Since the early 1980s a number of computer scientists and engineers, principally in Japan and the United States, have intensified efforts to develop computer systems endowed with what is frequently called artificial intelligence (AI). Such machines, the so-called fifth generation, would have the capability to reason, to recognize relationships, and to learn. In effect, they would be able to think like humans and improve their performance on the basis of experience. They would not have to be instructed step by step as to how to carry out specific tasks, as required by current computers.

Artificial intelligence

It is clear that VLSI technology can provide the basis of the hardware needed to build the next generation of computers. These machines will have an internal structure that allows parallel processing—namely, the simultaneous execution of several separate operations (*e.g.,* memory, logic, and control) by means of numerous integrated circuits in which millions of CPU, memory, and input/output circuits are working together. Virtually all present-day conventional digital computers perform their operations serially or sequentially: discrete input circuits feed data to individual memory cells, each of which in turn relays bits of information one at a time to a single CPU for processing, with the results then being transmitted to an external output device. Even though today's fastest computers—the so-called supercomputers—can carry out about 50,000,000 instructions per second, they work too slowly to approximate higher forms of human thought that involve making associations and generalizations very rapidly.

Major improvements in software such as natural-language programs also are helping to lay the groundwork for intelligent computers. Programs of this sort enable computers to understand written commands in ordinary English (see below *Applications of digital computers*). Already available too are single-chip signal processors that can be programmed to recognize simple, isolated spoken words. Computer manufacturers expect to produce voice-input devices capable of handling connected speech of larger vocabularies (including words with multiple meanings) as soon as more sophisticated speech-processing software is perfected.

Speech recognition

Other developments that may pave the way for computers with artificial intelligence are knowledge-based systems, which entail the storage and application of enormous amounts of specialized knowledge to particular complex, nonnumerical problems. This type of program has been used to help physicians diagnose certain forms of illness and to determine suitable treatments. Perhaps more important in AI research are attempts to develop computer programs that can prepare other programs. Such sophisticated software would transform the craft of programming and, in some cases, would eliminate it altogether.

## Digital computer operation and components

### BASIC FUNCTIONS AND OPERATIONAL UNITS

The basic functions of every digital computer are input, storage, control, processing, and output. The accompanying diagram illustrates these functions in simplified form. The digital computer receives data as patterns of digital signals and stores them in any of various main or auxiliary memory devices until needed. Also held in storage are the instructions that dictate the operations of the computer. The control function involves retrieving instructions from storage in appropriate sequence and relaying the proper commands to the processor, the arithmetic/logic unit (ALU). It also includes interfacing or synchronizing the operating speeds of the various input and output units to that of the processor. The processor functions somewhat like a calculator, adding, subtracting, multiplying, dividing, and listing. It carries out these operations, however, at extremely high speeds. In addition, the processor is able to make decisions—*i.e.,* select and compare on the basis of stored information. The output of the digital computer, like its input, takes many forms. Generally it is printed, put on cards or tape, stored in secondary memory units, displayed on the screen of a cathode-ray tube, or communicated to other remote devices.

A digital computer system is physically a collection of electromechanical and electronic components and devices assembled in cases (modules) and cabinets. The main memory, control elements, and ALU together make up the central processing unit, the heart of the system. The different input/output (I/O) devices and auxiliary storage units, some of which double as I/O devices, constitute peripheral equipment. Apparatus of this kind with direct connection to the CPU is said to be on-line; peripheral equipment working independently of it is termed off-line.

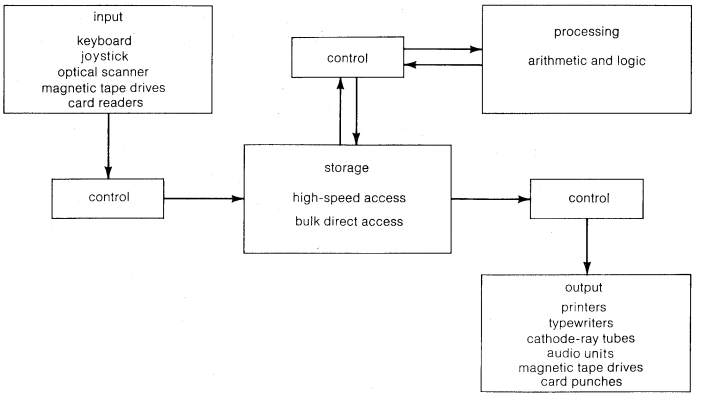
Peripheral equipment

**Input.** Input to the computer can be in a variety of forms—punched cards, punched paper tape, magnetic cards, or magnetic ink (bank checks); or the input may be made directly and in binary form from the main memory or peripheral storage units and various conversion units. By the early 1970s typed print and even handwritten script in limited vocabularies became serviceable as input to computers. Optical scanning was perfected several years later. Optical laser scanners such as those used with point-of-sale terminals in retail stores can read either bar-coded data or optical character fonts.

Input may also be taken from stylus-tablet devices that allow a form of freehand pointing and sketching action, or from movement sensors such as a joystick or a "mouse." Such devices permit a kind of graphic interaction in which a picture displayed by the computer on a television-type screen is modified by the program in response to the actions taken by the operator with the input device. This sort of interaction is an important aid for engineering and other applications. Voice input is being developed with wider vocabulary (see above). Furthermore, a device called the modem (modulator/demodulator) enables any computer or simple data terminal to communicate with other computers over a voice-grade telephone circuit.

**Storage.** A device for storing information in binary form must be designed so that it can be placed in one of two stable states corresponding to 1 or 0, and it has to be able to maintain this state indefinitely until changed. It also must provide access to the stored data for reading or modification at very high speeds (about 1,000,000 times a second). Finally, since hundreds of millions or more bits of data may have to be stored, the apparatus must be inexpensive. As all these requirements cannot be met satisfactorily by any single type of device at present, most computers use more than one kind of storage unit.

*Main memory.* Every digital computer has a main memory that holds data and instructions for immediate use by the central processing unit. The main memory receives this information from an input device or an auxiliary storage apparatus. In addition, it accepts and temporarily stores data from the control and arithmetic/logic units of the CPU. This primary storage device is generally a high-speed random-access memory, or RAM, whose ac-




Basic computer functions.

Semiconductor memory

cess time is independent of the sequence in which bits of information were stored. The prevailing technology uses a derivative of the metal-oxide-semiconductor transistor to furnish solid-state devices containing up to 256,000 bits of memory on each single chip.

*Auxiliary storage units.* These secondary storage devices constitute a basic part of a digital system's peripheral equipment. They can store much larger amounts of information than can a main memory but operate at slower speeds. The most common forms of secondary storage are magnetic tape and disk. Other systems include the magnetic bubble memory and the optical disk memory.

Magnetic tape storage

Magnetic tape was introduced as a data-storage medium in 1951, when it was employed as an auxiliary memory for UNIVAC I. For about the next decade almost all digital computers used magnetic tape for secondary storage, but by the mid-1960s magnetic disk and drum devices arrived to augment the tape units in mainframe systems designed for scientific and business data-processing applications that required rapid retrieval of stored data and programs. The main use of magnetic tape today is for archive and bulk information transfer from one machine to another. With home computers, it is used in audio tape-cassette recorders to provide low-cost auxiliary storage.

In a magnetic tape system the digits to be stored are recorded as changes in magnetization along tracks on a narrow ribbon of plastic film coated with iron oxide. There are typically nine tracks across such a tape that measures ½ inch (1.3 centimetres) wide and up to 2,400 feet (732 metres) long. The recording density along each track is usually 800, 1,600, or 6,250 bytes per inch, these being successive recent industry standards. The storage capacity of a full reel of tape is thus more than 100,000,000 bytes and with improved recording methods even higher densities are becoming possible. To read any particular entry on a tape, it is necessary to be able to start and stop the tape in a few thousandths of a second. The mechanical design of the tape-handling equipment is thus a major problem. But the main drawback with tapes is the access time, since it may be necessary to traverse the whole tape to find a desired stored item. Tape stores are generally used only for information that can be processed as it arises or for information that has been stored in logical order, so that the tape-handling unit does not spend most of its time searching for the item in question.

Magnetic disk unit

Magnetic disk technology was applied to secondary data storage in 1962. A magnetic disk is a flat, circular, metal or plastic plate coated on both sides with iron oxide or other readily magnetizable material. Input signals are recorded on the surface of the disk as magnetic patterns in circular tracks by a recording head while the disk is rotated by a drive unit. The heads, which are also used to read the magnetic impressions on the disk, can be positioned anywhere on the disk with great precision. A collection of disks (often 11 in number), called a disk pack, is mounted vertically on the spindle of a drive unit, which is equipped with multiple reading/writing heads. A single disk pack may hold as many as 300,000,000 bytes. Magnetic disk devices have a distinct advantage over tape units. They can directly access a block of data on a precise track without having to pass over a major portion of its contents sequentially, thereby reducing retrieval time to a fraction of a second. Small, flexible plastic disks, commonly called floppy disks, were developed in the 1970s. Floppy disks cannot store as much information or retrieve data as rapidly as conventional rigid disks, but they are adequate for use with personal microcomputers where low cost and ease of use are often of primary consideration.

The optical disk mass-storage system, introduced during the early 1980s, developed from videodisc technology. The most sophisticated system of this kind can store up to 4,000,000,000 bytes, far more than the highest capacity magnetic disk units. Information is encoded by a laser beam on the metallic recording medium of a videodisc platter as a pattern of pits. For information retrieval, the digitally coded data are tracked by an optical laser scanner.

**Control.** The operations of a digital computer system are regulated and integrated by a central control unit and in some cases several subordinate control elements. The central control unit constitutes a principal component of the CPU. It selects and calls up program instructions from the memory in proper sequence and interprets them so as to direct the movement of data through the system for processing and output.

In a certain sense the operation of a control unit is analogous to that of a miniature but complete computer whose output signals run the other parts of the system. In effect, each computer instruction is broken into a sequence of small operations known as microsteps (or microinstructions), the whole sequence being called the microprogram for that instruction. A computer designed as a microprogrammed machine has a pleasing simplicity in its structure and is relatively straightforward to understand, build, and extend.

**Processing.** All data in the main memory are transferred to the arithmetic/logic unit for processing, which involves basic arithmetic functions and certain logic operations. The latter includes the comparing of data and the selection of changes to the normal instruction sequencing.

The arithmetic/logic unit

The ALU consists of two basic components: registers and operational circuits. The registers are storage elements that temporarily hold binary digits representing the data being processed. They are composed of flip-flops, circuits capable of assuming either of two stable states representing 0 and 1. The operational circuits include such binary logic coincidence circuits as the AND gate, OR gate, NAND gate, NOR gate, and inverter. These elementary gates are combined in various arrangements to perform the arithmetic and logic operations.

**Output.** A wide array of devices, some of which can also be employed to input information, convert electrical signals from the ALU into a form usable by a human operator or a machine controlled by the computer. The most common output devices include video displays, high-speed printers, and console typewriters. A recent development in output equipment is an audio response unit that can convert data into spoken messages. Some of these units construct sentences with human speakers' words that have been digitalized and stored in a memory device. Others synthesize human speech, providing the appropriate response by generating signals similar to those produced in regular speech.

Audio response devices

## METHODS OF OPERATION

Most I/O devices are relatively slow (speeds are measured in milliseconds) compared to the speed of the central processing unit (in nanoseconds). A computer program that entails a good deal of data input and output is slowed down considerably while the CPU waits for I/O operations to be completed. There are several ways of eliminating this problem or at least of minimizing it. The first approach proposed was to prepare a batch of programs and data on a magnetic tape unit using a simple, inexpensive computer to collect the data from input devices and write the tape (*i.e.,* describe the jobs at hand and specify what has to be done to complete them). The main computer would then pick up its tasks from this tape and, in turn, write the results onto another tape for later printing. In this way the main computer would not be involved in any extremely slow I/O operations (*e.g.,* those of a card reader), and so would operate at nearly full speed.

A much better approach was devised some years later. It involves a single computer that prepares and batches jobs, runs through them, and prints the results. The computer shifts its attention from one activity to another periodically. This switching occurs in response to a signal, called an interrupt signal, from a remote I/O device indicating that it requires program attention. After transferring data and restarting the peripheral device, the computer resumes its main activity—that of running the program at the head of the batched sequence of jobs. This technique is known as multiprogramming and, in a more refined form, underlies the operation of practically all present-day digital computers. In multiprogramming, the operating system (see below) is a control program designed to schedule and run the several activities that the computer has to execute in a manner that makes reasonably efficient use of available resources and prevents the different

Multiprogramming

programs from interfering with one another. If several interconnected computers are involved, the same kind of controlled cooperation can be achieved while bringing more total computing power to bear. A style of operation known as time-sharing makes it possible to execute many different programs simultaneously, each in communication with a user at a terminal linked to the system. In a large modern time-sharing system, several hundred users may be active at one time, each user receiving immediate response to commands typed at his terminal. The total computing power, of course, is shared among the users, and a certain amount of it is lost in organizing the sharing operation itself, but the perceived effect is that all users are connected to a large mainframe that is working for each of them individually.

When a computer is used to react to situations and to exert some control over them, the delay time before the computer responds may be critical. The term real-time operation is applied when a computer has to keep up with external events such as controlling a segment of a chemical processing plant or the landing of an aircraft.

A conventional digital computer has a single instruction stream; that is to say, at any given instant only a single instruction is executed regardless of the number of active programs and users. Historically, increases in computer speed have been achieved by simply accelerating the rate at which instructions are carried out. At present, computing speeds are fast approaching that of light—the maximum speed at which an electronic impulse can pass from one point to another. Once this limit is attained, further increases in computing speed will have to come from computer designs that permit multiple instructions to be executed simultaneously. Since the early 1980s much research has been devoted to parallel computer architectures and to programming languages that can exploit such parallelism (see above). (A.M.P./N.E.W./S.F.We.)

## Digital computer programming

### PROGRAMMING FUNDAMENTALS

As has been indicated, the power of the digital computer rests not only on its speed and capacity for manipulating large quantities of coded data but also on its ability to perform complicated sequences of such actions under control of a stored program. This permits the same computer to work at a variety of tasks in sequence or even intermittently. Before the computer can work at all, however, a program or programs must be prepared in some suitable programming language and introduced into the computer's memory to instruct the computer exactly what to do. A computer task is first formulated and then expressed in an appropriate language, presumably one suited to the application and the thought processes of the formulator. The specification thus expressed is then translated, perhaps in several stages, into a coded program directly executable by the computer on which the task is to be run. The coded program is said to be in "machine code," or "machine language," while languages suitable for original formulation are referred to as "problem-oriented languages," or "high-level languages." But why are special problem-oriented languages needed at all? Why not use everyday English (or French for that matter) to instruct the computer what to do? The reason is obvious—the natural language of humans is imprecise. It gains understandability through many devices other than just the words themselves, as for example through inflection, hand waving, relation to previous conditions, and alternative phrasing. A compromise language between those of humans and machines must be as natural and easy to use as possible but have certain artificial characteristics and restraints in order to make it deterministic without requiring devices such as those mentioned above.

A wide variety of problem-oriented languages has been developed, making the process of successful program preparation depend more on understanding the problem than on understanding the computer that is to run it. Indeed, most of these languages are essentially machine-independent and can be used with any computer for which appropriate translation programs are available. Unfortunately some computer systems require minor alterations to the standard programming languages before they will accept them; some computer centres or installations also have their own particular rules or conventions, which are slight but crucial variations of the languages, all of which make it somewhat more difficult than might be imagined to transfer programs from one computer to another.

The way in which a computer system is operated varies according to its purpose. Early attempts to optimize the efficiency of system use on large computers resulted in the concept of "batch" mode: each job is submitted in its entirety (in the form, say, of a program on punched cards), and results are returned to the submitter at some later time (in the form, say, of printed tables). More refined versions permit the submitted job to call for program segments or data pools obtained from "within the system" (*i.e.,* stored on mass-storage devices available to the system) and for submission of jobs from locations removed from the system proper (*i.e.,* "remote batch").

The notion of the "batch" mode

An inherent property of any batch operation is that the user cannot interact with the program while it is running. He merely sets up the program, submits it, and, some time later, sees the results. A second and more modern mode of operation is the interactive mode, in which the user can interact directly with the running program. He can, for example, input data to the program, receive messages, modify the program's operation, and even stop the program if necessary. The interactive mode can be implemented on large multiprogrammed systems in which the host computer serves, for a fraction of a second, each of the many interactive jobs currently in the system (and, if time permits, runs some "background" batch jobs as well). It also can be implemented in a single-user system, as is done in most microcomputers. The interactive mode is better than the batch mode for both program development and use. In development, it allows the programmer to deal with problems as they arise rather than only after the program has concluded. In operation, an interactive program can be made more user-friendly, set up so as to prompt the user for information, report errors as they occur, and respond almost immediately to user requests.

Programs that are used in control applications (see below) can also be characterized as executing in the interactive mode, even though the communication/interaction is with another machine rather than with a human. Such a mode makes possible a fundamentally different style of use and so has implications for programming languages.

Computer systems are supplied with a variety of programs designed largely to assist the user to run jobs or to optimize system performance without unduly penalizing the user running jobs. This collection of programs constituting the operating system is as important to the operation of the computer system as the hardware itself. The operating system is usually supplied by the computer manufacturer or a firm specializing in program design. Successful operation of a computer system depends on the stable and reliable function of its operating system, which is usually "frozen" and not subject to changes, even to improve performance, until the implications for operation have been first explored. Current technology makes it possible to build some operating characteristics into a processor at the time of manufacture in the form of fixed programs introduced by customer orders. This leads to a further distinguishing feature known as firmware.

The operating system

During normal operation, the operating system is in general, overall control while user programs are being executed. It decides when a particular program should be run and what to do when a certain process terminates (subject to the ultimate control of human operators). In most cases, various utility programs that operate as independent units to facilitate the programming process are also available within the system. In this category are translators, which transform an entire program from one language to another; interpreters, which execute a program sequentially, translating at each step; editors, which allow text files to be entered or altered; or debuggers, which execute a program piecemeal, interpretively or not, and monitor various circumstances to permit the programmer to check to see whether the operation of the program is correct.

ranslating ıto ıachine nguage

Translators are classified as assemblers or compilers. An assembler is a translator that takes as input a program in a form close to machine language, called assembly language, in which each instruction step that the computer must obey is identified explicitly (but symbolically). The assembler translates this program into actual machine language, or a form that can be so rendered at the time the program is loaded into the computer's memory ready for execution. A compiler, by contrast, is a translator that takes a program written in a higher-level language and translates it into some other language—more often than not, the machine code of the target computer or something very like it. Since some higher-level languages are more or less specifically problem-oriented, there may be more than one stage of compilation. A program written in a special simulation language, for example, may be translated at one compilation stage into a more general-purpose problem-oriented language, which is then subjected to a second stage of compilation into some intermediate form similar to assembly language, and from there in a third stage to machine language. The user need be aware in detail only of the features of the highest level, although some awareness of the total process is helpful in designing programs that are efficient, both running and being translated. In particular, only the final code generation phase need be geared in detail to the particular hardware configuration on which the program will run, although general features such as the number of input/output units must be taken into account at higher levels.

Programming languages differ among one another by level of generality and intended class of applications, which are reflected in the data structures handled, operations allowed, file and communication features incorporated, and program structures assumed. Data structures are composed of elementary units of various data types, including integers, noninteger numbers, bit patterns, and strings of characters, and their composition into more complex structures. Operations are defined appropriate to the various data types and structures available; arithmetic operations on integers and noninteger numbers, matrix arithmetic operations on arrays of numbers, and logical, or Boolean, operations on bit patterns. The operations are invoked in program statements for combining specified units or structures, referred to by variables, and assigning the result to another variable or set of variables. Program structure has to do with how a program is organized for execution purposes and may involve concepts of unconditional or conditional transfer of control (to cause sequencing to pick up elsewhere in the program, perhaps only if some condition is satisfied), subroutines (procedural units treatable as self-contained), recursive procedures (procedures that invoke themselves), and block structure (program blocks outside of which variable names may be reused without ambiguity). Various control statements reflect program structures (*e.g.*, jump statements and subroutine calls). Finally, a programming language usually permits a variety of supporting, or nonexecutable, statements, which are really messages to the compiler regarding data formats and related elements.

The versatility of this approach to computing has resulted in an almost standardized system approach when converting a problem into a program. The problem is analyzed, and the input and output formats are designated. An algorithm is then designed. It specifies clearly and precisely the procedural sequence by which the input data are to be converted into the output solution to the problem, although the detail of how each step is to be accomplished may be omitted. There are several ways of expressing such an algorithm on paper. The most popular of these used to be the flowchart, a special type of diagram that shows clearly the logical sequence of procedures to be followed together with the interrelationships of different parts of the proposed program. Recently the flowchart has gone very much out of fashion due perhaps to the increased use of more structured languages. It has largely been replaced by outline programs prepared in a pseudo computer language that can be written quickly and easily and that is readily converted to a suitable computer language when the details are later added.

low- ıarts ıd rogram utlines

Many computer users have difficulties in transposing their problem into a usable computer algorithm. Block diagrams, flowcharts, and program outlines all facilitate the required attention to the interrelationships, logic, and sequence of problem components. Programming follows development of such an aid. The computer cannot exercise what in humans would be called judgment or common sense, and it must be meticulously instructed in the program as to how to handle every contingency: what to do if the answer of an intermediate calculation becomes zero or when particular exceptions to anticipated results occur. Each eventuality must be anticipated and analyzed completely or the program will not work satisfactorily. The outline algorithm is converted to statements in a language chosen as most suited for the specific problem and also as one that the particular computer system being used will accept (*i.e.*, for which it has a translator). In a succession of translation stages, the program is converted automatically to the specific machine code of the particular computer. Testing with sample data and debugging (correction of program errors) of the program follows, perhaps with the help of a debugging program. The program is then ready for production running, or submitted input, which will be converted to printed, graphic, or other form of output. Modifications may finally be made for alternative solutions or improvements.

## LANGUAGE CATEGORIES

Assembly languages

Although programs must ultimately be expressed in the relevant machine language in order to be accepted by a computer, machine language programs are long and tedious to write and provide programmers with many possibilities for committing errors. To alleviate some of these problems and to shorten the programmer's tasks, assembly, or mnemonic, languages were created. A mnemonic code is designed to assist the human memory in that it is a standard formulation of letters standing for a known word (*e.g., mpy* for multiply or *sto* for store). Programmer-invented mnemonics can also stand for the coded addresses of storage locations with reference to the meaning of the data stored at those locations, such as *price* and *quantity* for the corresponding values in a problem in economics. Codes of this type are converted into machine codes by an assembler, which is the translator for assembly language programs. An assembler then prepares a machine language program from a symbolic language program by substituting machine operation codes and absolute addresses for symbolic (mnemonic) code words used by the programmer. This speeds up the programming process considerably, since it is much easier to remember that SUB price means "subtract the value of the variable called price from the current value" than to remember that the operation code for subtract is 15 and that the value being referred to is in store location 1074. Assembly code is specific to a particular class of machines; consequently, assembly language programs are not easily transferred from one type of machine to another.

Higher-level languages are even easier to use because individual constructs in a high-level language normally result in several low-level machine instructions. Thus programs are shorter and can be much more natural in their expressions. High-level ideas, such as repeating large sections of code for as long as some condition is true, can be expressed easily without the programmer's needing to know anything about how the tests will be carried out or where temporary results will be stored. In addition, programs written in high-level languages can be more readily transferred from one machine to another.

Most high-level languages provide operations and data structures that are not directly supported by the underlying hardware. The trend in programming languages is toward higher levels of abstraction for both data and programs and for greater independence of modules. Object-oriented languages permit the programmer to work at a higher level, concerned more with what has to be done than with the details of how it is to be accomplished.

The variety of higher-level languages continues to proliferate as new applications and wider use of computers develop. The descriptions of the languages to follow are di-

vided into the following classes: (1) fundamental algorithmic and procedural languages, (2) variants for interactive use, (3) nonnumerical languages, (4) simulation languages, and (5) process- and numerical-control languages.

**Algorithmic and procedural languages.** *FORTRAN.* FORTRAN, originally an acronym for *for*mula *tran*slation, was developed for scientific problems and programs but is now also widely used in commercial and educational computer applications. FORTRAN II was the first programming language to be accepted worldwide as a powerful, convenient, and practical tool. The first version of FORTRAN II, which ran only on the IBM 704 processor, appeared in 1958. It was quickly copied and modified to run on other processors. By the mid-1960s many modified and extended versions of FORTRAN existed, and it was felt necessary to introduce some standardization to help allow programs to be transferred from one FORTRAN machine to another and run without modification. The American National Standards Institute (ANSI) adopted standards on FORTRAN in 1966. ANSI Standard FORTRAN IV, which included relaxations of some of the restrictions of FORTRAN II as well as some new features, was published later that same year. Thereafter, FORTRAN implementations were based on this standard until 1977, when a new standard, called FORTRAN 77, was introduced. This was a major extension to FORTRAN IV but remained essentially "upwards compatible" with FORTRAN IV; that is, correct ANSI FORTRAN IV programs (with minor exceptions) remain correct and acceptable under the new standard.

Uses of FORTRAN

FORTRAN is basically a programming system that includes the language and a compiler that permits programs to be written in a mathematical-type language. It is essentially machine-independent and procedure-oriented. The grammar, symbols, rules, and syntax used are, in general, simple, with easy-to-learn mathematical and English-language conventions. Simply stated, all FORTRAN-type languages treat arithmetic operations with commands that evaluate expressions and substitute the result for current values of variables. Also included are statements for transfer of control, looping (designating a set of statements to be executed more than once), and input/output.

*ALGOL.* ALGOL alternately stands for either *Algo*rithmic *L*anguage or *Alg*ebraic *O*riented *L*anguage. It is a result of international cooperation to obtain a standardized algorithmic language, which now exists as two major standards, ALGOL 60 and ALGOL 68.

Applications of ALGOL

ALGOL is principally used in the programming of scientific problems (though ALGOL 68 is a good general-purpose language), but it is also selected by many programmers as a reference and publication language and as a model for the invention of new artificial languages, compiling techniques, and mathematical structures. It has become an internationally accepted procedure for designing mathematical algorithms.

*Pascal.* Pascal is an ALGOL-related language whose name is derived from the scientist-philosopher Blaise Pascal. It was introduced in the early 1970s and was originally developed as a teaching tool for computer programming.

Not only is Pascal extremely effective in structured programming, but it also is one of the few procedure-oriented languages specially designed to aid programmers in pinpointing coding errors quickly. Capable of processing both numerical and textual data, this compact, highly efficient language is now extensively used by academic institutions as well as by business and industry, particularly for minicomputer and microcomputer applications such as typesetting and editorial operations and computer-aided design and manufacturing (see below *Design*).

*Ada.* Named for Augusta Ada Byron (Countess of Lovelace), a collaborator of the inventor Charles Babbage, Ada is a programming language based on Pascal but which includes additional features for modularity, portability, and concurrency. Ada was developed during the late 1970s under the sponsorship of the U.S. Department of Defense primarily to meet the special needs of military-related programs; however, it may be suitable for a variety of business, industrial, and other civilian applications.

*C.* Introduced at Bell Laboratories in 1974, C was originally developed for use in the UNIX operating system. It is a general-purpose programming language that is similar to but more flexible than Pascal. It features both the efficiency of a low-level language and the control and data structure of a high-level language, making it especially useful for writing systems programs.

BASIC:

```
10  X = 5
20  Y = 8
30  GOSUB 100
40  PRINT X, Y
50  END
100 REM SWAP X AND Y
110 Z = X
120 X = Y
130 Y = Z
140 RETURN
```

Pascal:

```
Program Swap(input, output);
    Var  x, y  :  Integer;
Procedure Exchange(a, b: Integer);
    Var  c  :  Integer;
    Begin
        c := a;
        a := b;
        b := c;
    End;
Begin
    x := 5;
    y := 8;
    Exchange(x,y);
    Writeln(x,y);
End.
```

C:

```
Main()
{
    int   x, y;
    x = 5;
    y = 8;
    Exchange(&x, &y);
    printf("%d, %d\n, x, y);
}
Exchange(a, b)
    int   *a, *b;
{
    int   c;
    c = *a;
    *a = *b;
    *b = c;
}
```

BASIC, Pascal, and C programs for exchanging items.

*COBOL.* COBOL is a procedure-oriented language originally designed to meet the needs of business programming. The name is an acronym for *Co*mmon-*B*usiness-*O*riented *L*anguage. COBOL was developed in 1959 and has been maintained and revised since then.

Advantages of COBOL

The advantages of using COBOL can be summarized as follows: COBOL programs are stated in precise, easily learned natural words and phrases; they can be read and understood by nontechnical people with a minimal background in data processing; programs written for one computer can be run on another computer with minimum modification to take advantage of the second computer's features; training time is negligible for the novice in programming; and program testing is simplified and can be completed by someone other than the original programmer. COBOL contains many important file-organization features and can deal with variable data length. Input/output procedures and report generators are its strong points.

*Program language 1.* PL/1 is a multipurpose programming language designed for solving both business and scientific problems. It incorporates many features of both FORTRAN and COBOL. FORTRAN's advantages are reflected in its simple, concise statements, while those of COBOL lie in its ability to manipulate and its easily input/output grouped files.

PL/1 also has a block structure, as first introduced in ALGOL, and is often classed as a language bringing "all things to all people," containing as it does the amalgamation of many features from its various predecessors.

**Interactive languages.** For the operation of computers through a terminal, or indeed when the computer is *in* the terminal (as in the case of home computers), it is often convenient to provide a single language for working the system, writing and running programs, changing them, finding flaws, and saving or fetching different versions of the programs. Interpretative execution allows the dialogue with the computer to be conducted solely in that language, with the result that there is less to learn and less confusion for an inexperienced user when bugs in the program are detected. Interpretative execution, however, slows down operations by a large factor and consequently is not very useful for bigger jobs. Furthermore, the languages that are amenable to this sort of use are themselves rather primitive and weak. Nevertheless, they still have their place.

Interpretative execution

*BASIC.* BASIC is an acronym for *B*eginner's *A*ll-purpose *S*ymbolic *I*nstruction *C*ode. It is a FORTRAN-like language that can be learned readily by the average non-computer-oriented individual. The simple conditions and attributes of the language allow it to be mastered with practice in a few days. It is commonly used for business, educational, and commercial purposes.

Each line of a program written in BASIC begins with a number (the line number) that identifies the line and specifies the order in which the statements are to be performed. The computer sorts out the program before running it, and, therefore, the statements need not be input in any specific order.

BASIC, though much less powerful and versatile than, say, ALGOL or FORTRAN, is considered to be adequate for many of the commercial and business problems that run on small business systems. BASIC permits conversational statements, free-style input, segmenting of complex statements, six significant digits of accuracy, easy and safe program modification, and editing functions that permit combinations of two or more programs into one and allow selection from a library of stored programs or functions such as procedures for solving simultaneous equations, curve fitting, and statistical analysis.

*Forth.* Forth is presently gaining popularity as an interactive language because it runs relatively fast, is very expressive, and encourages a good style of programming. The name derives from an intention to provide a language for fourth-generation computers. However, Forth is in reality quite machine-independent. It uses a notation known as reverse polish in which an operator is always preceded by its arguments. Thus a + b is written a b +, and (a + b)*c is written a b + c *. The style, though at first sight hard to follow, can actually be learned quickly. The language is readily extended by the user and is finding application in many varied areas, including word processing, instrumentation and control, and video games.

*Logo.* Logo is a simple, interactive language. Though compact enough to run on most microcomputers, it embodies powerful programming facilities. Logo permits modern control structures, including recursion, and has powerful graphics capabilities. It is used extensively for teaching programming to children.

Purpose of list-processing languages

**Nonnumerical languages.** Nonnumerical languages are oriented specifically for the manipulation of data, especially nonnumerical data, whose length and structure change considerably during the calculation of a problem solution. It is because the data are usually not numerical that the concepts and terms referred to as symbol manipulation and nonnumerical data processing are most often associated with list-processing languages. String-manipulation languages that deal only with strings of characters are a special case of list-processing languages. Most general list-processing languages are used as research tools and not necessarily for production-line programming.

A compiler for some programming languages is an example of a program that uses symbol manipulation ideas; it will adopt list-processing techniques in some form or other. List-processing procedures are oriented more to techniques of programming than to particular applications except perhaps in the areas of information retrieval, algebraic manipulation, and artificial intelligence.

A list is defined simply as a set of items given in a specific order. Some characteristics of lists are: (1) they can be of variable length subject principally to gross machine limitations, and the lengths can vary during the computation; (2) items of many different types, such as numbers and alphanumeric symbols, can appear on lists; (3) items can be added at the beginning, end, or middle, and any item can be deleted from the list; (4) any number of distinct lists can be created or developed from a program; and (5) lists can be referenced by the program and can be copied, modified, examined, and analyzed by programs. Elements of lists can themselves be lists, allowing the recursive decomposition of large constructs into smaller parts.

*SNOBOL.* SNOBOL is an acronym for *Strin*g-*O*riented Sym*b*olic *L*anguage. It has significant applications in program compilation and generation of symbolic equations. It is a programming language that provides facilities for the manipulation of strings of characters by pattern-matching expressions. SNOBOL is particularly applicable for programs associated with text editing, linguistics, compiling, and symbolic manipulation of algebraic expressions.

*LISP.* LISP, standing for *Lis*t *P*rocessor, differed radically from other languages available at the time it was introduced in 1960 and has remained something of a puzzle to many programmers who have mastered a more traditional language such as FORTRAN. It is in fact deceptively simple. The basic entity is an s-expression (short for symbolic expression), which is either an atomic symbol or a list structure of the kind described earlier. S-expressions represent both the data and the functions to be applied to them, and the result of evaluating some s-expression is simply another s-expression. Thus a program and its input and output are all of the same form.

LISP elements

Running a LISP program consists of evaluating the s-expression for it, a particularly elegant feature being that the s-expression corresponds to a formula and not to a procedure—a matter of great satisfaction for mathematicians. To be sure, some of this elegance is dissipated in practice by the use of an escape mechanism (called PROG) in the language that provides some of the more usual ALGOL-like constructs, such as local variables and assignments. A LISP program using PROG is no longer a formula, but it may deliver the required results with greater speed or less use of memory.

*PROLOG.* This language, whose name derives from *Pro*gramming-in-*Log*ic, is gaining popularity in many areas of application. With PROLOG the emphasis is on description rather than on action, more on what is to be done than on how it is to be done. To find, for instance, the greater of two input numbers, one would describe what the greater-of relation meant and then query it with the given numbers as data. In a conventional imperative language such as FORTRAN, one would give instead a sequence of statements that took the given numbers, compared them, and then emitted the bigger one. The difference between these two approaches varies with the job at hand and may be minor or profound. PROLOG works by accepting a query and trying to find values for the variables in it that are consistent with the definitions given in the program. It does this by searching the definitions for each condition, generating new queries as it goes for those that match. This style of computing is called pattern-directed role-based programming and is useful in several applications of artificial intelligence (*e.g.,* constructing expert, or knowledge-based, systems).

**Simulation languages.** Simulation can be considered a representation of a changing, or dynamic, system developed to simplify manipulation and study while using a computer for computations, comparisons, and analyses. The system in its essence is a set of interacting components and processes. The interactions may be internal or they may be linked to an external environment. A model is a representation of a system in which the processes or interactions bear a close resemblance or relationship to those of the specific system being simulated or studied. As a consequence, computer models used in simulation are seldom highly abstract or strictly mathematical. Manipulation of a simulation system concerns the acceptance of inputs and the generation of outputs similar or analogous to those of the system represented.

Though almost any type of programming language can be used to implement a simulation model, various specific languages have been developed for simulation applications. They contain status descriptors, which are definitions of the essential elements of a model; procedures for modifying the state of a model; and procedures for controlling the dynamic performance and observations of model behaviour. Most simulation languages develop data structures, data transformations, sequencing of transformations, output routines, and operational features. In addition, extensive error checking is available to prevent misuse of the language mechanically (*i.e.,* hand coding or keypunching), as well as routines for debugging a model and locating logical errors.

Components of simulation languages

*SIMSCRIPT.* SIMSCRIPT is a proprietary general-purpose simulation language and system. It is based on the notion that the state of a system is definable and can be described in terms of entities (*i.e.,* the specific objects or things of which a system is composed) and attributes (*i.e.,* those properties that are associated with the entities as well), and as sets (*i.e.,* groups of entities). All entities must be specific and explicit with a complete list of their attributes and possible set memberships; this is the major

prerequisite in the development of a simulation model. The state of a system once described is changed by the occurrence of an event that is a user-defined subroutine written with either SIMSCRIPT or FORTRAN statements. Entities can be created or destroyed; set memberships of individual entities can be altered; and numerical values of attributes can be changed. Because entities and attributes must be individually located, much of the SIMSCRIPT language is devoted to providing convenient and flexible methods for performing storage and retrieval functions.

Components of GPSS

*General-Purpose System Simulator (GPSS).* This simulation language has three basic components: blocks, transactions, and equipment. The language develops simulation models through the use of and in the terms of block diagrams. Such diagrams are graphic devices that portray the logical and physical flow of transactions or basic information through a system. Such block types in the simulation block diagram are very specific, and from them models are built. Punched cards or other entry program segments define the properties of each block in the model. Temporary GPSS elements are formed into transactions and moved from one block to another within the simulation model. Each transaction has eight parameters associated with it and eight possible priority levels. As transactions enter blocks, various subroutines associated with those particular block types are interpretatively executed, thereby causing a modification of one or more status descriptors.

**Process- and numerical-control languages.** Languages have been developed to facilitate the use of computers controlling industrial processes and machine tools. These vary widely, because each application can be very different. Only one example, APT (Automatically Programmed Tools), is discussed here. It is not necessarily typical of such languages, but it is fundamental. Using APT, the designer describes his tool and the desired part in a high-level, geometrically oriented language. A preprocessor program accepts the high-level language and digests it into a simpler, formalized internal representation. The central program, which is tool-independent, converts the material, tool, and geometrical information into tool motion commands. A postprocessor program prepares the tool motion information in a format suitable for the particular control mechanism being used. (C.J.S./A.M.P./N.E.W./S.F.We.)

Programming machine tools with APT

## Applications of digital computers

In the United States, the Bureau of the Census and military test establishments were among the initial users of computers for data analysis. Others who made early use of the computer's ability to perform rapid calculations were university researchers in science and engineering. Soon industry and business began employing computers for a multiplicity of operations. Worldwide growth in the value and number of installed computer systems was rapid. Today, computers affect nearly every aspect of daily life. Their applications cover so wide a field that only a few categories can be surveyed here. These are: control, communication, simulation, design, and artificial intelligence.

### CONTROL

The digital computer has proved highly effective in applications in which it is employed to adjust the performance of a system or device without human intervention. There are a myriad of such applications, the most important being the control of industrial processes and machinery.

**Process control.** Process control involves the use of digital computers to provide automation of continuous industrial operations, such as those of petroleum refineries, automobile assembly plants, and electric power stations. With the development of increasingly reliable and inexpensive models during the late 1960s, digital machines quickly supplanted analog computers in many industrial control systems. The digital controllers offer a distinct advantage over their analog counterparts. They can be programmed readily to handle a broad range of separate tasks, and their programs can also be revised relatively easily to accommodate any changes in operation without having to modify the hardware of the control system.

Several different process-control procedures have been devised. Distributed, hierarchical control, for example, is extensively used in manufacturing plants and other industrial facilities where precise and rapid control of processes is essential. It entails the use of digital computers in control situations at several different levels in an integrated fashion. In such a system, mainframes handle overall plant-level operations; they determine the optimal production schedules to meet specific quotas. This information is transmitted to the next level down—to minicomputers that provide supervisory control. Besides keeping track of inventories of raw materials and manufactured items, these computers perform optimization calculations, the results of which are relayed to lower-level microcomputers that directly control individual production units.

Distributed, hierarchical control

In certain factories, such a production unit might consist of a robot mechanism controlled by a microprocessor. These one-arm robotic devices, capable of approximating the articulation and movement of the human arm and hand, have largely automated automobile assembly lines. They are used for such functions as welding, spray painting, and lifting automobile bodies. The mechanical arms can be programmed by physically moving them through the desired motions. The movements are recorded in the microprocessor's memory so that they can be repeated precisely. Some advanced, high-performance industrial robots have optical sensors that enable them to locate and identify components on a conveyor belt and make certain that the dimensions of these items meet specifications.

Computer-controlled industrial robots

**Numerical control.** Another computer application in the manufacturing field is the numerical control of such machine tools as grinders, lathes, and drill presses. In contrast to process control, numerical control provides automation of discrete operations. In most modern control systems of this kind, the dimensions and other specifications of the component to be milled are fed into a minicomputer, which then generates numerical control instructions and transmits them directly through communication lines to the milling machine.

**Transportation control applications.** Automated systems combining digital computers, radar units, and auxiliary electronic equipment have been developed to handle the ever-increasing volume of air traffic. These systems aid air traffic controllers in directing the continuous flow of incoming and outgoing airplanes at large airports. They can pinpoint the position of every plane within 50 miles (80 kilometres) of the airfield and produce warning signals when two planes come too close together.

Many of the operational systems of the aircraft itself are computer-controlled. Most commercial passenger planes are equipped with digital computers programmed for automatic flight control and power distribution regulation. Highly sophisticated computerized inertial navigation systems are utilized on board such spacecraft as the Space Shuttle; they also are used for deploying weather and communications satellites and guiding ballistic missiles.

Computer control of aircraft and spacecraft

The management of rail yards has been facilitated substantially by computerized systems that coordinate the signaling and switching functions of classification yards where freight trains are sorted and assembled. Electronic scanners read colour-coded identification labels on all freight cars entering a classification yard and send the information to yard computers that assign the cars to the appropriate track. Automatic computer control has also been adopted by many passenger rail lines to improve service. In certain cases, both the operation of the trains and the ticketing of passengers are completely regulated by a central computer.

Computers are being applied to the problem of automotive traffic control in numerous cities. In one type of control system, traffic flow on surface streets is regulated by computer-synchronized traffic lights. In another type, congestion on expressways, highways, and other major roads is alleviated by controlled access. Roadside sensors monitor vehicles approaching an entrance ramp; information from these electronic devices is analyzed by a computer at a remote traffic control centre to determine whether or not to allow access to vehicles waiting at a signal light on the ramp.

**Other control applications.** A large number of instruments employed in scientific research, including particle accelerators, radiation detectors, electron microscopes, and radio telescopes, are so complex that their operations have to be controlled and checked by computers. The same holds true for various kinds of sensitive instruments used in hospitals and clinics for diagnosis, treatment, and monitoring (*e.g.,* continuous observation of vital signs during anesthesia).

The availability of inexpensive microprocessors has given rise to many kinds of computer-controlled consumer products. A few common examples include programmable microwave ovens, self-tuning television sets, and cameras with automatic focusing. Automakers have been quick to introduce computerized ignition and fuel systems designed to improve fuel economy and performance. Some home security systems depend on direct digital control of sensors and alarms.

## COMMUNICATION

Communication applications involve the acquisition, analysis, organization, storage, and dissemination of information by computers. They also include the use of digital computers to control such facilities as telephone switching networks and microwave relay stations for satellite communications.

Computer data files

Government institutions, including law enforcement agencies, rely heavily on computers to maintain enormous data files. With the aid of computers, government tax offices throughout much of the world process and record information from millions of tax returns filed by individuals and business firms each year. The U.S. Federal Bureau of Investigation likewise analyzes and stores large volumes of information (*e.g.,* fingerprints and arrest records) in computer data bases. Computerized data-base systems also have been developed for use in hospitals and medical research centres. Some of these provide access to patient case histories, while others are designed to serve as disease registries. In recent years computerized data retrieval services have become available for certain professions. Scientists, for example, can subscribe to a service that furnishes bibliographic references to the latest research papers and reports published in their respective fields. Information services for personal computer users have appeared in North America, Europe, and Japan. These systems allow subscribers to access educational and entertainment materials as well as news and business reports from central data bases by means of telephone lines or television networks.

Computer-assisted instruction

Computers have begun to play an increasingly important role in education. Besides providing convenient access to reference and instructional materials, they serve as efficient teaching aids. More and more institutions of learning, including elementary schools, universities, and industrial and military training centres, use some form of computer-assisted instruction (CAI). Programs have been prepared in almost all major subject areas for many different age levels. Those written for remedial instruction of basic mathematical and language skills have been particularly effective. In a CAI system students typically interact with the computer by means of a video display terminal. This interaction may take the form of reading text on the display screen and typing answers on a keyboard. Input may be by means of some other device, such as a joystick or a "mouse," which allows pointing, drawing, or selecting options presented on the screen in the form of a "menu." Graphic output may take the form of still images, simple animation, or video playback. On the whole, students find it easier to understand how, for example, a piston works when shown a representation of its movement on the screen, together with an accompanying description, rather than by reading several pages of text alone. This helps to make the learning experience more enjoyable, with students readily assimilating facts and principles that they may have failed to understand when presented in other more traditional ways. A major advantage of CAI systems is that instruction is ordinarily on a one-to-one basis. The computer is able to analyze the responses or questions of a student and automatically modify the speed and method of presentation according to the student's particular ability level.

Insurance firms, stock brokerages, and many other businesses make use of computers for record keeping and exchange of data within the company or between companies. Here again, computer graphics is making its mark. Managers, in general, are not interested in studying page upon page of figures before making a decision; communication of ideas is greatly improved when facts are presented in clear, concise graphic or pictorial form.

EFT systems

Another form of computer communication is being established by the banking industry. Known as electronic funds transfer (EFT), it is becoming the standard means of conducting financial transactions. This system permits the movement of money via high-speed communications lines. If fully implemented, EFT would virtually eliminate the use of cash, checks, and credit cards. In this electronic banking system, salaries, social security payments, and other income are credited to a user's account. Payment of utility bills, rent, or mortgage loans is also made directly, with the amount of the outlay deducted from the account balance. Another feature of EFT is the use of remote point-of-sale terminals linking retail stores and banks that allows purchases to be charged against bank accounts.

## SIMULATION

As noted earlier, simulation is the imitative representation of the behaviour of a dynamic system (actual or postulated) by that of another system—namely, a computer-based model. The main objectives of this process are to test hypotheses and to design or experiment with systems that may not yet be available for direct observation or experimentation.

Advantages of simulation systems

Flexibility is provided because it is easier, less costly, and safer to change a representation of a system than to change the system itself. Modifications can be anticipated during model design and can be completed in various runs of a computerized simulation system.

A simulation system may be described either as continuous—represented by mathematical equations (especially sets of differential equations)—or discrete, in which major states are distinct and individually identifiable. If the simulation is deterministic, it contains elements whose behaviour is described a priori. On the other hand, if the system is modeled stochastically (*i.e.,* exhibiting a succession of random variations), pseudorandom variations in the essential elements (such as timing or quantity of deliveries) may be introduced. In the simulation of a supermarket checkout, for example, the effect on customer-waiting times of various checkout arrangements could be examined. Is it useful to have one checkout position reserved for customers with fewer than eight items? Or would having an extra person at each counter to pack the customers' grocery bags be more effective? These and other options can be examined within the framework of a computer simulation and the possible consequences thoroughly analyzed without disrupting normal operations.

Digital simulation has been employed by virtually every segment of industry to find solutions to a multitude of problems relating resource allocation to production scheduling. Biomedical researchers construct computerized models of basic physiological processes to better understand their functioning. Similarly, digital simulation affords chemists an opportunity to study in some detail complex reactions (*e.g.,* the interaction of ions in molecule formation) that they cannot otherwise observe. High-speed digital machines with enormous data-storage capacity have permitted meteorologists to build computer models of rapidly changing global weather patterns with which they can formulate reasonably accurate short-range forecasts. Earth scientists have evaluated and refined the theory of plate tectonics by programming computers to vary certain parameters in equations that govern pertinent geophysical processes and phenomena and then comparing the different possibilities with collected observational data.

Man–machine simulation models

Man–machine simulation systems include operational and behavioral models that require the active participation of a human being and the allocation of functions between the computer portion of the model and the person.

Many man–machine simulation models relate to training exercises in which a person must face a simulation of a real-life situation—as in flight simulation where complex computer systems accurately model the operation of an aircraft. The pilot sits in a mock-up of an actual cockpit and "flies" the simulator by interacting with the regular cockpit controls, all operations of which are monitored and interpreted by the computer system. The computer continuously updates the view from the cockpit windows and the readings on the dials, as well as causing the cockpit to physically tilt and twist in response to maneuvers or simulated atmospheric turbulence. The person controlling the simulator is able to duplicate the effects of changing weather conditions or engine malfunctions, enabling the pilot to experience various emergency situations without danger to life or aircraft. Man–machine simulation is also used in education systems (*e.g.,* management games) and behavioral experimentation, such as a study of psychological behaviour in specific controlled environments.

### DESIGN

Another highly productive application of digital computers is their use in design work. Commonly called computer-aided design (CAD), it has become very important in the field of engineering. Many industrial CAD systems have been developed over the years and are now employed in designing bridges, power plants, integrated circuits, machine parts, aircraft, and cars. Applications of design systems outside the field of engineering include designing patterns for clothing and page layouts for books, magazines, and newspapers. The style of the systems in the different design areas is extremely varied, although at a hardware level most systems consist of a computer with one or more remote workstations featuring a video display, keyboard, and interactive graphics input devices (often a digitizing tablet and stylus or a hand-sized device with interrupt buttons called a puck).

In some systems the emphasis is mainly on the production of accurate, possibly realistic drawings of components or assemblies (such as in architectural rendering). This may necessitate the use of a high-performance colour graphics terminal and software to perform perspective and shading calculations. For other applications it may be the relationship between components that is of greater interest, and the representation on the screen is only of secondary importance. For these applications less sophisticated display devices may be required, but more information will need to be held in the computer model.

The creation of a suitable computer-based representation of the object being designed can be a difficult and time-consuming part of the overall design process, particularly when the object to be modeled is three-dimensional. In some systems the designer may use a tablet and puck to trace plans or sketch a design, or he may compose a picture by selecting and positioning items from a catalog of basic objects. Thus, for example, a person designing the layout of an integrated circuit could indicate positions at which transistors or diodes are to be placed; the design program would draw the appropriate symbol on the screen and make the necessary adjustments to the computer model. In this application, the computer would probably also consult specified design rules (*e.g.,* minimum component separation) to check whether this is a valid placement of the particular component.

Once an initial model has been entered into the computer, it will almost certainly be necessary to modify the design in some way. It is at this stage that the real benefits of a computer-based system become apparent. A number of people can work on the same design at once, making simple, clean alterations to the model without large amounts of redrawing being necessary, and have immediate access to the master design held in the computer memory rather than working from paper plans, which may be out-of-date. Programs for analyzing the properties of the object being designed, such as its weight, strength, or electrical characteristics, may also be available.

CAE systems

In some systems the CAD process is integrated with computer-aided manufacturing (CAM), the other component of what is now commonly referred to as computer-aided engineering (CAE). By combining the design and manufacturing operations, the CAE system significantly increases productivity and reduces the time required to develop new products. In this type of system the computer assists the engineer in developing and revising a design and then prepares instructions for numerically controlled machine tools and orders the materials needed for the implementation of the design.

### ARTIFICIAL INTELLIGENCE

Efforts to compose music and to create visual art forms with computers date from the 1950s. While the works produced in this manner reflect the ingenuity of the individuals who programmed the computers, they cannot be regarded as manifestations of any genuine creativity on the part of the machines themselves. Computers have demonstrated a proficiency in theorem proving, chess playing, and other highly logical tasks. Their success at these activities is attributable partly to an ability to manipulate symbolic information at exceedingly high speeds and partly to a better understanding of the act of reasoning itself.

As yet, no computer system capable of approximating man's ability to reason, generalize, discover meanings, or learn from past experience has been developed. Recent research projects in artificial intelligence (AI), nonetheless, have yielded some results of immediate practical value. The most significant of these are related to decision making, natural-language processing, computer vision, and robotics.

**Decision making.** Knowledge-based software systems, sometimes called expert systems, provide computers with the capacity to make decisions for solving complex nonnumerical problems. These systems comprise hundreds (in some cases, thousands) of "if–then" logic rules, along with various rules of thumb, formulated with knowledge gleaned from leading authorities in a given field. An example of such a heuristic program is MYCIN, which physicians can consult for advice on treating certain bacterial infections of the blood. Using MYCIN, a computer arrives at its recommendations by first making a plausible guess as to the patient's condition on the basis of observed symptoms and then proceeding to see how well that tentative diagnosis fits all known facts about the behaviour of the microorganism thought to be involved. Once the computer has identified the cause of the infection, it considers the kinds of antibiotics available and makes its recommendation, usually offering several alternative forms of therapy. If requested, the computer will explain its reasoning, specifying the particular logic rules it followed in making its decision, and cite pertinent medical literature as well.

Heuristic programs

Various other useful expert systems have been devised. These include programs formulated to assist geologists in locating petroleum and ore deposits; to help chemists interpret data obtained through mass spectroscopy; and to aid computer sales representatives in configuring computer systems to meet customer requirements.

**Natural-language processing.** Programs that enable computers to comprehend and produce ordinary language (*e.g.,* English or Japanese)—either written or spoken—constitute another form of artificial intelligence with considerable potential for practical application. Programs capable of translating from one language to another on a relatively primitive level have been in use for a number of years; however, there are none as yet that approach the sophistication and proficiency of human translators. Also, practical natural-language systems have been developed for querying data bases on specific subjects. Significant progress is expected in the near future in both these applications, as well as in the production of natural-language components for expert systems.

The difficulties encountered in natural-language processing arise from the inherent ambiguity of everyday language and the fact that intelligent use of language depends on an understanding of the subject being discussed and the purpose of the discussion. This means that even though existing practical systems are capable of dealing with a wide range of grammatical structures, they tend to be limited to dealing with a specific subject in a specific

way, such as providing factual information in response to questions about a company ledger. Adapting a natural-language system to an individual user's needs is therefore often time-consuming and expensive.

**Computer vision.** To achieve vision, a computer needs an artificial eye and a program with which to process the data received by the eye into some meaningful interpretation of what is being seen. An eye can be quite simply furnished with a type of television camera that delivers to the computer a digitized signal representing the scene it is picking up. However, programs that enable the computer to analyze this signal and determine what is in the scene are still in a relatively early stage of development. Frequently the scene takes many seconds (even hours) for the computer to analyze, thus making it virtually impossible to react in real time to the familiar human-scale, natural world. Faster computers and further research are expected to alleviate this problem. In some cases, rapid analysis is possible but only for very specialized situations. The inspection of welds made by a (robot or human) welder, the checking of shapes cut by a wood-forming machine, and the microscopic examination of the surfaces of metallurgical samples are examples of computer-vision-associated activities that can be achieved with current technology.

Much useful work has been done on pattern recognition, an area of computer vision in which the ability to react in real time is neither required nor appropriate. Pattern-recognition research has led to the development of document readers capable of reading printed pages of text, transforming them into encoded character sequences, and storing them in the computer's memory in less time than the text material could ever be typed by a human. Pattern recognition has found numerous other applications. Computers, used in conjunction with microscopes, have been employed to detect and identify human chromosome and cell disorders. Computer processing of images transmitted by orbiting Earth satellites has permitted scientists to monitor weather movements and to gather global land use data. In addition, computer-based image-processing techniques are employed in converting old motion pictures in black and white into colour and in matching fingerprints taken from the scene of a crime with those in fingerprint archives.

Appli-
cations of
pattern
recog-
nition

**Robotics.** An application of artificial intelligence of growing commercial importance is industrial automation in the form of computer-controlled robots. The first practical industrial robot, designed to remove parts from a die-casting machine, was introduced during the 1960s. By the mid-1980s, many thousands of computerized robotic devices had come into use, principally in automobile and aircraft assembly plants. The majority of these industrial robots are mechanical manipulators, single-arm, stationary devices that can be programmed to carry out rather simple, repetitive assembly-line jobs (*e.g.*, welding, riveting, drilling, and spray painting). Much current AI research in robotics is aimed at developing more versatile units—mobile robots equipped with electronic vision systems and tactile sensors (for mechanical "hands"). Such devices would be capable of performing more complicated tasks and ideally those too difficult or too dangerous for humans, as for example transporting objects weighing many tons around a factory, accurately measuring machine components by touch, or handling radioactive materials and other highly toxic substances for extended periods of time.

Mechanical
manip-
ulators

BIBLIOGRAPHY. H. DOMINIC COVVEY and NEIL HARDING MCALISTER, *Computer Consciousness: Surviving the Automated 80s* (1980, reissued 1982), a popular introduction to computer operation, selection, and terminology; ALFRED BORK, *Learning with Computers* (1981), an introduction to the use of computers in education; WALTER M. MATHEWS (ed.), *Monster or Messiah?: The Computer's Impact on Society* (1980), essays by scholars in the humanities; DAVID H. FLAHERTY, *Privacy and Government Data Banks: An International Perspective* (1979), a study of data collection and dissemination by governments in the United States, the United Kingdom, Canada, West Germany, and Sweden; ANDREW E. WESSEL, *The Social Use of Information—Ownership and Access* (1976), an argument for greater public access to information retrieval systems; TOM FORESTER (ed.), *The Microelectronics Revolution* (1980, reissued 1981), a collection of essays exploring the impact of microcomputers on society; EDWARD A. FEIGENBAUM and PAMELA MCCORDUCK, *The Fifth Generation,* rev. ed. (1984), a discussion of artificial intelligence and the Japanese efforts to produce fifth-generation computers; PAMELA MCCORDUCK, *Machines Who Think* (1979), a survey of attempts to create artificial intelligence. HERMAN H. GOLDSTINE, *The Computer from Pascal to von Neumann* (1972, reprinted 1980), a history of computers and a personal account by one of the pioneers; *A History of Computing in the Twentieth Century,* ed. by N. METROPOLIS, J. HOWLETT, and GIAN-CARLO ROTA (1980), conference proceedings containing firsthand accounts of the history and development of modern computers; KATHERINE DAVIS FISHMAN, *The Computer Establishment* (1981, reissued 1982), a history of the computer industry from 1950 to 1980; GRANINO A. KORN and THERESA M. KORN, *Electronic Analog and Hybrid Computers,* 2nd rev. ed. (1972), which discusses the uses of and circuit requirements for analog and hybrid computers; RAYMOND M. KLINE, *Digital Computer Design* (1977), a thorough treatment for the reader with a knowledge of electronics; IVAN TOMEK, *Introduction to Computer Organization* (1981), an introduction to logic design and structure of major computer components—beneficial for the beginning student; MORTON H. LEWIN, *Logic Design and Computer Organization* (1983), a text for students of design; G.W. GORSLINE, *Computer Organization: Hardware/Software,* 2nd ed. (1986), an introductory discussion of computer organization, software, and practical applications; *History of Programming Languages,* ed. by RICHARD L. WEXELBLAT (1981), conference proceedings containing papers detailing the historical development of several major computer programming languages; ELLIS HOROWITZ, *Fundamentals of Programming Languages,* 2nd ed. (1984), a thorough treatment of programming language issues. For reference, consult CHARLES J. SIPPL and ROGER J. SIPPL, *Computer Dictionary and Handbook,* 3rd ed. (1980); *Encyclopedia of Computer Science and Engineering,* 2nd ed., edited by ANTHONY RALSTON (1983); and *Information Industry Market Place: An International Directory of Information Products and Services* (annual), published as *Information Trade Directory* outside North and South America. Useful information can be found in ROBERT A. STERN and NANCY STERN, *An Introduction to Computers and Information Processing,* 2nd ed. (1985).

(A.M.P./N.E.W./S.F.We.)

# Confucius and Confucianism

Confucianism, a Western term that has no counterpart in Chinese, is a world view, a social ethic, a political ideology, a scholarly tradition, and a way of life. Sometimes viewed as a philosophy and sometimes as a religion, Confucianism may be understood as an all-encompassing humanism that neither denies nor slights Heaven. East Asians may profess themselves to be Shintōists, Taoists, Buddhists, Muslims, or Christians, but, by announcing their religious affiliations, seldom do they cease to be Confucians.

Although often grouped with the major historical religions, Confucianism differs from them by not being an organized religion. Nonetheless, it spread to all East Asian countries under the influence of Chinese literate culture and exerted a profound influence on East Asian spiritual life as well as on East Asian political culture. Both the theory and practice of Confucianism have indelibly marked the patterns of government, society, education, and family of East Asia. It is an exaggeration to characterize traditional Chinese life and culture as Confucian, but Confucian ethical values have for well over 2,000 years served as the source of inspiration as well as the court of appeal for human interaction between individuals, communities, and nations in the Sinitic world.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 825.

This article is divided into the following sections:


Confucius 653
  The life and thought of Confucius 653
    The historical context
    The life of Confucius
    The *Analects* as the embodiment of Confucian ideas
Confucianism 656
  Formation of the classical Confucian tradition 656
    Mencius: The paradigmatic Confucian intellectual
    Hsün-tzu: The transmitter of Confucian scholarship
    The Confucianization of politics
    The Five Classics
    Tung Chung-shu: The Confucian visionary
    Confucian ethics in the Taoist and Buddhist context
  The Confucian revival 659
    The Sung masters
    Confucian learning in Chin, Yüan, and Ming
    The age of Confucianism: Yi-dynasty Korea, Tokugawa Japan, and Ch'ing China
  Modern transformation 661
Bibliography 662


## Confucius

### THE LIFE AND THOUGHT OF CONFUCIUS

The story of Confucianism does not begin with Confucius (Latinized form of K'ung-fu-tzu, Master K'ung; 551–479 BC). Nor was Confucius the founder of Confucianism in the sense that Buddha was the founder of Buddhism and Christ the founder of Christianity. Rather Confucius considered himself a transmitter who consciously tried to reanimate the old in order to attain the new. He proposed retrieving the meaning of the past by breathing vitality into seemingly outmoded rituals. Confucius' love of antiquity was motivated by his strong desire to understand why certain rituals, such as the ancestral cult, reverence for Heaven, and mourning ceremonies, had survived for centuries. His journey into the past was a search for roots, which he perceived as grounded in humanity's deepest needs for belonging and communicating. He had faith in the cumulative power of culture. The fact that traditional ways had lost vitality did not, for him, diminish their potential for regeneration in the future. In fact, Confucius' sense of history was so strong that he saw himself as a conservationist responsible for the continuity of the cultural values and the social norms that had worked so well for the civilization of the Chou dynasty.

By courtesy of the Museum of Fine Arts, Boston, Weld-Fenollosa Collection



Confucius, screen painting by Kanō Tanyū, 17th century. In the Museum of Fine Arts, Boston.

**The historical context.** The scholarly tradition envisioned by Confucius can be traced to the sage-kings of antiquity. Although the earliest dynasty confirmed by archaeology is the Shang dynasty (18th–12th century BC), the historical period that Confucius claimed as relevant was much earlier. Confucius may have initiated a cultural process known in the West as Confucianism, but he and those who followed him considered themselves part of a tradition, later identified by Chinese historians as the *ju-chia,* "scholarly tradition," that had its origins two millennia previously, when the legendary Yao and Shun created a civilized world through moral persuasion.

The "feudal" ritual system

Confucius' hero was Chou Kung, or the Duke of Chou (d. 1094 BC), who was said to have helped consolidate and refine the "feudal" ritual system. This system was based on blood ties, marriage alliances, and old covenants as well as on newly negotiated contracts and was an elaborate system of mutual dependence. The appeal to cultural values and social norms for the maintenance of interstate as well as domestic order was predicated on a shared political vision, namely, that authority lies in universal kingship, heavily invested with ethical and religious power by the mandate of Heaven, and that social solidarity is achieved not by legal constraint but by ritual observance. Its implementation enabled the Chou dynasty to survive in relative peace and prosperity for more than five centuries.

Inspired by the statesmanship of Chou Kung, Confucius harboured a lifelong dream to be in a position to emulate the duke by putting into practice the political ideas that he had learned from the ancient sages and worthies. Although Confucius never realized his political dream, his conception of politics as moral persuasion became more and more influential.

Heaven and the mandate of Heaven

The idea of Heaven, unique in Chou cosmology, was compatible with the concept of the Lord-on-High in the Shang dynasty. The Lord-on-High may have referred to the progenitor of the Shang royal lineage so that the Shang kings could claim their position as divine descendants, as the emperors of Japan later did, but Heaven to the Chou kings was a much more generalized anthropomorphic God. They believed that the mandate of Heaven (the functional equivalent of the will of the Lord-on-High) was not constant and that there was no guarantee that the descendants of the Chou royal house would be entrusted with kingship, for "Heaven sees as the people see and Heaven hears as the people hear"; thus the virtues of the kings were essential for the maintenance of their power and authority. This emphasis on benevolent rulership, expressed in numerous bronze inscriptions, was both a reaction to the collapse of the Shang dynasty and an affirmation of a deep-rooted world view.

Partly because of the vitality of the feudal ritual system and partly because of the strength of the royal household itself, the Chou kings were able to control their kingdom for several centuries. In 771 BC, however, they were forced to move their capital eastward to present-day Lo-yang to avoid barbarian attacks from Central Asia. Real power thereafter passed into the hands of feudal lords. Since the surviving line of the Chou kings continued to be recognized in name, they still managed to exercise some measure of symbolic control. By Confucius' time, however, the feudal ritual system had been so fundamentally undermined that the political crises also precipitated a profound sense of moral decline: the centre of symbolic control could no longer hold the kingdom from total disintegration.

Confucius' response was to address himself to the issue of learning to be human. In so doing he attempted to redefine and revitalize the institutions that for centuries had been vital to political stability and social order: the family, the school, the local community, the state, and the kingdom. Confucius did not accept the status quo, which held that wealth and power spoke the loudest. He felt that virtue, both as a personal quality and as a requirement for leadership, was essential for individual dignity, communal solidarity, and political order.

**The life of Confucius.** Confucius' life, in contrast to his tremendous importance, seems starkly undramatic, or, as a Chinese expression has it, it seems "plain and real." The plainness and reality of Confucius' life, however, underlines that his humanity was not revealed truth but an expression of self-cultivation, of the ability of human effort to shape its own destiny. The faith in the possibility of ordinary human beings to become awe-inspiring sages and worthies is deeply rooted in the Confucian heritage, and the insistence that human beings are teachable, improvable, and perfectible through personal and communal endeavour is typically Confucian.

The perfectibility of man

Although the facts about Confucius' life are scanty, they do establish a precise time frame and historical context. Confucius was born in the 22nd year of the reign of Duke Hsiang of Lu (551 BC). The traditional claim that he was born on the 27th day of the eighth lunar month has been questioned by historians, but September 28 is still widely observed in East Asia as Confucius' birthday. It is an official holiday, "Teachers' Day," in Taiwan.

Confucius was born in Ch'ü-fu in the small feudal state of Lu in what is now Shantung Province, which was noted for its preservation of the traditions of ritual and music of the Chou civilization. His family name was K'ung and his personal name Ch'iu, but he is referred to as either K'ung-tzu or K'ung-fu-tzu (Master K'ung) throughout Chinese history. The adjectival "Confucian," derived from the Latinized Confucius, is not a meaningful term in Chinese, nor is the term Confucianism, which was coined in Europe as recently as the 18th century.

Confucius' ancestors were probably members of the aristocracy who had become virtual poverty-stricken commoners by the time of his birth. His father died when Confucius was only three years old. Instructed first by his mother, Confucius then distinguished himself as an indefatigable learner in his teens. He recalled toward the end of his life that at age 15 his heart was set upon learning. A historical account notes that, even though he was already known as an informed young scholar, he felt it appropriate to inquire about everything while visiting the Grand Temple.

Confucius had served in minor government posts managing stables and keeping books for granaries before he married a woman of similar background when he was 19. It is not known who Confucius' teachers were, but he made a conscientious effort to find the right masters to teach him, among other things, ritual and music. Confucius' mastery of the six arts—ritual, music, archery, charioteering, calligraphy, and arithmetic—and his familiarity with the classical traditions, notably poetry and history, enabled him to start a brilliant teaching career in his 30s.

The purposes of learning and teaching

Confucius is known as the first teacher in China who wanted to make education available to all men and who was instrumental in establishing the art of teaching as a vocation, indeed as a way of life. Before Confucius, aristocratic families had hired tutors to educate their sons in specific arts, and government officials had instructed their subordinates in the necessary techniques, but he was the first person to devote his whole life to learning and teaching for the purpose of transforming and improving society. He believed that all human beings could benefit from self-cultivation. He inaugurated a humanities program for potential leaders, opened the doors of education to all, and defined learning not merely as the acquisition of knowledge but also as character building.

For Confucius the primary function of education was to provide the proper way of training noblemen (*chün-tzu*), a process that involved constant self-improvement and continuous social interaction. Although he emphatically noted that learning was "for the sake of the self" (the end of which was self-knowledge and self-realization), he found public service a natural consequence of true education. Confucius confronted learned hermits who challenged the validity of his desire to serve the world; he resisted the temptation to "herd with birds and animals," to live apart from the human community, and opted to try to transform the world from within. For decades Confucius was actively involved in politics, wishing to put his humanist ideas into practice through governmental channels.

Political career and voluntary exile

In his late 40s and early 50s Confucius served first as a magistrate, then as an assistant minister of public works, and eventually as minister of justice in the state of Lu. It is likely that he accompanied King Lu as his chief minister on one of the diplomatic missions. Confucius' political career was, however, short-lived. His loyalty to the King alienated him from the power holders of the time, the large Chi families, and his moral rectitude did not sit well with the King's inner circle, who enraptured the King with sensuous delight. At 56, when he realized that his superiors were uninterested in his policies, Confucius left the country in an attempt to find another feudal state to which he could render his service. Despite his political frustration he was accompanied by an expanding circle of students during this self-imposed exile of almost 12 years. His reputation as a man of vision and mission spread. A guardian of a border post once characterized him as the "wooden tongue for a bell" of the age, sounding Heaven's prophetic note to awaken the people (*Analects,* 3:24). Indeed, Confucius was perceived as the heroic conscience who knew realistically that he might not succeed but, fired by a righteous passion, continuously did the best he could. At the age of 67 he returned home to teach and to preserve his cherished classical traditions by writing and editing. He died in 479 BC at the age of 73. According to the *Records of the Historian* 72 of his students mastered the "six arts," and those who claimed to be his followers numbered 3,000.

**The Analects as the embodiment of Confucian ideas.** The *Lun-yü* (*Analects*), the most revered sacred scripture in the Confucian tradition, was probably compiled by the second generation of Confucius' disciples. Based primarily on the Master's sayings, preserved in both oral and written transmissions, it captures the Confucian spirit in form and content in the same way that the Platonic dialogues embody Socratic pedagogy.

The *Analects* has often been viewed by the critical mod-

ern reader as a collection of unrelated conversations randomly put together. This impression may have resulted from the mistaken conception of Confucius as a mere commonsense moralizer who gave practical advice to students in everyday situations. If a person approaches the *Analects* as a communal memory, a literary device on the part of those who considered themselves beneficiaries of the Confucian Way to continue the Master's memory and to transmit his form of life as a living tradition, he comes close to what it has been revered for in China for centuries. Dialogues are used to show Confucius in thought and action, not as an isolated individual but as the centre of relationships. Actually the sayings of the *Analects* reveal Confucius' personality—his ambitions, his fears, his joys, his commitments, and above all his self-knowledge.

The purpose of the *Analects*

The purpose, then, in compiling these distilled statements centring on Confucius seems not to have been to present an argument or to record an event but to offer an invitation to readers to take part in an ongoing conversation. Through the *Analects* Confucians for centuries learned to reenact the awe-inspiring ritual of participating in a conversation with Confucius.

One of Confucius' most significant personal descriptions is the short autobiographical account of his spiritual development found in the *Analects:*

> At 15 I set my heart on learning; at 30 I firmly took my stand; at 40 I had no delusions; at 50 I knew the Mandate of Heaven; at 60 my ear was attuned; at 70 I followed my heart's desire without overstepping the boundaries of right. (2:4)

Confucius' life as a student and teacher exemplified his idea that education was a ceaseless process of self-realization. When one of his students reportedly had difficulty describing him, Confucius came to his aid:

> Why did you not simply say something to this effect: he is the sort of man who forgets to eat when he engages himself in vigorous pursuit of learning, who is so full of joy that he forgets his worries, and who does not notice that old age is coming on? (7:18)

Confucius was deeply concerned that the culture (*wen*) he cherished was not being transmitted and that the learning (*hsüeh*) he propounded was not being taught. His strong sense of mission, however, never interfered with his ability to remember what had been imparted to him, to learn without flagging, and to teach without growing weary. What he demanded of himself was strenuous:

> It is these things that cause me concern: failure to cultivate virtue, failure to go deeply into what I have learned, inability to move up to what I have heard to be right, and inability to reform myself when I have defects. (7:3)

What he demanded of his students was the willingness to learn: "I do not enlighten anyone who is not eager to learn, nor encourage anyone who is not anxious to put his ideas into words (7:8).

The Confucian community

The community that Confucius created was a scholarly fellowship of like-minded men of different ages and different backgrounds from different states. They were attracted to Confucius because they shared his vision and to varying degrees took part in his mission to bring moral order to an increasingly fragmented polity. This mission was difficult and even dangerous. Confucius himself suffered from joblessness, homelessness, starvation, and occasionally life-threatening violence. Yet his faith in the survivability of the culture that he cherished and the workability of the approach to teaching that he propounded was so steadfast that he convinced his followers as well as himself that Heaven was on their side. When Confucius' life was threatened in K'uang, he said:

> Since the death of King Wen [founder of the Chou dynasty] does not the mission of culture (*wen*) rest here in me? If Heaven intends this culture to be destroyed, those who come after me will not be able to have any part of it. If Heaven does not intend this culture to be destroyed, then what can the men of K'uang do to me? (9:5)

This expression of self-confidence informed by a powerful sense of mission may give the impression that there was presumptuousness in Confucius' self-image. Confucius, however, made it explicit that he was far from attaining sagehood and that all he really excelled in was "love of learning" (5:27). To him, learning not only broadened his knowledge and deepened his self-awareness but also defined who he was. He frankly admitted that he was not born endowed with knowledge, nor did he belong to the class of men who could transform society without knowledge. Rather, he reported that he used his ears widely and followed what was good in what he had heard and used his eyes widely and retained in his mind what he had seen. His learning constituted "a lower level of knowledge" (7:27), a level that was presumably accessible to the majority of human beings. In this sense Confucius was neither a prophet with privileged access to the divine nor a philosopher who had already seen the truth but a teacher of humanity who was also an advanced fellow traveler on the way to self-realization.

As a teacher of humanity Confucius stated his ambition in terms of concern for human beings: "To bring comfort to the old, to have trust in friends, and to cherish the young" (5:25). Confucius' vision of the way to develop a moral community began with a holistic reflection on the human condition. Instead of dwelling on abstract speculations such as man's condition in the state of nature, Confucius sought to understand the actual situation of a given time and to use that as his point of departure. His aim was to restore trust in government and to transform society into a moral community by cultivating a sense of humanity in politics and society. To achieve that aim, the creation of a scholarly community, the fellowship of *chün-tzu* (noblemen), was essential. In the words of Confucius' disciple Tseng-tzu, the true nobleman

The fellowship of *chün-tzu*

> must be broad-minded and resolute, for his burden is heavy and his road is long. He takes humanity as his burden. Is that not heavy? Only with death does his road come to an end. Is that not long? (8:7)

The fellowship of *chün-tzu* as moral vanguards of society, however, did not seek to establish a radically different order. Its mission was to redefine and revitalize those institutions that for centuries were believed to have maintained social solidarity and enabled people to live in harmony and prosperity. An obvious example of such an institution was the family.

It is related in the *Analects* that Confucius, when asked why he did not take part in government, responded by citing a passage from an ancient classic, the *Shu Ching* ("Classic of History"), "Simply by being a good son and friendly to his brothers a man can exert an influence upon government!" to show that what a person does in the confines of his home is politically significant (2:21). This maxim is based on the Confucian conviction that cultivation of the self is the root of social order and that social order is the basis for political stability and universal peace.

The assertion that family ethics is politically efficacious must be seen in the context of the Confucian conception of politics as "rectification" (*cheng*). Rulers should begin by rectifying their own conduct; that is, they are to be examples who govern by moral leadership and exemplary teaching rather than by force. Government's responsibility is not only to provide food and security but also to educate the people. Law and punishment are the minimum requirements for order; the higher goal of social harmony, however, can only be attained by virtue expressed through ritual performance. To perform rituals, then, is to take part in a communal act to promote mutual understanding.

The importance of filial piety

One of the fundamental Confucian values that ensures the integrity of ritual performance is *hsiao* (filial piety). Indeed, Confucius saw filial piety as the first step toward moral excellence, which he believed lay in the attainment of the cardinal virtue, *jen* (humanity). To learn to embody the family in the mind and heart is to become able to move beyond self-centredness or, to borrow from modern psychology, to transform the enclosed private ego into an open self. Filial piety, however, does not demand unconditional submissiveness to parental authority but recognition of and reverence for the source of life. The purpose of filial piety, as the ancient Greeks expressed it, is to enable both parent and child to flourish. Confucians see it as an essential way of learning to be human.

Confucians, moreover, are fond of applying the family metaphor to the community, the country, and the universe. They prefer to address the emperor as the son of

Heaven, the king as ruler-father, and the magistrate as the "father-mother official" because to them the family-centred nomenclature implies a political vision. When Confucius said that taking care of family affairs is itself active participation in politics, he had already made it clear that family ethics is not merely a private concern; the public good is realized by and through it.

Confucius defined the process of becoming human as being able to "conquer yourself and return to ritual" (12:1). The dual focus on the transformation of the self (Confucius is said to have freed himself from four things: "opinionatedness, dogmatism, obstinacy, and egoism" [9:4]) and on social participation enabled Confucius to be loyal (*chung*) to himself and considerate (*shu*) of others (4:15). It is easy to understand why the Confucian "golden rule" is "Do not do unto others what you would not want others to do unto you!" (15:23). Confucius' legacy, laden with profound ethical implications, is captured by his "plain and real" appreciation that learning to be human is a communal enterprise:

> A man of humanity, wishing to establish himself, also establishes others, and wishing to enlarge himself, also enlarges others. The ability to take as analogy of what is near at hand can be called the method of humanity. (6:30)

## Confucianism

### FORMATION OF THE CLASSICAL CONFUCIAN TRADITION

According to Han-fei-tzu (d. 233 BC), shortly after Confucius' death his followers split into eight distinct schools, all claiming to be the legitimate heir to the Confucian legacy. Presumably each school was associated with or inspired by one or more of Confucius' disciples. Yet the Confucians did not exert much influence in the 5th century BC. Although the mystic Yen Yüan (or Yen Hui), the faithful Tseng-tzu, the talented Tzu Kung, the erudite Tzu-hsia, and others may have generated a great deal of enthusiasm among the second generation of Confucius' students, it was not at all clear at the time that the Confucian tradition was to emerge as the most powerful one in Chinese history.

Mencius (*c.* 371–*c.* 289 BC) complained that the world of thought in the early Warring States period (475–221 BC) was dominated by the collectivism of Mo-tzu and the individualism of Yang Chu (440–*c.* 360 BC). The historical situation a century after Confucius' death clearly shows that the Confucian attempt to moralize politics was not working; the disintegration of the Chou feudal ritual system and the rise of powerful hegemonic states reveal that wealth and power spoke the loudest. The hermits (the early Taoists), who left the world to create a sanctuary in nature in order to lead a contemplative life, and the realists (proto-Legalists), who played the dangerous game of assisting ambitious kings to gain wealth and power so that they could influence the political process, were actually determining the intellectual agenda. The Confucians refused to be identified with the interests of the ruling minority because their social consciousness impelled them to serve as the conscience of the people. They were in a dilemma. Although they wanted to be actively involved in politics, they could not accept the status quo as the legitimate arena in which to exercise authority and power. In short, they were in the world but not of it; they could not leave the world, nor could they effectively change it.

**Mencius: The paradigmatic Confucian intellectual.** Mencius is known as the self-styled transmitter of the Confucian Way. Educated first by his mother and then allegedly by a student of Confucius' grandson, Mencius brilliantly performed his role as a social critic, a moral philosopher, and a political activist. He argued that cultivating a class of scholar-officials who would not be directly involved in agriculture, industry, and commerce was vital to the well-being of the state. In his sophisticated argument against the physiocrats (those who advocated the supremacy of agriculture), he intelligently employed the idea of the division of labour to defend those who labour with their minds, observing that service is as important as productivity. To him Confucians served the vital interests of the state as scholars not by becoming bureaucratic functionaries but by assuming the responsibility of teaching the ruling minority humane government (*jen-cheng*) and the kingly way (*wang-tao*). In dealing with feudal lords, Mencius conducted himself not merely as a political adviser but also as a teacher of kings. Mencius made it explicit that a true man cannot be corrupted by wealth, subdued by power, or affected by poverty.

Mencius' defense of scholar-officials

To articulate the relationship between Confucian moral idealism and the concrete social and political realities of his time, Mencius began by exposing as impractical the prevailing ideologies of Mo-tzu's collectivism and Yang Chu's individualism. Mo-tzu's collectivism rested on the advocacy of universal love. Mencius contended, however, that the result of the Mohist admonition to treat a stranger as intimately as one's own father would be to treat one's own father as indifferently as one would treat a stranger. Yang Chu, on the other hand, advocated the primacy of the self. Mencius contended, however, that excessive attention to self-interest would lead to political disorder. Indeed, in Mohist collectivism fatherhood becomes a meaningless concept, and so does kingship in Yang Chu's individualism.

Mencius' strategy for social reform was to change the language of profit, self-interest, wealth, and power by making it part of a moral discourse, with emphasis on rightness, public-spiritedness, welfare, and influence. Mencius, however, was not arguing against profit. Rather, he instructed the feudal lords to look beyond the narrow horizon of their palaces and to cultivate a common bond with their ministers, officers, clerks, and the seemingly undifferentiated masses. Only then, Mencius contended, would they be able to preserve their profit, self-interest, wealth, and power. He encouraged them to extend their benevolence and warned them that this was crucial for the protection of their families.

Mencius' appeal to the common bond among all people as a mechanism of government was predicated on his strong "populist" sense that the people are more important than the state and the state more important than the king and that the ruler who does not act in accordance with the kingly way is unfit to rule. Mencius insisted that an unfit ruler should be criticized, rehabilitated, or, as the last resort, deposed. Since "Heaven sees as the people see; Heaven hears as the people hear," revolution, or literally the change of the mandate, in severe cases is not only justifiable but is a moral imperative.

Mencius' "populist" conception of politics

Mencius' "populist" conception of politics was predicated on his philosophical vision that human beings can perfect themselves through effort and that human nature is good. While he acknowledged the role of biological and environmental factors in shaping the human condition, he insisted that human beings become moral simply by willing to be so. According to Mencius, willing entails the transformative moral act insofar as the propensity of humans to be good is automatically activated whenever they decide to bring it to their conscious attention.

Mencius' moral idealism

Mencius taught that all people have the spiritual resources to deepen their self-awareness and strengthen their bonds with others. Biologic and environmental constraints notwithstanding, men always have the freedom and the ability to refine and enlarge their Heaven-endowed nobility (their "great body"). The possibility of continuously refining and enlarging the self is vividly illustrated in Mencius' description of degrees of excellence:

> He who commands our liking is called good (*shan*).
> He who is sincere with himself is called true (*hsin*).
> He who is sufficient and real is called beautiful (*mei*).
> He whose sufficiency and reality shine forth is called great (*ta*).
> He whose greatness transforms itself is called sagely (*sheng*).
> He whose sageliness is beyond our comprehension is called spiritual (*shen*). (VIIB:25)

Furthermore, Mencius asserted that if men fully realize the potential of their hearts, they will understand their nature; by understanding their nature, they will know Heaven. Learning to be fully human, in this Mencian perspective, entails the cultivation of human sensitivity to embody the whole universe as one's lived experience:

> All the 10,000 things are there in me. There is no greater joy for me than to find, on self-examination, that I am true to

myself. Try your best to treat others as you would wish to be treated yourself, and you will find that this is the shortest way to humanity. (VIIA:4)

**Hsün-tzu: The transmitter of Confucian scholarship.** If Mencius brought Confucian moral idealism to fruition, Hsün-tzu (*c.* 300–*c.* 230 BC) conscientiously transformed Confucianism into a realistic and systematic inquiry on the human condition, with special reference to ritual and authority. Widely acknowledged as the most eminent of the notable scholars who congregated in Chi-hsia, the capital of the wealthy and powerful Ch'i state in the mid-3rd century BC, Hsün-tzu distinguished himself in erudition, logic, empiricism, practical-mindedness, and argumentation. His critique of the so-called 12 philosophers gave an overview of the intellectual life of his time. His penetrating insight into the shortcomings of virtually all the major currents of thought propounded by his fellow thinkers helped to establish the Confucian school as a forceful political and social persuasion. His principal adversary, however, was Mencius, and he vigorously attacked Mencius' view that human nature is good as naive moral optimism.

True to the Confucian and, for that matter, Mencian spirit, Hsün-tzu underscored the centrality of self-cultivation. He defined the process of Confucian education, from nobleman to sage, as a ceaseless endeavour to accumulate knowledge, skills, insight, and wisdom. In contrast to Mencius, Hsün-tzu stressed that human nature is evil. Because he saw human beings as prone by nature to pursue the gratification of their passions, he firmly believed in the need for clearly articulated social constraints. Without constraints, social solidarity, the precondition for human well-being, would be undermined. The most serious flaw he perceived in the Mencian commitment to the goodness of human nature was the practical consequence of neglecting the necessity of ritual and authority for the well-being of society.

Hsün-tzu's moral pessimism

Hsün-tzu singled out the cognitive function of the mind (human rationality) as the basis for morality. Men become moral by voluntarily harnessing their desires and passions to act in accordance with society's norms. Although this is alien to human nature, it is perceived by the mind as necessary for both survival and well-being. Like Mencius, Hsün-tzu believed in the perfectibility of all human beings through self-cultivation, in humanity and rightness as cardinal virtues, in humane government as the kingly way, in social harmony, and in education. But his view of how these could actually be achieved was diametrically opposed to that of Mencius. The Confucian project, as shaped by Hsün-tzu, defines learning as socialization. The authority of ancient sages and worthies, the classical tradition, conventional norms, teachers, governmental rules and regulations, and political officers are all important for this process. A cultured person is by definition a fully socialized member of the human community, who has successfully sublimated his instinctual demands for the public good.

Hsün-tzu's tough-minded stance on law, order, authority, and ritual seems precariously close to that of the Legalists, whose policy of social conformism was designed exclusively for the benefit of the ruler. His insistence on objective standards of behaviour may have ideologically contributed to the rise of authoritarianism, which resulted in the dictatorship of the Ch'in (221–206 BC). As a matter of fact, two of the most influential Legalists, the theoretician Han-fei-tzu from the state of Han and the Ch'in minister Li Ssu (*c.* 280–208 BC), were his pupils. Yet Hsün-tzu was instrumental in the continuation of Confucianism as a scholarly enterprise. His naturalistic interpretation of Heaven, his sophisticated understanding of culture, his insightful observations on the epistemological aspect of the mind and social function of language, his emphasis on moral reasoning and the art of argumentation, his belief in progress, and his interest in political institutions so significantly enriched the Confucian heritage that he was revered by the Confucians as the paradigmatic scholar for more than three centuries.

The paradigmatic Confucian scholar

**The Confucianization of politics.** The short-lived dictatorship of the Ch'in marked a brief triumph of Legalism. In the early years of the Western Han (206 BC–AD 25), however, the Legalist practice of absolute power of the emperor, complete subjugation of the peripheral states to the central government, total uniformity of thought, and ruthless enforcement of law were replaced by the Taoist practice of reconciliation and noninterference. This practice is commonly known in history as the Huang-Lao method, referring to the art of rulership attributed to the Yellow Emperor (Huang-ti) and the mysterious founder of Taoism, Lao-tzu. Although a few Confucian thinkers, such as Lu Chia and Chia I, made important policy recommendations, Confucianism before the emergence of Tung Chung-shu (*c.* 179–*c.* 104 BC) was not particularly influential. Nonetheless, the gradual Confucianization of Han politics began soon after the founding of the dynasty.

By the reign of Wu Ti (the Martial Emperor, 141–87 BC), who was by temperament a Legalist despot, Confucianism was deeply entrenched in the central bureaucracy. It was manifest in such practices as the clear separation of the court and the government, often under the leadership of a scholarly prime minister, the process of recruiting officials through the dual mechanism of recommendation and selection, the family-centred social structure, the agriculture-based economy, and the educational network. Confucian ideas were also firmly established in the legal system as ritual became increasingly important in governing behaviour, defining social relationships, and adjudicating civil disputes. Yet it was not until the prime minister Kung-sun Hung (d. 121 BC) had persuaded Wu Ti to announce formally that the *ju* school alone would receive state sponsorship that Confucianism became an officially recognized Imperial ideology and state cult.

Confucianism shaping bureaucratic procedures

As a result Confucian Classics became the core curriculum for all levels of education. In 136 BC Wu Ti set up at court five Erudites of the Five Classics (see below *The Five Classics*) and in 124 BC assigned 50 official students to study with them, thus creating a de facto Imperial university. By 50 BC enrollment at the university had grown to an impressive 3,000, and by AD 1 a hundred men a year were entering government service through the examinations administered by the state. In short, those with a Confucian education began to staff the bureaucracy. In the year 58 all government schools were required to make sacrifices to Confucius, and in 175 the court had the approved version of the Classics, which had been determined by scholarly conferences and research groups under Imperial auspices for several decades, carved on large stone tablets. (These stelae, which were erected at the capital, are today well preserved in the museum of Sian.) This act of committing to permanence and to public display the content of the sacred scriptures symbolized the completion of the formation of the classical Confucian tradition.

**The Five Classics.** The compilation of the *Wu Ching* (The Five Classics) was a concrete manifestation of the coming of age of the Confucian tradition. The inclusion of both pre-Confucian texts, the *Shu Ching* ("Classic of History") and the *Shih ching* ("Classic of Poetry"), and contemporary Ch'in-Han material, such as certain portions of the *Li chi* ("Record of Rites"), suggests that the spirit behind the establishment of the core curriculum for Confucian education was ecumenical. The Five Classics can be described in terms of five visions: metaphysical, political, poetic, social, and historical.

The metaphysical vision, expressed in the *I Ching* ("Classic of Changes"), combines divinatory art with numerological technique and ethical insight. According to the philosophy of change, the cosmos is a great transformation occasioned by the constant interaction of two complementary as well as conflicting vital energies, yin and yang. The universe, which resulted from this great transformation, always exhibits both organismic unity and dynamism. The nobleman, inspired by the harmony and creativity of the universe, must emulate this pattern by aiming to realize the highest ideal of "unity of man and Heaven" through ceaseless self-exertion.

Metaphysical, political, poetic, social, and historical visions

The political vision, contained in the *Shu Ching,* presents kingship in terms of the ethical foundation for a humane government. The legendary Three Emperors (Yao, Shun, and Yü) all ruled by virtue. Their sagacity, *hsiao* (filial piety), and dedication to work enabled them to create a

political culture based on responsibility and trust. Their exemplary lives taught and encouraged the people to enter into a covenant with them so that social harmony could be achieved without punishment or coercion. Even in the Three Dynasties (Hsia, Shang, and Chou) moral authority, as expressed through ritual, was sufficient to maintain political order. The human continuum, from the undifferentiated masses to the enlightened people, the nobility, and the sage-king, formed an organic unity as an integral part of the great cosmic transformation. Politics means moral persuasion, and the purpose of the government is not only to provide food and maintain order but also to educate.

The poetic vision, contained in the *Shih ching,* underscores the Confucian valuation of common human feelings. The majority of verses give voice to emotions and sentiments of communities and persons from all levels of society expressed on a variety of occasions. The basic theme of this poetic world is mutual responsiveness. The tone as a whole is honest rather than earnest and evocative rather than expressive.

The social vision, contained in the *Li chi,* shows society not as an adversarial system based on contractual relationships but as a community of trust with emphasis on communication. Society organized by the four functional occupations—the scholar, farmer, artisan, and merchant—is, in the true sense of the word, a cooperation. As a contributing member of the cooperation each person is obligated to recognize the existence of others and to serve the public good. It is the king's duty to act kingly and the father's duty to act fatherly. If the king or father fails to behave properly, he cannot expect his minister or son to act in accordance with ritual. It is in this sense that a chapter in the *Li chi* entitled the "Great Learning" specifies, "From the Son of Heaven to the commoner, all must regard self-cultivation as the root." This pervasive consciousness of duty features prominently in all Confucian literature on ritual.

The historical vision, presented in the *Ch'un-ch'iu* ("Spring and Autumn Annals"), emphasizes the significance of collective memory for communal self-identification. Historical consciousness is a defining characteristic of Confucian thought. By defining himself as a lover of antiquity and a transmitter of its values, Confucius made it explicit that a sense of history is not only desirable but is necessary for self-knowledge. Confucius' emphasis on the importance of history was in a way his reappropriation of the ancient Sinitic wisdom that reanimating the old is the best way to attain the new. Confucius may not have been the author of the *Ch'un-ch'iu,* but it seems likely that he applied moral judgment to political events in China proper from the 8th to the 5th century BC. In this unprecedented procedure he assumed a godlike role in evaluating politics by assigning ultimate historical praise and blame to the most powerful and influential political actors of the period. This practice inspired not only the innovative style of the grand historian Ssu-ma Ch'ien (*c.* 145–*c.* 85 BC), but it was also widely employed by others writing dynastic histories in Imperial China.

**Tung Chung-shu: The Confucian visionary.** Like Ssu-ma Ch'ien, Tung Chung-shu (*c.* 179–*c.* 104 BC) also took the *Ch'un-ch'iu* absolutely seriously. His own work, *Ch'un-ch'iu fan-lu* ("Luxuriant Gems of the Spring and Autumn Annals"), however, is far from being a book of historical judgment. It is a metaphysical treatise in the spirit of the *I Ching.* A man extraordinarily dedicated to learning (he is said to have been so absorbed in his studies that for three years he did not even glance at the garden in front of him) and strongly committed to moral idealism (one of his often-quoted dicta is "rectifying rightness without scheming for profit; enlightening his Way without calculating efficaciousness"), Tung was instrumental in developing a characteristically Han interpretation of Confucianism.

Intellectual syncretism of the Han cultural elite

Despite Wu Ti's pronouncement that Confucianism alone would receive Imperial sponsorship, Taoists, yin-yang cosmologists, Mohists, Legalists, shamanists, practitioners of seances, healers, magicians, geomancers, and others all contributed to the cosmological thinking of the Han cultural elite. Indeed, Tung himself was a beneficiary of this intellectual syncretism, for he freely tapped the spiritual resources of his time in formulating his own world view: that human actions have cosmic consequences.

Tung's inquiries on the meaning of the five agents (metal, wood, water, fire, and earth), the correspondence of human beings and the numerical categories of Heaven, and the sympathetic activation of things of the same kind, as well as his studies of cardinal Confucian values such as humanity, rightness, ritual, wisdom, and trustworthiness, enabled him to develop an elaborate world view integrating Confucian ethics with naturalistic cosmology. What Tung accomplished was not merely a theological justification for the emperor as the "Son of Heaven"; rather, his theory of mutual responsiveness between Heaven and man provided the Confucian scholars with a higher law by which to judge the conduct of the ruler.

Despite Tung's immense popularity, his world view was not universally accepted by Han Confucian scholars. A reaction in favour of a more rational and moralistic approach to the Confucian Classics, known as the "Old Text" school, had already set in before the fall of the Western Han. Yang Hsiung (*c.* 53 BC–AD 18) in the *Fa-yen* ("Model Sayings"), a collection of moralistic aphorisms in the style of the *Analects,* and the *T'ai-hsüan ching* ("Classic of the Supremely Profound Principle"), a cosmological speculation in the style of the *I Ching,* presented an alternative world view. This school, claiming its own recensions of authentic classical texts allegedly rediscovered during the Han period and written in an "old" script before the Ch'in unification, was widely accepted in the Eastern Han (AD 25–220). As the institutions of the Erudites and the Imperial university expanded in the Eastern Han, the study of the Classics became more refined and elaborate. Confucian scholasticism, however, like its counterparts in Talmudic and biblical studies, became too professionalized to remain a vital intellectual force.

Confucian scholasticism

Yet Confucian ethics exerted great influence on government, schools, and society at large. Toward the end of the Han as many as 30,000 students attended the Imperial university. All public schools throughout the land offered regular sacrifices to Confucius, and he virtually became the patron saint of education. Many Confucian temples were also built. The Imperial courts continued to honour Confucius from age to age; a Confucian temple eventually stood in every one of the 2,000 counties. As a result, the teacher, together with Heaven, Earth, the emperor, and parents, became one of the most respected authorities in traditional China.

**Confucian ethics in the Taoist and Buddhist context.** Incompetent rulership, faction-ridden bureaucracy, a mismanaged tax structure, and domination by eunuchs toward the end of the Eastern Han first prompted widespread protests by the Imperial university students. The high-handed policy of the court to imprison and kill thousands of them and their official sympathizers in AD 169 may have put a temporary stop to the intellectual revolt, but the downward economic spiral made the life of the peasantry unbearable. The peasant rebellion led by Confucian scholars as well as Taoist religious leaders of faith-healing sects, combined with open insurrections of the military, brought down the Han dynasty and thus put an end to the first Chinese empire. As the Imperial Han system disintegrated, barbarians invaded from the north. The plains of northern China were fought over, despoiled, and controlled by rival groups, and a succession of states was established in the south. This period of disunity, from the early 3rd to the late 6th century, marked the decline of Confucianism, the upsurge of Neo-Taoism, and the spread of Buddhism.

Fall of the Han dynasty

The prominence of Taoism and Buddhism among the cultural elite and the populace in general, however, did not mean that the Confucian tradition had disappeared. In fact, Confucian ethics was by then virtually inseparable from the moral fabric of Chinese society. Confucius continued to be universally honoured as the paradigmatic sage. The outstanding Taoist thinker Wang Pi (226–249) argued that Confucius, by not speculating on the nature of the Tao, had an experiential understanding of it superior to Lao-tzu's. The Confucian Classics remained the foundation of all literate culture, and sophisticated commen-

taries were being produced throughout the age. Confucian values continued to dominate in such political institutions as the central bureaucracy, the recruitment of officials, and local governance. The political forms of life also were distinctively Confucian. When a barbarian state adopted a Sinicization policy, notably the case of the Northern Wei (386–535), it was by and large Confucian in character. In the south systematic attempts were made to strengthen family ties by establishing clan rules, genealogical trees, and ancestral rituals based on Confucian ethics.

Revival of Confucian learning in the T'ang dynasty

The reunification of China by the Sui (581–618) and the restoration of lasting peace and prosperity by the T'ang (618–907) gave a powerful stimulus to the revival of Confucian learning. The publication of a definitive, official edition of the *Wu Ching* with elaborate commentaries and subcommentaries and the implementation of Confucian rituals at all levels of governmental practice, including the compilation of the famous T'ang legal code, were two outstanding examples of Confucianism in practice. An examination system was established based on literary competence. This system made the mastery of Confucian Classics a prerequisite for political success and was, therefore, perhaps the single most important institutional innovation in defining elite culture in Confucian terms.

The T'ang dynasty, nevertheless, was dominated by Buddhism and, to a lesser degree, by Taoism. The philosophical originality of the dynasty was mainly represented by monk-scholars such as Chi-tsang (549–623), Hsüan-tsang (602–664), and Chih-i (538–597). An unintended consequence in the development of Confucian thought in this context was the prominent rise of the metaphysically significant Confucian texts, notably *Chung yung* ("Doctrine of the Mean") and *I-chuan* ("The Great Commentary of the Classic of Changes"), which appealed to some Buddhist and Taoist thinkers. A sign of a possible Confucian turn in the T'ang was Li Ao's (d. *c.* 844) essay on "Returning to Nature" that foreshadowed features of Sung (960–1279) Confucian thought. The most influential precursor of a Confucian revival, however, was Han Yü (768–824). He attacked Buddhism from the perspectives of social ethics and cultural identity and provoked interest in the question of what actually constitutes the Confucian Way. The issue of *Tao-t'ung,* the transmission of the Way or the authentic method to repossess the Way, has stimulated much discussion in the Confucian tradition since the 11th century.

## THE CONFUCIAN REVIVAL

The Buddhist conquest of China and the Chinese transformation of Buddhism, a process entailing the introduction, domestication, growth, and appropriation of a distinctly Indian form of spirituality, lasted for at least six centuries. Since Buddhist ideas were introduced to China via Taoist categories and since the development of the Taoist religion benefited from having Buddhist institutions and practices as models, the spiritual dynamics in medieval China was characterized by Buddhist and Taoist values. The reemergence of Confucianism as the leading intellectual force thus involved both a creative response to the Buddhist and Taoist challenge and an imaginative reappropriation of classical Confucian insights. Furthermore, after the collapse of the T'ang dynasty, the grave threats to the survival of Chinese culture from the Khitan, the Juchen (Chin), and later the Mongols prompted the literati to protect their common heritage by deepening their communal critical self-awareness. To enrich their personal knowledge as well as to preserve China as a civilization-state, they explored the symbolic and spiritual resources that made Confucianism a living tradition.

**The Sung masters.** The Sung dynasty (960–1279) was militarily weak and much smaller than the T'ang, but its cultural splendour and economic prosperity were unprecedented in human history. The Sung's commercial revolution produced flourishing markets, densely populated urban centres, elaborate communication networks, theatrical performances, literary groups, and popular religions—developments that tended to remain unchanged into the 19th century. Technological advances in agriculture, textiles, lacquer, porcelain, printing, maritime trade, and weaponry demonstrated that China excelled in the fine arts as well as in the sciences. The decline of the aristocracy, the widespread availability of printed books, the democratization of education, and the full implementation of the examination system produced a new social class, the gentry, noted for its literary proficiency, social consciousness, and political participation. The outstanding members of this class, such as the classicists Hu Yüan (993–1059) and Sun Fu (992–1057), the reformers Fan Chung-yen (989–1052) and Wang An-shih (1021–86), the writer-officials Ou-yang Hsiu (1007–72) and Su Shih (pen name of Su Tung-p'o; 1036–1101), and the statesman-historian Ssu-ma Kuang (1019–86), contributed to the revival of Confucianism in education, politics, literature, and history and collectively to the development of a scholarly official style, a way of life informed by Confucian ethics.

The Sung literati

The Confucian revival, understood in traditional historiography as the establishment of the lineage of the *Tao-hsueh* ("Learning of the Tao"), nevertheless can be traced through a line of Neo-Confucian thinkers from Chou Tun-i (1017–73) by way of Shao Yung (1011–77), Chang Tsai (1020–77), the brothers Ch'eng Hao (1032–85) and Ch'eng I (1033–1107), and the great synthesizer Chu Hsi (1130–1200). These men developed a comprehensive humanist vision in which cultivation of the self was integrated with social ethics and moral metaphysics. In the eyes of the Sung literati this new philosophy faithfully restored the classical Confucian insights and successfully applied them to the concerns of their own age.

Chou Tun-i ingeniously articulated the relationship between the "great transformation" of the cosmos and the moral development of human beings. In his metaphysics, humanity, as the recipient of the highest excellence from Heaven, is itself a centre of cosmic creativity. He developed this all-embracing humanism by a thought-provoking interpretation of the Taoist diagram of T'ai Chi ("Great Ultimate"). Shao Yung elaborated on the metaphysical basis of human affairs, insisting that a disinterested numerological mode of analysis is most appropriate for understanding the "supreme principles governing the world." Chang Tsai, on the other hand, focused on the omnipresence of *ch'i* ("vital energy"). He also advocated the oneness of *li* ("principle"; comparable to the idea of Natural Law) and the multiplicity of its manifestations, which is created as the principle expresses itself through the "vital energy." As an article of faith he pronounced in the "Western Inscription": "Heaven is my father and Earth is my mother, and even such a small being as I finds a central abode in their midst. Therefore that which fills the universe I regard as my body and that which directs the universe I consider as my nature. All people are my brothers and sisters, and all things are my companions."

Neo-Confucian philosophy

This theme of mutuality between Heaven and human beings, consanguinity between man and man, and harmony between man and nature was brought to fruition in Ch'eng Hao's definition of humanity as "forming one body with all things." To him the presence of T'ien-li ("Heavenly Principle") in all things as well as in human nature enables the human mind to purify itself in a spirit of reverence. Ch'eng I, following his brother's lead, formulated the famous dictum, "self-cultivation requires reverence; the extension of knowledge consists in the investigation of things." By making special reference to *ko-wu* ("investigation of things"), he raised doubts about the appropriateness of focusing exclusively on the illumination of the mind in self-cultivation, as his brother seems to have done. The learning of the mind as advocated by Ch'eng Hao and the learning of the principle as advocated by Ch'eng I became two distinct modes of thought in Sung Confucianism.

Chu Hsi, clearly following Ch'eng I's School of Principle and implicitly rejecting Ch'eng Hao's School of Mind, developed a pattern of interpreting and transmitting the Confucian Way that for centuries defined Confucianism not only for the Chinese but for the Koreans and the Japanese as well. If, as quite a few scholars have advocated, Confucianism represents a distinct form of East Asian spirituality, it is the Confucianism shaped by Chu Hsi. Chu Hsi virtually reconstituted the Confucian tradition, giving it new structure, new texture, and new meaning.

He was more than a synthesizer; through conscientious appropriation and systematic interpretation he gave rise to a new Confucianism, known as Neo-Confucianism in the West but often referred to as Li Hsüeh ("Learning of the Principle") in modern China.

Chu Hsi's restructuring of the scriptural tradition

The "Doctrine of the Mean" and the "Great Learning," two chapters in the *Li chi,* had become independent treatises and, together with the *Analects* and *Mencius,* had been included in the core curriculum of Confucian education for centuries before Chu Hsi's birth. But by putting them into a particular sequence, the "Great Learning," the *Analects, Mencius,* and the "Doctrine of the Mean," synthesizing their commentaries, interpreting them as a coherent humanistic vision, and calling them the Four Books, Chu Hsi fundamentally restructured the Confucian scriptural tradition. The Four Books, placed above the Five Classics, became the central texts for both primary education and civil service examinations in traditional China from the 14th century. Thus they have exerted far greater influence on Chinese life and thought in the past 600 years than any other book.

As an interpreter and transmitter of the Confucian Way Chu Hsi identified which early Sung masters belonged to the lineage of Confucius and Mencius. His judgment, later widely accepted by governments in East Asia, was based principally on philosophical insight. Chou Tun-i, Chang Tsai, and the Ch'eng brothers, the select four, were Chu Hsi's cultural heroes. Shao Yung and Ssu-ma Kuang were originally on his list, but Chu Hsi apparently changed his mind, perhaps because of Shao's excessive metaphysical speculation and Ssu-ma's obsession with historical facts.

Up until Chu Hsi's time the Confucian thinking of the Sung masters was characterized by a few fruitfully ambiguous concepts, notably the Great Ultimate, principle, vital energy, nature, mind, and humanity. Chu Hsi defined the process of the investigation of things as a rigorous discipline of the mind to probe the principle in things. He recommended a twofold method of study: to cultivate a sense of reverence and to pursue knowledge. This combination of morality and wisdom made his pedagogy an inclusive approach to humanist education. Reading, sitting quietly, ritual practice, physical exercise, calligraphy, arithmetic, and empirical observation all had a place in his pedagogical program. Chu Hsi reestablished the White Deer Grotto in present Kiangsi Province as an academy. It became the intellectual centre of his age and provided an instructional model for all schools in East Asia for generations to come.

Chu Hsi was considered the preeminent Confucian scholar in Sung China, but his interpretation of the Confucian Way was seriously challenged by his contemporary, Lu Chiu-yüan (Lu Hsiang-shan, 1139–93). Claiming that he appropriated the true wisdom of Confucian teaching by reading Mencius, Lu criticized Chu Hsi's theory of the investigation of things as fragmented and ineffective empiricism. Instead he advocated a return to Mencian moral idealism by insisting that establishing the "great body" (*i.e.,* Heaven-endowed nobility) is the primary precondition for self-realization. To him the learning of the mind as a quest for self-knowledge provided the basis upon which the investigation of things assumed its proper significance. Lu's confrontation with Chu Hsi in the famous meeting at the Goose Lake Temple in 1175 further convinced him that Confucianism as Chu Hsi had shaped it was not Mencian. Although Lu's challenge remained a minority position for some time, his learning of the mind later became a major intellectual force in Ming China (1368–1644) and Tokugawa Japan (1603–1867).

Lu Chiu-yüan's return to Mencian moral idealism

**Confucian learning in Chin, Yüan, and Ming.** For about 150 years, from the time the Sung court moved its capital to the South and reestablished itself there in 1127, North China was ruled by three conquest dynasties, the Liao (907–1125), Hsi Hsia (1038–1227), and Chin (1115–1234). Although the bureaucracies and political cultures of both Liao and Hsi Hsia were under Confucian influence, no discernible intellectual developments helped to further the Confucian tradition there. In the Juchen Chin dynasty, however, despite the paucity of information about the Confucian renaissance in the Southern Sung, the Chin scholar-officials continued the classical, artistic, literary, and historiographic traditions of the North and developed a richly textured cultural form of their own. Chao Ping-wen's (1159–1232) combination of literary talent and moral concerns and Wang Jo-hsü's (1174–1243) scholarship in Classics and history, as depicted in Yüan Hao-wen's (1190–1257) biographical sketches and preserved in their collected works, compared well with the high standards set by their counterparts in the South.

When the Mongols reunited China in 1279, the intellectual dynamism of the South profoundly affected the northern style of scholarship. Although the harsh treatment of scholars by the conquest Yüan (Mongol) dynasty (1206–1368) seriously damaged the well-being of the scholarly community, outstanding Confucian thinkers nevertheless emerged throughout the period. Some opted to purify themselves so that they could repossess the Way for the future; some decided to become engaged in politics to put their teaching into practice.

Hsü Heng's instruction of the Mongols

Hsü Heng (1209–81) took a practical approach. Appointed by Kublai, the Great Khan in Marco Polo's *Description of the World,* as the president of the Imperial Academy and respected as the leading scholar in the court, Hsü conscientiously introduced Chu Hsi's teaching to the Mongols. He assumed personal responsibility for educating the sons of the Mongol nobility to become qualified teachers of Confucian Classics. His erudition and skills in medicine, legal affairs, irrigation, military science, arithmetic, and astronomy enabled him to be an informed adviser to the conquest dynasty. He set the tone for the eventual success of the Confucianization of Yüan bureaucracy. In fact, it was the Yüan court that first officially adopted the Four Books as the basis of the civil service examination, a practice that was to be observed until 1905. Thanks to Hsü Heng, Chu Hsi's teaching prevailed in the Mongol period, but it was significantly simplified.

The hermit-scholar, Liu Yin (1249–93), on the other hand, allegedly refused Kublai Khan's summons in order to maintain the dignity of the Confucian Way. To him education was for self-realization. Loyal to the Chin culture in which he was reared and faithful to the Confucian Way that he had learned from the Sung masters, Liu Yin rigorously applied philological methods to classical studies and strongly advocated the importance of history. Although true to Chu Hsi's spirit, by taking seriously the idea of the investigation of things, he put a great deal of emphasis on the learning of the mind. Liu Yin's contemporary, Wu Cheng (1249–1333), further developed the learning of the mind. He fully acknowledged the contribution of Lu Chiu-yüan to the Confucian tradition, even though as an admirer of Hsü Heng he considered himself a follower of Chu Hsi. Wu assigned himself the challenging task of harmonizing the difference between Chu and Lu. As a result, he reoriented Chu's balanced approach to morality and wisdom to accommodate Lu's existential concern for self-knowledge. This prepared the way for the revival of Lu's learning of the mind in the Ming (1368–1644).

Trend toward moral subjectivity

The thought of the first outstanding Ming Confucian scholar, Hsüeh Hsüan (1389–1464), already revealed the turn toward moral subjectivity. Although a devoted follower of Chu Hsi, Hsüeh's *Records of Reading* clearly shows that he considered the cultivation of "mind and nature" to be particularly important. Two other early Ming scholars, Wu Yü-pi (1391–1469) and Ch'en Hsien-chang (1428–1500), helped to define Confucian education for those who studied the Classics not simply in preparation for examinations but as learning of the "body and mind." They cleared the way for Wang Yang-ming (1472–1529), the most influential Confucian thinker after Chu Hsi.

As a critique of excessive attention to philological details characteristic of Chu Hsi's followers, Wang Yang-ming allied himself with Lu Chiu-yüan's learning of the mind. He advocated the precept of uniting thought and action. By focusing on the transformative power of the will, he inspired a generation of Confucian students to return to the moral idealism of Mencius. His own personal example of combining teaching with bureaucratic routine, administrative responsibility, and leadership in military campaigns demonstrated that he was a man of deeds.

Despite his competence in practical affairs, Wang's primary concern was moral education, which he felt had to be grounded in the "original substance" of the mind. This he later identified as *liang-chih* ("good conscience"), by which he meant innate knowledge or a primordial existential awareness possessed by every human being. He further suggested that good conscience as the Heavenly Principle is inherent in all beings from the highest spiritual forms to grass, wood, bricks, and stone. Because the universe consists of vital energy informed by good conscience, it is a dynamic process rather than a static structure. Human beings can learn to regard Heaven and Earth and the myriad things as one body by extending their good conscience to embrace an ever-expanding network of relationships.

Wang Yang-ming's dynamic idealism

Wang Yang-ming's dynamic idealism, as Wing-tsit Chan, the dean of Chinese philosophy in North America, characterized it, set the Confucian agenda for several generations in China. His followers, such as the communitarian Wang Chi (1498–1583), who devoted his long life to building a community of the like-minded, and the radical individualist Li Chih (1527–1602), who proposed to reduce all human relationships to friendship, broadened Confucianism to accommodate a variety of life-styles.

Among Wang's critics, Liu Tsung-chou (1578–1645) was perhaps the most brilliant. His *Human Schemata* (*Jen-p'u*) offered a rigorous phenomenological description of human mistakes as a corrective to Wang Yang-ming's moral optimism. Liu's student Huang Tsung-hsi (1610–95) compiled a comprehensive biographical history of Ming Confucians based on Liu's writings. One of Huang's contemporaries, Ku Yen-wu (1613–82), was also a critic of Wang Yang-ming. He excelled in his studies of political institutions, ancient phonology, and classical philology. While Ku was well-known in his time and honoured as the patron saint of "evidential learning" in the 18th century, his contemporary Wang Fu-chih (1619–92) was discovered 200 years later as one of the most sophisticated original minds in the history of Confucian thought. His extensive writings on metaphysics, history, and the Classics made him a thorough critic of Wang Yang-ming and his followers.

**The age of Confucianism: Yi-dynasty Korea, Tokugawa Japan, and Ch'ing China.** Among all the dynasties, Chinese and foreign, the long-lived Yi (Chosŏn) in Korea (1392–1910) was undoubtedly the most thoroughly Confucianized. Since the 15th century, when the aristocracy (*yangban*) defined itself as the carrier of Confucian values, the penetration of court politics and elite culture by Confucianism had been unprecedented. Even today, as manifested in political behaviour, legal practice, ancestral veneration, genealogy, village schools, and student activism, the vitality of the Confucian tradition is widely felt in South Korea.

Yi T'oegye (1501–70), the single most important Korean Confucian, helped shape the character of Yi Confucianism through his creative interpretation of Chu Hsi's teaching. Critically aware of the philosophical turn engineered by Wang Yang-ming, T'oegye transmitted the Chu Hsi legacy as a response to the advocates of the learning of the mind. As a result, he made Yi Confucianism at least as much a true heir to Sung learning as Ming Confucianism was. Indeed, his *Discourse on the Ten Sagely Diagrams,* an aid for educating the king, offered a depiction of all the major concepts in Sung learning. His exchange of letters with Ki Taesung (1527–72) in the famous Four-Seven debate, which discussed the relationship between Mencius' four basic human feelings—commiseration, shame, modesty, and right and wrong—and seven emotions, such as anger and joy, raised the level of Confucian dialogue to a new height of intellectual sophistication.

In addition, Yi Yulgok's (1536–84) challenge to T'oegye's re-presentation of Chu Hsi's Confucianism, from the perspective of Chu's thought itself, significantly enriched the repertoire of the learning of the principle. The leadership of the central government, supported by the numerous academies set up by aristocratic families and by institutions such as the community compact system and the village schools, made the learning of the principle not only a political ideology but also a common creed in Korea.

In Japan, Chu Hsi's teaching, as interpreted by T'oegye, was introduced to Yamazaki Ansai (1618–82). A distinctive feature of Yamazaki's thought was his recasting of native Shintōism in Confucian terminology. The diversity and vitality of Japanese Confucianism was further evident in the appropriation of Wang Yang-ming's dynamic idealism by the samurai-scholars, notably Kumazawa Banzan (1619–91). It is, however, in Ogyū Sorai's (1666–1728) determination to rediscover the original basis of Confucian teaching by returning to its pre-Confucian sources that a true exemplification of the independent-mindedness of Japanese Confucians is found. Indeed, Ogyū's brand of ancient learning with its particular emphasis on philological exactitude foreshadowed a similar scholarly movement in China by at least a generation. Although Tokugawa Japan was never as Confucianized as Yi Korea had been, virtually every educated person in Japanese society was exposed to the Four Books by the end of the 17th century.

Developments in Japan

The Confucianization of Chinese society reached its apex during the Ch'ing (1644–1911/12) when China was again ruled by a conquest (Manchu) dynasty. The Ch'ing emperors outshone their counterparts in the Ming in presenting themselves as exemplars of Confucian kingship. They transformed Confucian teaching into a political ideology, indeed a mechanism of control. Jealously guarding their Imperial prerogatives as the ultimate interpreters of Confucian truth, they undermined the freedom of scholars to transmit the Confucian Way by imposing harsh measures, such as literary inquisition. It was Ku Yen-wu's classical scholarship rather than his insights on political reform that inspired the 18th-century evidential scholars. Tai Chen, the most philosophically-minded philologist among them, couched his brilliant critique of Sung learning in his commentary on "The Meanings of Terms in the *Book of Mencius.*" Tai Chen was one of the scholars appointed by the Ch'ien-lung emperor in 1773 to compile an Imperial manuscript library. This massive scholarly attempt, *The Complete Library of the Four Treasures,* is symbolic of the grandiose intent of the Manchu court to give an account of all the important works of the four branches of learning—the Classics, history, philosophy, and literature—in Confucian culture. The project comprised more than 36,000 volumes with comments on about 10,230 titles, employed as many as 15,000 copyists, and took 20 years to complete. The Ch'ien-lung emperor and the scholars around him may have expressed their cultural heritage in a definitive form, but the Confucian tradition was yet to encounter its most serious threat.

## MODERN TRANSFORMATION

At the time of the first Opium War (1839–42) East Asian societies had been Confucianized for centuries. The continuous growth of Mahāyāna Buddhism throughout Asia and the presence of Taoism in China, shamanism in Korea, and Shintōism in Japan did not undermine the power of Confucianism in government, education, family rituals, and social ethics. In fact, Buddhist monks were often messengers of Confucian values, and the coexistence of Confucianism with Taoism, shamanism, and Shintōism actually characterized the syncretic East Asian religious life. The impact of the West, however, so fundamentally undermined the Confucian roots in East Asia that it has come to be widely debated whether or not Confucianism can remain a viable tradition in modern times.

The impact of the West

Beginning in the 19th century, Chinese intellectuals' faith in the ability of Confucian culture to withstand the impact of the West became gradually eroded. This loss of faith may be perceived in Lin Tse-hsü's (1785–1850) moral indignation against the British, followed by Tseng Kuo-fan's (1811–72) pragmatic acceptance of the superiority of Western technology, K'ang Yu-wei's (1858–1927) sweeping recommendation for political reform, and Chang Chih-tung's (1837–1909) desperate, eclectic attempt to save the essence of Confucian learning, which, however, eventually led to the anti-Confucian iconoclasm of the so-called May Fourth Movement in 1919. The triumph of Marxism-Leninism as the official ideology of the People's Republic of China in 1949 relegated Confucian rhetoric to the background. The modern Chinese intelligentsia, however, maintained unacknowledged, sometimes uncon-

scious, continuities with the Confucian tradition at every level of life—behaviour, attitude, belief, and commitment. Indeed, Confucianism remains an integral part of the psycho-cultural construct of the contemporary Chinese intellectual as well as of the Chinese peasant.

The emergence of Japan and the other newly industrialized Asian countries (South Korea, Taiwan, Hong Kong, and Singapore) as the most dynamic region of economic development since World War II has generated much scholarly interest. Labeled the "Sinitic World in Perspective," "The Second Case of Industrial Capitalism," the "Eastasia Edge," or "the Challenge of the Post-Confucian States," this phenomenon has raised questions about how the typical East Asian institutions, still suffused with Confucian values—such as a paternalistic government, an educational system based on competitive examinations, the family with emphasis on loyalty and cooperation, and local organizations informed by consensus—have adapted themselves to the imperatives of modernization.

Some of the most creative and influential intellectuals in contemporary China have continued to think from Confucian roots. Hsiung Shih-li's ontological reflection, Liang Shu-ming's cultural analysis, Fung Yu-lan's reconstruction of the learning of the principle, Ho Lin's new interpretation of the learning of the mind, T'ang Chün-i's philosophy of culture, Hsü Fu-kuan's social criticism, and Mou Tsung-san's moral metaphysics are noteworthy examples. Although some of the most articulate intellectuals in the People's Republic of China criticize their Confucian heritage as the embodiment of authoritarianism, bureaucratism, nepotism, conservatism, and male chauvinism, others in China, Taiwan, Hong Kong, Singapore, and North America have imaginatively established the relevance of Confucian humanism to China's modernization. The revival of Confucian studies in South Korea, Taiwan, Hong Kong, and Singapore has been under way for more than a generation, though Confucian scholarship in Japan remains unrivaled. Confucian thinkers in the West, inspired by religious pluralism and liberal democratic ideas, have begun to explore the possibility of a third epoch of Confucian humanism. They uphold that its modern transformation, as a creative response to the challenge of the West, is a continuation of its classical formulation and its medieval elaboration. Scholars in mainland China have also begun to explore the possibility of a fruitful interaction between Confucian humanism and democratic liberalism in a socialist context.

Continuity and criticism in the late 20th century

**BIBLIOGRAPHY.** The study of Confucius and Confucianism, not only as a historically significant inquiry but also as a philosophically meaningful and challenging endeavour, has come of age in the English-speaking world since the 1970s. In addition to the following bibliography, see WING-TSIT CHAN, *An Outline and an Annotated Bibliography of Chinese Philosophy,* rev. ed. (1969); and LAURENCE G. THOMPSON, *Chinese Religion in Western Languages: A Comprehensive and Classified Bibliography of Publications in English, French, and German Through 1980* (1985).

*Classical age:* For a major study of Confucius, see H.G. CREEL, *Confucius: The Man and the Myth* (1949, reissued 1975), also published as *Confucius and the Chinese Way* (1949, reprinted 1960). Also see HERBERT FINGARETTE, *Confucius—The Secular as Sacred* (1972), which perceives the Confucian idea of ritual as a philosophical issue; BENJAMIN I. SCHWARTZ, *The World of Thought in Ancient China* (1985), which approaches Confucius and Confucianism as a challenging intellectual enterprise in comparative studies of great civilizations; and D.C. LAU, *The Analects (Lun Yü)* (1979, reissued 1986), and *Mencius* (1970), modern translations from Chinese. I.A. RICHARDS, *Mencius on the Mind: Experiments in Multiple Definition* (1932, reissued 1983); and EZRA POUND (trans.), *The Classic Anthology Defined by Confucius* (1954, reprinted 1976), also published as *The Confucian Odes* (1959), are literary achievements. RICHARD WILHELM (trans.), *The I Ching: or, Book of Changes,* 3rd ed. (1967, reprinted 1981; originally published in German, 1924), is unsurpassed in its richness of primary sources and clarity of presentation. For scholarly interpretations of classical Confucian thought, see DONALD J. MUNRO, *The Concept of Man in Early China* (1969); HOMER H. DUBS, *Hsüntze: The Moulder of Ancient Confucianism* (1927, reissued 1966); TU WEI-MING, *Centrality and Commonality: An Essay on Chung-yung* (1976); and HELLMUT WILHELM, *Heaven, Earth, and Man in "The Book of Changes"* (1977). JOHN K. SHRYOCK, *The Origin and Development of the State Cult of Confucius* (1932, reprinted 1966), is a pioneering attempt to study Confucianism as a Chinese national institution.

*The Confucian tradition:* Important primary sources, all translated from Chinese, can be found in WING-TSIT CHAN (trans.), *Reflections on Things at Hand: The Neo-Confucian Anthology* (1967), writings compiled by Chu Hsi and Lü Tsu-ch'ien, *Neo-Confucian Terms Explained: The Pei-Hsi Tzu-I* (1986), writings by Ch'en Ch'un, and *Instructions for Practical Living, and Other Neo-Confucian Writings* (1963), writings by Wang Yang-ming; and JULIA CHING (ed. and trans.), *The Records of Ming Scholars* (1987), excerpts from writings by Huang Tsung-hsi. Several symposium volumes dedicated to the study of the Neo-Confucian form of life have been published, including WM. THEODORE DE BARY (ed.), *Self and Society in Ming Thought* (1970), and *The Unfolding of Neo-Confucianism* (1975); WM. THEODORE DE BARY and IRENE BLOOM (eds.), *Principle and Practicality: Essays in Neo-Confucianism and Practical Learning* (1979); HOK-LAM CHAN and WM. THEODORE DE BARY (eds.), *Yüan Thought: Chinese Thought and Religion Under the Mongols* (1982); and WM. THEODORE DE BARY and JAHYUN KIM HABOUSH (eds.), *The Rise of Neo-Confucianism in Korea* (1985). For an impressive collection of essays on Chu Hsi, see WING-TSIT CHAN (ed.), *Chu Hsi and Neo-Confucianism* (1986). Studies on major thinkers include CHI-YUN CHEN, *Hsün Yüeh (A.D. 148–209): The Life and Reflection of an Early Medieval Confucian* (1975); JAMES T.C. LIU, *Ou-yang Hsiu: An Eleventh-Century Neo-Confucianist* (1967; originally published in Chinese, 1963); A.C. GRAHAM, *Two Chinese Philosophers: Ch'êng Ming-tao and Ch'êng Yi-ch'uan* (1958, reprinted 1978); HOYT CLEVELAND TILLMAN, *Utilitarian Confucianism: Ch'en Liang's Challenge to Chu Hsi* (1982); WINSTON WAN LO, *The Life and Thought of Yeh Shih* (1974); JULIA CHING, *To Acquire Wisdom: The Way of Wang Yang-ming* (1976); TU WEI-MING, *Neo-Confucian Thought in Action: Wang Yang-ming's Youth (1472–1509)* (1976); EDWARD T. CH'IEN, *Chiao Hung and the Restructuring of Neo-Confucianism in the Late Ming* (1986); and DAVID S. NIVISON, *The Life and Thought of Chang Hsüeh-ch'eng, 1738–1801* (1966). Monographs on significant issues include WM. THEODORE DE BARY, *Neo-Confucian Orthodoxy and the Learning of the Mind-and-Heart* (1981), and *The Liberal Tradition in China* (1983); DANIEL K. GARDNER, *Chu Hsi and the Ta-hsueh: Neo-Confucian Reflection on the Confucian Canon* (1986); JOHN W. DARDESS, *Confucianism and Autocracy: Professional Elites in the Founding of the Ming Dynasty* (1983); BENJAMIN A. ELMAN, *From Philosophy to Philology: Intellectual and Social Aspects of Change in Late Imperial China* (1984); and TU WEI-MING, *Humanity and Self-Cultivation: Essays in Confucian Thought* (1979). Three studies in comparative philosophy and religion are noteworthy: DAVID E. MUNGELLO, *Leibniz and Confucianism: The Search for Accord* (1977); JULIA CHING, *Confucianism and Christianity* (1977); and JACQUES GERNET, *China and the Christian Impact: A Conflict of Cultures* (1985; originally published in French, 1982).

*Modern transformation:* Confucianism as it exists in the 20th century is discussed in WING-TSIT CHAN, *Religious Trends in Modern China* (1953, reissued 1969). The thesis that Confucian humanism is incompatible with modernization defined in terms of industrial capitalism was first formulated in MAX WEBER, *The Religion of China: Confucianism and Taoism* (1951; originally published in German, 1922). JOSEPH R. LEVENSON, *Confucian China and Its Modern Fate: A Trilogy,* 3 vol. in 1 (1965, reissued 1968), further develops the claim that Confucianism could not survive the challenge of Western science and technology. Critical reflections on the Weberian and Levensonian interpretation include HAO CHANG, *Liang Ch'i-ch'ao and Intellectual Transition in China, 1890–1907* (1971); CHARLOTTE FURTH (ed.), *The Limits of Change: Essays on Conservative Alternatives in Republican China* (1976); and THOMAS A. METZGER, *Escape from Predicament: Neo-Confucianism and China's Evolving Political Culture* (1977). The reasons for iconoclastic attacks on the Confucian tradition are explored in LIN YÜ-SHENG, *The Crisis of Chinese Consciousness: Radical Antitraditionalism in the May Fourth Era* (1979); and KAM LOUIE, *Critiques of Confucius in Contemporary China* (1980). Studies of modern Confucian personalities include KUNG-CHUAN HSIAO, *A Modern China and a New World: K'ang Yu-wei, Reformer and Utopian, 1858–1927* (1975); HAO CHANG, *Chinese Intellectuals in Crisis: Search for Order and Meaning (1890–1911)* (1987); JOEY BONNER, *Wang Kuo-wei: An Intellectual Biography* (1986); and GUY S. ALITTO, *The Last Confucian: Liang Shu-ming and the Chinese Dilemma of Modernity,* 2nd ed. (1986). For contemporary manifestations of the Confucian tradition, see IRENE EBER (ed.), *Confucianism: The Dynamics of a Tradition* (1986); and TU WEI-MING, *Confucian Ethics Today: The Singapore Challenge* (1984).

(T.W.-m.)

# Conservation of Natural Resources

Although the idea of conservation is probably as old as the human species, the use of the word in its present context is relatively recent. Over the years conservation has acquired many connotations: to some it has meant the protection of wild nature, to others the sustained production of useful materials from the resources of the Earth. The most widely accepted definition, presented in 1980 in *World Conservation Strategy* by the International Union for Conservation of Nature and Natural Resources, is that of "the management of human use of the biosphere so that it may yield the greatest sustainable benefit while maintaining its potential to meet the needs and aspirations of future generations." The document defines the objectives of the conservation of living resources as: maintenance of essential ecological processes and life-support systems, preservation of genetic diversity, and guarantee of the sustainable use of species and ecosystems. More generally, conservation involves practices that perpetuate the resources of the Earth on which human beings depend and that maintain the diversity of living organisms that share the planet. This includes such activities as the protection and restoration of endangered species, the careful use or recycling of scarce mineral resources, the rational use of energy resources, and the sustainable use of soils and living resources.

Conservation is necessarily based on a knowledge of ecology, the science concerned with the relationship between life and the environment, but ecology itself is based on a wide variety of disciplines, and conservation involves human feelings, beliefs, and attitudes as well as science and technology.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 355.

This article is divided into the following sections:

Concepts important to conservation 663
The need for natural resources 663
The importance of conservation 664
The history of conservation 665
Early practices 665
Recent history 666
Types of natural resources 667
Renewable resources 668
Nonrenewable resources 669
Management of natural resources 670
Managing nonliving resources 670
Managing living resources 674
International problems of resource management 679
The pollution of natural resources 681
Conservation and growth 683
Bibliography 685

## Concepts important to conservation

### THE NEED FOR NATURAL RESOURCES

Like "conservation" itself, the term natural resources has undergone an expansion in meaning as a result of a greater understanding of the relationship of human beings with the world they inhabit. Early in the 20th century natural resources were viewed primarily as sources of useful commodities. They were the raw materials in the environment that were used or capable of being used by people for some purpose: minerals and fuels, forest and grazing resources, wildlife, fisheries, and the like. In a restricted sense, the term is still used in this way. More recently, however, the concept of natural resources has been broadened to include the total natural environment—the entire surface layer of the planet—because all parts of the Earth's surface are of use and of value in that they contribute to the production of the necessities and amenities that people require or demand. Thus, when considered in this respect, the atmosphere, oceans, deserts, and polar regions have all become valuable resources that must be managed with care to provide for the future.

**Primary and secondary needs.** Humanity's primary needs, or natural-resource requirements for existence, include energy in the form of organic foods that are digestible, are capable of being assimilated, and contain adequate amounts of proteins, fats, carbohydrates, vitamins, and minerals; water with a relatively low content of dissolved salts and free from toxic or injurious substances; air that contains an adequate quantity of oxygen but no harmful materials; and an external source of energy for heating and cooking, as well as various materials from which clothing and shelter can be fashioned to provide warmth in cold weather and coolness during excessively hot weather. These basic human needs were supplied to mankind's earliest ancestors by wild vegetation and animal life, by springs and streams, and by the atmosphere and the Sun. Later, the discovery of fire made possible the heating of shelters and also made available a greater choice of foods—those that could be rendered more palatable by being heated.

The primary needs of people are the same today; however, with increased populations and depleted supplies of wild resources, a secondary category of needs has developed. These include those materials or energy sources needed to maintain an urban civilization. In addition to such needs, there is also a wide range of natural-resource "wants." These include the materials, experiences, or space needed to make existence more enjoyable.

Natural-resource "wants"

*Agricultural and urban development.* The development of agriculture enabled people to produce greater amounts of food on a more reliable basis and from smaller areas of land than previously had been possible. The direct dependence on the availability of wild food materials was lost. Greater supplies of food made it possible to provide for greater numbers of people in the agricultural regions, and these numbers soon exceeded the capacity of the original or still-existing natural environment to supply their primary needs. Thus, the first secondary needs developed—farming tools and, later, domestic animals to help use the tools more effectively and, for the latter, the food supplies necessary to keep them alive. In time, to keep the agricultural soils productive, the need for fertilizers of various kinds developed. Increasing dependence upon foods that could not be eaten raw generated the need for materials from which cooking and eating utensils could be fashioned. Thus, requirements for a wide variety of nonliving natural resources developed along with the rise of agricultural lands and settled villages.

With the growth of civilization and the concentration of people into cities, natural-resource requirements increased as secondary needs expanded. It became essential to organize and direct agriculture over large areas in order to provide for urban inhabitants. Effective transportation from farmlands to cities became essential, as did metals and all kinds of other minerals, stones, and timber suitable for the construction of buildings, ships, and vehicles; in addition, greater numbers of domestic animals (cows, hogs, sheep, etc.) were required. Human wants were further increased as the greater leisure of civilized life enabled part of the population to look beyond the problem of mere survival. Thus, a desire for contact with wild nature developed as the urban population became increasingly separated from it through city life.

*Industrial and technological growth.* The greatest expansion of human requirements for natural resources followed the Industrial Revolution during the latter half of the 18th and first half of the 19th centuries and the scientific and technological revolutions that succeeded it in the 20th. Resources that were of no value only a relatively few decades ago are now used—for example, beryllium for rockets and uranium for nuclear fuel. Coal, natural gas, and petroleum—resources that were scarcely used only a century ago—are now consumed in enormous quantities. The amount of food now demanded has involved an enormous input into the agricultural economy of chemical materials and sophisticated farming implements as well as an input of fuel energy that often exceeds the energy value of the food raised. Moreover, because both the demand for luxuries and the degree of wastefulness are excessive, not all consumption of resources is related to the supplying of needs.

Excessive resource demands and waste

In the confusion of needs, wants, and waste related to the current use of natural resources, it is notable that the primary human requirements are the same as they were in primitive times. It would be possible for mankind to survive, although in greatly reduced numbers, by using only wild vegetation and animal life to meet resource needs. This is because the living resources of the Earth contain all the basic requirements for human survival. The need for other resources is the result of the desire to live in greater numbers and at a standard of living considerably higher than that previously enjoyed. By reducing population growth in the future, it would be possible to enjoy a highly developed technology, a high material standard of living, and a wide range of wants and luxuries while still placing little strain upon the Earth's available resources. But, with the growing human population, with an expanding technology that becomes ever more demanding, and with the growing demands for material goods, the pressure on the Earth's natural resources increases steadily. Whether or not available quantities of these resources are sufficient to meet humanity's growing wants and needs is uncertain.

**Rational use of resources.** In its present usage, the conservation of natural resources includes a wide range of subsidiary concepts. One such concept is that of the rational use of the environment, which includes the preservation of certain areas or resources in an essentially undisturbed condition because they either are of scientific interest, have aesthetic appeal, or have recreational value. Preservation also serves an ecological purpose by maintaining the function of the total environment, such as the protection of forests to assure a sustained yield of water into urban reservoirs or the protection of estuaries in order to perpetuate an ocean fishery. But the preservation or the protection of natural resources is not the only concern of conservation; rational use also implies the direct use of resources for their commodity or recreational values. Thus, the harvesting of forest crops, the grazing of grasslands by livestock, the catching of fish, and the hunting of wild animals can be considered a legitimate part of the rational use of natural resources when they are carried out in such a way that the resource is perpetuated and not endangered. Such activities involve another concept, that of sustained yield. Sustained yield is the understanding that, for example, hunting and fishing should take only the annual surplus of individuals so as not to endanger the breeding stock of game animals and fishes. For another example, the cutting of trees and the grazing of grasses should remove only the annual increment or that portion realistically capable of being replaced over a period of years through the operation of natural processes with human assistance when needed.

Sustained yield

**Multiple use and restoration.** Also important to the concept of conservation is the recognition that natural resources have multiple values. In addition to its value as livestock forage, grass, for example, also supports wild animal life, holds soil in place, maintains the productivity of soil, keeps soil and water relationships in proper balance, and helps guarantee streamflow or yields of water to underground channels. Grasslands, moreover, have aesthetic, recreational, and scientific values. All of the many values of grassland must be considered before a decision is made to use a grassland for a particular purpose. Ideally, an area of land can serve many purposes simultaneously or sequentially; *i.e.,* can have multiple uses.

Another of the more hopeful aspects of conservation is the concept of restoration. A forest that has been cut or burned can, with care, regenerate itself. Areas that have been mined and left barren often can be revegetated with reasonable expenditures of money and effort. Depleted animal or plant populations recover their original abundance if suitably protected in an adequate habitat. The restoration of natural vegetation depends upon the ecological process of succession, in which plants with varying degrees of tolerance to extreme conditions in an environment invade a disturbed or barren area and replace one another until a stable, self-perpetuating community is achieved. Restoration is possible, however, only as long as species are protected and the genetic diversity of life is maintained. When species become extinct the restoration of past conditions becomes impossible.

## THE IMPORTANCE OF CONSERVATION

**Values to mankind.** Conservation is essential to human survival. Because life depends upon the proper functioning of the biosphere—the relatively narrow zone of air, water, soil, and rock in which all life on Earth exists—the ultimate purpose of conservation is to maintain the biosphere in a healthy operating condition. Although it is known that green plants supply oxygen to the atmosphere, that plants and animals recycle nutrients, and that plants and animals help maintain the fertility of soils, many of the elements that contribute to the proper functioning of the biosphere have not yet been identified. Because mankind lives with such environmental uncertainties, an attitude of care and protection toward the Earth's living resources is necessary.

The role of conservation in maintaining the biosphere

Certain aspects of conservation, however, such as the prevention of pollution, have a more narrow and immediate importance. There are numerous illustrations of the serious effect of pollutants in air, water, or soil on human health and survival; for example, the accumulation of sulfur dioxide in the air of London during the 1950s led to many deaths that probably would not have otherwise occurred. The dumping of mercury-containing wastes in waters around Japan caused the death of many people and destroyed the health of others, and continuing accumulations of such toxic metals as lead, cadmium, and arsenic in air and water threaten widespread damage to human health.

*Economic value.* Unless viewed in terms of human survival, the economic value of conservation is sometimes difficult to demonstrate. Although the floating plants of the ocean, the microscopic phytoplankton, are of little direct economic value to people, for example, their elimination from the food chain would quickly destroy the world's marine fisheries, which are a major source of human food; in time, even the world's oxygen supply would be severely depleted.

As explained below, much of the apparent conflict in the economics of conservation results from the difference between short-term or individual interests and the long-term interests of groups or of all mankind. The long-term economic community benefits to be derived from stable and productive farmlands and forests are considerable when compared with farms that are exploited, eroded, and abandoned or with forests that are cut, burned, and allowed to deteriorate. Short-term economic considerations, however, may lead individuals or communities to exploit their farms and forests for maximum profit at minimum cost and then move on, leaving the deteriorated lands behind.

*Aesthetic and recreational value.* Appreciation of wild nature as a source of aesthetic pleasure and the use of wildlands and wild-animal resources for recreational enjoyment have long been recognized as among the more important values of conservation. Outdoor-based recreational activities, such as fishing, hunting, boating, swimming, picnicking, sunbathing, hiking, and skiing, are related to the continued existence of natural or near-natural environments as the sites for these activities. Although it is

The appreciation of wild nature

almost impossible to evaluate aesthetic and recreational values in terms of their psychological or sociological importance, because they may vary from one culture to another, evidence indicates that, as personal affluence and the freedom from the sheer struggle for survival increase, the demand for outdoor recreation and outdoor space also increases. Even in the absence of government activity, it has become financially attractive for private investors to provide facilities and opportunities that exploit outdoor recreational resources.

*Scientific value.* Conservation is also of great scientific value. Because relatively little is known about the past, present, and possible future of the biosphere, natural outdoor laboratories, including areas of undisturbed nature, must be maintained in order to conduct the studies needed to acquire knowledge. Moreover, there are many natural resources with undiscovered scientific and technological values. If, for example, all apparently worthless rosy periwinkle plants had been destroyed, an important drug used in treating leukemia would not have been discovered. Because each wild plant and animal contains a storehouse of genetic and biochemical information, the loss of a single species might result in the loss of information that could ultimately have great value for mankind's welfare or survival.

**Conflicting attitudes and issues.** Although the importance of conservation may seem obvious, most of the world's people live too close to the margin of existence to exercise concern for anything more than their immediate survival and well-being. Planning for the future becomes difficult when the present itself is in doubt, and activities that could help tomorrow's generations may seem quixotic to those for whom survival is at stake. Thus, while conservation has made great strides in some areas of the world, it is still too soon for people to have any feeling of security about the future of the environment.

*Short-term versus long-term views.* It is often regarded as essential to the survival or the enrichment of an individual or a group to use resources in such a way as to realize immediate gains or profits. Such activities, however, may impair the future productivity of an area of land, exterminate a species, or destroy the usefulness of a site for any other purpose. In such a situation the short-term, private view conflicts with the long-term, public view. Though many would argue that the public view should be more conservation oriented, emphasizing proper safeguards to prevent deterioration of the environment, there are, nevertheless, times when governments take the short-term view in the face of real or imagined economic or political crises; they may, for example, authorize widespread destruction of resources as a temporary expedient to achieve a military goal or to strengthen the public treasury. But crises will tend to become self-perpetuating if the destruction of resources weakens the country ecologically and economically. Thus, continued, unrestricted population growth in a country poorly equipped to manage its natural resources creates a continuing sense of crisis, because ever-expanding immediate needs are commonly met at the cost of future productivity and environmental stability.

Conflicting governmental roles

As long as human populations were small and the pressures upon the environment were limited, conflicts between long-term and short-term interests made little difference. Deteriorated lands could be abandoned and new lands found because there was sufficient time to permit natural repair of environmental damage. Presently, however, with great and increasing numbers of people on a planet of limited capacity, conservationists are insisting that the difference between short- and long-term points of view be resolved in favour of actions that guarantee the survival of mankind.

*Technological issues.* Technological progress is another important reason for many conflicts in matters pertaining to conservation. Although technology can be a boon, it can also be poorly related to environmental realities. Through the use of technology great environmental changes can quickly be brought about. Although these changes are usually intended to be beneficial, they frequently occur in natural environments in which all things are ecologically related to one another. As a consequence, the changes may produce side effects that were not anticipated or that were discounted as being of little importance, thereby disrupting other human activities or the environment as a whole. Examples of such situations include the polluting effects of certain industries or the spread of waterborne diseases following the construction of major irrigation projects.

*Use of global resources.* A further area of conflict lies in attitudes toward resources that are held in common, such as the atmosphere and oceans. In instances in which the use of such resources is essentially free to the user, and the power to control use does not rest with any recognized authority, the resource often deteriorates. Although each fisherman may feel that his individual activities have very little effect on the resources of the ocean, the effect of the activities of all fishermen may threaten the existence of those resources. Similarly, each automobile driver does not feel that he is contributing much to the pollution of the global atmosphere, but all automobiles throughout the world contribute a total level of pollution that most critics feel cannot long be tolerated. When such situations exist, a recognized controlling authority is usually seen to be necessary.

## The history of conservation

### EARLY PRACTICES

For most of its history, the human species has lived by hunting animals and gathering wild plant foods. By extrapolation from studies of peoples who today live by such methods, it can be suggested that the relationship of hunter-gatherers with nature was relatively benign. It can also be suggested that people acquire and pass on through oral tradition a remarkable amount of knowledge about the plants and animals with which they associate and on which they depend. A number of breakthroughs in modern medicine, for example, have come from observing the therapeutic uses that traditional tribal cultures make of various wild plants. It is also known, however, that in prehistoric times people did modify their natural environment. Many grassland areas throughout the world have come to exist because people used fire as an aid to hunting or to modify vegetation to make it more suitable to their needs. Early hunting and gathering cultures contributed to the extermination of some animal species, although this seems to have been more of an exception than a general practice. For the most part, early humanity lived in an equable balance with the natural environment, if for no other reason than necessity. If they had done serious damage, people could not have survived.

Early mankind and the environment

Agriculture has been practiced only during the last 10,000–12,000 years, and urban civilization has been in existence only during the last 6,000 years. With urban life came pressure upon the natural environment and upon agricultural lands that was sometimes excessive. In the Asian homelands of Western agriculture there is widespread evidence of serious soil erosion during ancient times. Destruction of vegetation and the spread of deserts followed the rise of early urban civilizations in many areas of the Middle East and North Africa.

**Ancient conservation practices.** Certain conservation practices did develop in early civilizations, however. Some species of animals were protected by religious taboos; religious sanctions prevented the destruction of forest groves and sacred mountains. The use of organic fertilizer to maintain soil fertility is found among many more recent primitive peoples and has had a long history in Western agriculture. The Bible is filled with various injunctions governing the use of land and resources that have a conservation function. Civilizations such as those of the Phoenicians and the Incas developed sophisticated techniques of terracing to prevent soil erosion on hillsides and to make more effective use of water for irrigation. The earliest civilizations also show evidence of the creation of reserves or parks to protect wildlife or natural areas. Although they were hunting preserves for the use of royalty, they also served a conservation function.

As civilization developed, the accumulation of human experience led to increasingly sound land-use practices, evidence of which is found in the written descriptions of

Early development of sound land use

Roman agriculture and, later, in the well-tended irrigated fields and gardens developed during the height of Muslim culture. The agricultural landscapes of preindustrial western Europe, Japan, and China reflected great skill in the conservation of soil resources. Irrigated lands in the Nile Valley and volcanic soils in tropical Southeast Asia have been kept fertile and productive over thousands of years.

In preindustrial times, however, concern over wild nature was not widespread, largely because it was viewed as vast and inexhaustible relative to the domain, the numbers, and the power of human beings. This view was a justifiable one, because the 500,000,000 people who inhabited the world in 1600 lacked the energy sources and the machinery to effect great environmental changes. Moreover, most of the Earth's surface was sparsely settled.

**Conservation during and following the age of exploration.** Starting with the voyages of discovery in the 15th century, the influence of European culture was spread over the world. By the 17th century Europeans were equipped with an increasingly powerful technology and a growing ability to modify large areas of the Earth and to subdue less aggressive peoples. During this period the attitudes of explorers and colonists were oriented more toward immediate personal aggrandizement of the lands they visited and settled than toward any concern for the long-term health and productivity of the newly discovered countries. Soil erosion as well as the destruction of natural vegetation and wildlife accompanied the spread of European colonization in the Americas, Australia, and Africa. Nevertheless, during the same period, various conservation ideas and practices were being promoted. Forest conservation, for example, developed sound beginnings in 17th-century England and France, in part because of the disappearance of natural forests as a result of the increasing demand for wood fuel for industrial uses. As early as the 18th century in eastern North America, such men as Thomas Jefferson already had sound ideas for land management and conservation, and a general interest in and concern for wildlife was developing.

Environmental depredations in the 19th century

The 19th century, however, witnessed unusually severe environmental depredations. In Australia, for example, livestock populations were allowed to increase to levels far above what the natural forage could support. Although millions of animals died during drought periods, the process of overforaging damaged the range lands to such a degree that they have not yet recovered. In southern Africa many forms of wildlife were hunted to extinction, and most of the larger mammals were reduced to numbers that endangered their survival. It was in North America, however, that the changes were most dramatic. The great herds of wildlife that inhabited the plains and prairies vanished as the numbers of bison, elk, antelope, and deer were reduced by hunters. Even the larger predatory animals were nearly exterminated, and some of them—varieties of grizzly bear, cougar, and wolf—subsequently became extinct. Many types of birds that once had occurred in great abundance—*e.g.,* the passenger pigeon, Carolina parakeet, and heath hen—were wiped out. Logging and fires combined to menace the once luxurious forests of New England, the states surrounding the Great Lakes, and the South. The grasslands were overgrazed, and in some areas such as California native vegetation was eliminated over most of its range and replaced by species of European and Asian origin. It was in the West Indies and other islands throughout the world that changes were most marked. Native plant and animal species were eradicated and replaced by exotic invaders. By contrast, in the long-settled areas of Europe and Asia, changes were much less marked, as conservation-oriented systems of land management persisted.

**Rise of the modern conservation movement.** It could have been predicted that the modern conservation movement would have its beginnings not in the settled lands of the Old World but in those areas of the New World where, within the memory of a single generation, there had been extreme changes in the landscape and in the abundance of wildlife. The reaction to the destruction of natural resources in those areas precipitated the formation and growth of the conservation movement. As early as 1832, George Catlin, a U.S. artist and author, first proposed the idea of national parks encompassing major areas in which Indians and wild country could both be preserved. In the same decade the botanist William Bartram and the ornithologist John James Audubon were arousing an interest in wildlife and its conservation. A little later, the writers Ralph Waldo Emerson and Henry David Thoreau presented strong arguments concerning the importance of the continued survival of wild nature to the psychological well-being of mankind. Thoreau became one of the first literary advocates of wilderness conservation. The first textbook on conservation, *Man and Nature,* by George Perkins Marsh, appeared in the 1860s. In the same period the author and naturalist John Muir settled in California and became a leading advocate of wilderness preservation.

The establishment of national parks

In 1872 the U.S. Congress proclaimed the Yellowstone region of Wyoming as a national park and also established for the first time a national-government role in the protection and administration of such areas. In 1891 the first of the U.S. forest reserves, forerunners of the system of national forests, was proclaimed in the area around Yellowstone National Park.

Conservation as a national movement owes much to Pres. Theodore Roosevelt and his immediate advisers. Roosevelt's chief forester, Gifford Pinchot, is credited with having first used the term "conservation" in its present context. Pinchot was to become the leader of the nation's Forest Service and, along with Roosevelt, he advocated a utilitarian, "wise use" approach to conservation. In this, Pinchot and Roosevelt came into conflict with representatives of another school of thought, called preservationist and represented by Muir, over the building of a dam in the canyon of the Tuolumne River in Yosemite National Park, to provide a water supply for San Francisco. The philosophy of Roosevelt and Pinchot prevailed and the dam was built, but Muir's ideas lived on with the Sierra Club, of which he was a founder.

## RECENT HISTORY

The recent history of conservation has been marked by a great expansion of government roles in protecting the environment and by a growth of public interest in and support for this process. National-park systems, dedicated to the preservation of wild nature and to the provision of outdoor recreation space, have grown rapidly, and national-forest systems, dedicated to the multiple use of wild-land resources, have also become firmly established. In the United States the conservation of wildlife became a cause of national interest and led to the establishment of a far-ranging system of wildlife refuges and the gradual restoration of most wild animal species to levels approaching, in some cases exceeding, their primitive abundance. On private lands, however, and on government or public-domain lands not specifically reserved as national forests, parks, or refuges, deterioration continued, reaching a peak in the 1930s, when it became widely recognized that those range lands in the public domain had been disastrously overgrazed and that many privately owned farmlands had been depleted or exhausted. Firm control over the management of lands in the public domain and federal intervention to establish soil conservation on privately owned lands were accepted as appropriate activities for the national government.

**Spread of modern conservation practices.** Conservation ideas spread widely, being most readily accepted by those countries that had experienced sudden environmental changes. By the 1920s national parks were to be found on all continents. In 1924 the Soviet Union established the first of its now extensive system of natural reserves (*zapovedniki*). Conservation-oriented management of forest lands, which grew more from its origins in Europe than from practices in the United States, also became more widely accepted throughout the world. The scientific basis for the management of wild grazing lands for the sustained production of forage for livestock was established in U.S. national forests in 1913 and soon spread to other countries. Aldo Leopold in the United States, in 1933, wrote a textbook on game management, in which the conservation and management of wild animal life for such recreational purposes as sport hunting and fishing and for

direct commodity values on a sustained basis received particular emphasis. Leopold's work drew heavily on earlier studies of animal ecology by Charles Sutherland Elton in England; in fact, the establishment in Europe of wildlife reserves and protective laws as well as the managing of lands to produce sustained crops of wildlife long preceded even Elton's work. Subsequently, the management of wild animals in extensive wilderness areas made major strides in Africa, which possesses unusual wildlife resources, and in the Soviet Union, which retains large areas of wildland.

The production of sustained "crops" of wildlife

**New conservation problems and approaches.** After World War II the field of conservation expanded as new problems arose and as some older approaches proved to have been inadequate. With growing populations and increasing pressures on land and resources, planning for their use by taking into account only a single factor or a few factors at the most was found to be highly unsatisfactory. One such instance was the development of more effective synthetic pesticides for use in the control of disease-carrying insects as well as those that prey most heavily upon agricultural crops. The initial results were remarkable. In some countries, where the insecticide DDT was used to control malaria-bearing mosquitoes, the disease was reduced from being an important cause of human illness and mortality to a low and manageable level. Similarly, agricultural pests were drastically reduced, and crop yields soared in many regions. Eventually, however, it was discovered that the pesticides had unexpected and severe consequences on the environment, and by the 1970s their use anywhere for any purpose was open to serious debate.

All forms of pollution also became a matter of major significance as populations and industrial activities increased after World War II. Air in major cities became toxic; water supplies in many heavily populated areas were contaminated. Nuclear radiation had become a major cause for concern by the 1950s and early 1960s when it was found that radioactive materials from test explosions of atomic and hydrogen bombs spread throughout the entire biosphere instead of being confined to the immediate areas in which the tests were conducted.

In response to the need for a much more integrated approach to environmental problems and to natural-resource management than existed at the time, many countries established ministries for the environment or their equivalent, and in 1969 the United States, by the National Environmental Policy Act, established a national Council on Environmental Quality to oversee and help coordinate those activities of government departments that could have an effect upon the environment.

International environmental problems

By 1970, however, the problems of the environment had become international in scope. The oceans were seriously polluted, and no single country could control the situation. Pesticides and other toxic materials spread by air and water currents throughout the world were causing or threatening to cause environmental damage everywhere. But the need for an international approach to conservation problems found most nations generally unprepared to cope with the situation. Conservation-oriented recommendations aimed at controlling the use of radioactive materials, heavy metals, toxic pesticides, or the dumping of petroleum at sea could not be enforced internationally. The need to regulate the exploitation of marine resources was widely acknowledged, but such regulation was ineffective without an empowered international authority.

In recognition of these problems many international conferences were held, new treaties and conventions were proposed, and the need for regulatory power over the environment at an intergovernmental level was stated frequently. The World Health Organization and the World Meteorological Organization began a global program to monitor pollution levels. The United Nations Educational, Scientific and Cultural Organization (UNESCO) launched a major scientific program directed toward the problems of "Man and the Biosphere," and an international conference on environmental problems was held in Stockholm in June 1972. Following the conference, the United Nations General Assembly established the UN Environment Programme (UNEP) to act on the recommendations of the Stockholm meeting. The UNEP surveyed the status of many aspects of the world's environment and natural resources, subsequently publishing its findings in numerous reports. In 1980 the International Union for Conservation of Nature and Natural Resources, with the support of UNEP and the World Wildlife Fund, published *World Conservation Strategy.* This document, which presented worldwide strategies for the rational use of resources, has served as the basis for many national conservation plans. But many critics feared that, until the nations of the world were more willing to delegate greater authority to international organizations and to support them financially, little progress toward the solution of global problems could be expected. In existing conditions of international relations, this left each nation to attempt to do what it could within its own boundaries.

## Types of natural resources

In classifying natural resources it has been traditional to distinguish between those that are renewable and those that are nonrenewable. The former were once considered to be the living resources—*e.g.,* forests, wildlife, and the like—because of their ability to regenerate through reproduction. The latter were considered to be nonliving mineral or fuel resources, which, once used, did not replace themselves. In practice this separation is not entirely satisfactory, for reasons that will be dealt with in a later section. There are, nevertheless, certain aspects of conservation that apply specifically to nonliving resources:

1. Beneficiation is the upgrading of a resource that was once too uneconomical to develop. It usually depends upon technological improvements, such as those that make possible the concentration of a dispersed fuel or mineral so that it can be more easily handled, transported, or processed.
2. Maximization is the aggregate of those measures that avoid waste and increase the production of a resource.
3. Substitution involves the use of common resources in place of rare ones, as, for example, the use of aluminum in place of less abundant copper for a variety of products.
4. Allocation is the determination of the most appropriate use for a resource and the assignment of the resource to that purpose. In market economies allocation is usually controlled by the pricing mechanism: if the demand for a particular purpose is high, then the price of a resource to be used for that purpose will also be high; this high price will in turn make it more likely that the resource will be used mostly for that purpose. In government-controlled economies a resource may be reserved only for what are considered to be its most important uses.
5. Recycling, one of the most promising methods for conservation of mineral resources, involves the concentration of used or waste materials, their reprocessing (if this is required), and their subsequent reutilization in place of new materials. If carried out in an organized and consistent manner, recycling can greatly reduce the drain on supplies of minerals. It is also appropriate for products derived from living resources, such as the reuse of wood and paper as well as the reclamation of organic fertilizers from sewage.

Because all natural resources form a continuum, from those that are most renewable in the short term to those that are least renewable, they do not readily lend themselves to a single system of classification. It is useful, therefore, to examine the various types of natural resources in relation to their cycling time; *i.e.,* the length of time required to replace a given quantity of a resource that has been utilized with an equivalent quantity in a similarly useful form. From this point of view, renewable resources can be considered as those with short cycling times and nonrenewable resources as those with very long cycling times. Any resource can be nonrenewable, however, if the demand and rate of utilization exceed its cycling capacity.

The concept of cycling time

Two kinds of natural resources, pasture grass and coal, can be used to illustrate the concept of cycling time. When grass is grazed by livestock or mowed, a crop of it is removed. If provision is made to protect the fertility and structure of the soil and to leave enough seed or adequate roots and vegetative parts to produce new growth, then a

grass crop can be removed from a pasture each year for an indefinite period of time. Removal of one year's crop does not diminish the supply available for the next year if the land is cared for properly. The cycling time for this resource may be one year in areas in which climate limits growth, or it may be less than a year if growth can be continuous.

By contrast, the coal resources of the Earth were built up over millions of years. Most were laid down during the Carboniferous Period of geologic time (from 345,000,000 to 280,000,000 years ago), when climates were warm. Extensive swamp forests covered large areas of the Earth, and conditions were favourable for plant debris to accumulate in extensive deposits without decomposing and breaking down organically. Subsequently, heat and pressure generated by the deposition of other materials on top of the organic debris and by movements of the Earth's crust transformed the plant remains into coal. Organic debris is still being produced in swamps and marshes, and over millions of years this, too, could become transformed into coal. The time scale is so great, however, that, for human purposes, coal can be considered as a nonrenewable resource. Thus, only the supplies presently available in the Earth's crust can be counted on for future use.

## RENEWABLE RESOURCES

**Plants and animals.** The most clearly recognizable renewable resources are those consisting of, or produced by, living things. Agricultural crops, animal forage, forest crops, wild and domestic animals—all can continue to reproduce and regenerate their populations as long as environmental conditions remain favourable and an adequate seed source or breeding stock is maintained. Moreover, all can be cropped or harvested without diminishing their supply, provided that the cropping does not exceed the reproduction or growth rate. If it does, the resources will be depleted; and, if the rate of cropping continuously exceeds the rate of replacement or regrowth, the resource ceases to be renewable, and the species involved are reduced to the point of extinction. A renewable resource thus can be said to be "mined"—that is, it is removed at a rate that does not permit renewal. The renewability of a living resource is further endangered if the environment required by that resource is allowed to deteriorate or disappear. Sheep in a mountain pasture are a renewable resource only as long as the pasture produces vegetation that will nourish and support the sheep. If the pasture is overgrazed, the vegetation destroyed, and the soil eroded, sheep cease to be a renewable resource in that locality.

*Species renewability.* The renewability of a living resource varies with the species and with the areas involved. Thus, annual plants, from which a high percentage of cultivated field crops are derived, grow to maturity each year and then die back. They are annually renewable and can be cropped at a relatively high rate. Perennial plants, such as fast-growing poplar trees, may have a much slower rate of renewability, although this depends upon the purposes for which they are used. If seedlings are in demand, they can be cropped annually, and a new supply can be grown from protected seed sources. More realistically, however, the cycling time for these trees depends on the length of time required for them to grow to maturity and to produce seeds from which a new crop can be grown. Certain conifers, such as the Monterey pine, can reach a size adequate to yield useful timber and other wood products in less than 30 years. Other conifers, desirable because of the quality of their mature wood, may be cropped only on a 100-year cycle. Old-growth redwood trees can almost be considered a nonrenewable resource, because the time required to produce their equivalent may be from 500 to several thousand years, well beyond the limits for which people are prepared to plan. Redwood forests used for timber production are managed on the basis of a much shorter cycle and the cutting of younger trees. Such management does not provide for the replacement of 1,000-year-old specimens.

Renewability of annual and perennial plants

*Landscape renewability.* A distinction should also be made between renewable species and the communities or landscapes they occupy. Although it takes about 100 years to replace a mature coniferous tree, certain types of coniferous forests that are managed for timber production can be logged almost indefinitely if the annual level maintains a sustained yield. When all the interrelationships among such factors as soils and plant and animal life are considered, a natural forest that has not previously been disturbed by people may be far less renewable in its totality than the individual species within it. In other words, although the procedures employed in the harvesting of timber may assure a sustained yield, they may, nevertheless, be disruptive to other forms of life in the forest community, in which case species that are intolerant of such disturbance may disappear. It may be exceedingly difficult, therefore, to regrow a new community that resembles the original primitive forest if the area is to be disturbed periodically by timber cutting. It was this consideration, among others, that engendered the need to protect national parks, wilderness areas, and undisturbed research reserves. From certain viewpoints a wilderness is a nonrenewable resource; if it is seriously disrupted by human activity, it could require hundreds of years to recover its appearance and the various natural combinations of plant and animal life that contributed to its original wilderness value. The redwood-forest wilderness, as previously noted, could require many thousands of years for complete recovery following a major disturbance. Certain tropical rain forests that were disturbed more than 400 years ago still have not regained their original balance of species and do not resemble undisturbed, primary rain forest in the same region.

**Ecosystems.** Resources that contain a combination of interacting living and nonliving components are called ecosystems. It is impossible to separate an ecosystem into its living and nonliving components, because the whole constitutes a dynamic system in which there is a flow of energy from sunlight, gases from the atmosphere, and minerals and water from the soil. As a natural resource, soil, in turn, is also a combination of living and nonliving components: it consists of atmospheric gases, water, living and dead organic materials, and more or less finely divided mineral substances. Moreover, soil is a product of the interaction between the living and the nonliving environment. The living components of soil fit the definition of renewable resources, within the limitations that have been noted, and the mineral components fit the definition of nonrenewable resources. As long as the living components of soil remain healthy and continue to function, the mineral components are recycled from the soil, through the organic life within it (*e.g.*, bacteria and other microorganisms), and back to the soil following the decay and breakdown of dead organic materials. Because most forms of terrestrial life are dependent upon it for their continued existence, soil must be maintained in a renewable state. Mining soil, or using it in such a way that its fertility is exhausted and it is washed or blown away by too-rapid erosion, reduces the likelihood that life can continue to exist in the area affected.

Ecosystems defined as resources

**Solar energy.** The supply of solar energy represents an inexhaustible resource in relation to human time scales, and it is not affected by human activities. The potential lifetime of the Sun is in hundreds of millions of years, barring cosmic accidents, and throughout its lifetime the amount of energy reaching the Earth from the Sun could be capable of meeting all human needs. That energy supply, however, depends on the condition of the atmosphere, which can be affected by human activities. For example, scientists warn that even a limited nuclear war would darken the sky with smoke and other particulate matter, causing a marked drop in surface temperatures known as nuclear winter.

Solar energy can be captured directly, as in the space heating of buildings, the heating of fluids in solar collectors, or the conversion of that energy to electricity using photovoltaic cells. It can also be captured indirectly by powering the hydrologic cycle, thereby making solar energy available as water power, or by photosynthesis, thereby converting solar energy stored in plant tissues. Plants, in turn, contribute biomass, which can be converted to alcohol fuels or burned directly to provide heat. In the 1980s,

for example, wood fuel provided more of the energy used in the United States than did nuclear power. In developing countries, wood is often the major supplier of energy. In whatever form it is used, solar energy can be expected to play a growing role in meeting human energy needs.

**Water.** Water may also be considered an inexhaustible resource because the total supply of water in the biosphere is not affected by human activities. Water is not destroyed by human uses, although it may be held for a time in combination with other chemicals. To be useful, however, water must be in a particular place and of a certain quality, and so it must be regarded as a renewable, and often scarce, resource, with recycling times that depend on its location and use.

Water that falls from the atmosphere as various types of precipitation and then runs off the land surface to form streams and rivers that eventually reach the ocean generally operates on a one-year-renewable cycle known as the hydrologic cycle. From the ocean the water is evaporated by solar energy and returned to the atmosphere, from which it again falls as rain or some other form of precipitation. In certain locations, however, water has a much longer cycling time; after entering the ground from rainfall, it may percolate slowly through underground channels until it reaches underground reservoirs. In certain arid regions the total water supply may be underground water that accumulated during past ages, when the climate of the region was more humid. Since that time there may have been little or no addition to this supply because of the existing climatic conditions. Because its cycling time may be extremely long and dependent upon the frequency with which wet and dry climates alternate in a particular region, such a water resource can be virtually nonrenewable.

Hydrologic cycle as an annual renewable cycle

**Air.** Air is also an inexhaustible resource in the sense that the uses made of it have little effect on its total quantity. The quality of air, however, as measured by its chemical composition or physical state, is subject to human interference. For life to exist on Earth there must be a proper balance among the nitrogen, oxygen, carbon dioxide, water vapour, and other components of the atmosphere. A layer of the gas ozone, for example, must be maintained in the upper atmosphere to screen out damaging ultraviolet light from the Sun. The accumulation of toxic materials in the air must be kept to a minimum, and the concentration of solid and liquid particles in the atmosphere must not be allowed to reach a level that interferes with the flux of solar radiation. All of these factors are affected by human activity and by the effects of this activity on other forms of life.

## NONRENEWABLE RESOURCES

As noted above, renewable resources include resources with widely different cycling times, some so long as to make the resources essentially nonrenewable. Fossil and nuclear fuels and minerals also exhibit a wide range of properties that affect their management. Fossil fuels, such as coal and petroleum, are the least renewable of such resources because they are effectively exhausted by use and because their rate of formation is exceedingly slow. Most minerals, on the other hand, are not destroyed by use; thus, in a sense, they are renewable and inexhaustible because they can be recycled for further use. But useful supplies of these minerals in accessible locations are exhaustible, and thus they are nonrenewable for human purposes.

**Fossil fuels.** Fossil fuels are those organic materials that have been converted from their original form by physical and chemical processes within the Earth's crust into a solid mineral state (coal), a liquid (petroleum), or a gas (natural gas). If these substances are completely burned (oxidized) when used as fuel, the end products are carbon dioxide, water, and heat energy. These cannot be reconstituted into organic substances without either elaborate synthesis in a chemical laboratory or the natural photosynthetic processes of green plants. Thus, burning destroys fossil fuels as useful energy sources. Fossil fuels are also used for purposes other than fuel. Coal and petroleum are used industrially for the manufacture of a wide variety of carbon-containing materials, such as plastics, synthetic fibres, medicines, and food.

On the basis of existing knowledge of the amount of fossil fuels in the Earth's crust, it has been predicted that the availability of petroleum and natural gas could be greatly depleted within a century if used at the rates anticipated. Although coal supplies are greater, projected rates of use indicate that they cannot be expected to be easily available for more than a few centuries. These predictions can be changed, of course, if rates of use change, which is expected to happen with, for instance, the further development of renewable energy resources.

**Nuclear fuels.** Although nuclear fuels are inorganic substances, like fossil fuels they are destroyed when they are used in the production of heat energy. Unlike fossil fuels, however, they are also destroyed by spontaneous disintegration, through natural radioactivity. Uranium, for example, ultimately changes to lead, but the rate of change is very slow; its half-life (the length of time it takes for half the atoms of a given amount of a radioactive substance to disintegrate) is 7,600,000,000 years. Of the naturally occurring nuclear fuels, uranium and thorium, only uranium-235 can be used directly in a nuclear reactor for the production of power. The more common form of uranium, uranium-238, must be converted into plutonium before it can be used as a nuclear fuel. Similarly, thorium must be transformed into uranium-233 before it is usable for fuel purposes.

Uranium and thorium

Although supplies of uranium and thorium are relatively abundant, they are exhaustible and nonrenewable. By replacing nuclear fission with nuclear fusion as a power source, however, deuterium (an isotope of hydrogen) can be used as the fuel. Because it occurs in substantial quantities in seawater, deuterium is practically an inexhaustible resource for human purposes. However, many technological difficulties remain to be overcome before nuclear fusion can be used as a commercial supply of energy.

**Minerals.** Certain minerals, such as iron and aluminum, are so widely distributed throughout the crust of the Earth that the amounts exceed any foreseeable human needs. Other minerals, such as the precious metals (*e.g.*, gold, platinum, silver), are much more limited in their distribution and quantity. The usefulness of a mineral, however, depends upon its accessibility and concentration; therefore, minerals that are highly dispersed throughout the Earth are essentially unavailable, even though their total quantity may be great.

*Ores and reserves.* Most efforts to obtain minerals are directed toward finding mineral ores, which are deposits in which the concentration and quantity of a mineral are such that it can be extracted profitably. Mineral reserves are those that are known to exist or can reasonably be inferred to exist from geologic evidence. It follows that certain deposits of minerals that are not now considered either ores or reserves may become so if the technology for their extraction improves, if the supplies of energy available for their extraction increases, or if their economic value increases.

It is known from past experience that ore deposits can be exhausted. Because the gold mines around Virginia City, Nev., and the tin mines of Cornwall no longer can produce significant quantities of minerals, they can be considered virtually exhausted resources. The rich iron ores of the Mesabi Range in Minnesota have largely been depleted, and mining activity in this area has shifted to lower grade iron-ore deposits. The supplies available from any mineral deposit, at least on dry land, are exhaustible and nonrenewable because the geologic processes that led to the formation of those deposits operate slowly and over long periods of time.

Some mineral deposits, however, are renewable. Manganese ore, for example, is relatively scarce on dry land but is continuously formed in nodules on the ocean floor, as are cobalt, nickel, and copper. The rate at which the nodules of manganese, cobalt, and nickel are growing through chemical precipitation from seawater currently exceeds the rate at which these minerals are being used. Although the nodules are not yet being collected, the technology for doing so is being developed. Then these metals will be considered as renewable natural resources only so long as the rate of use does not exceed the rate of formation.

Oceanic mineral deposits

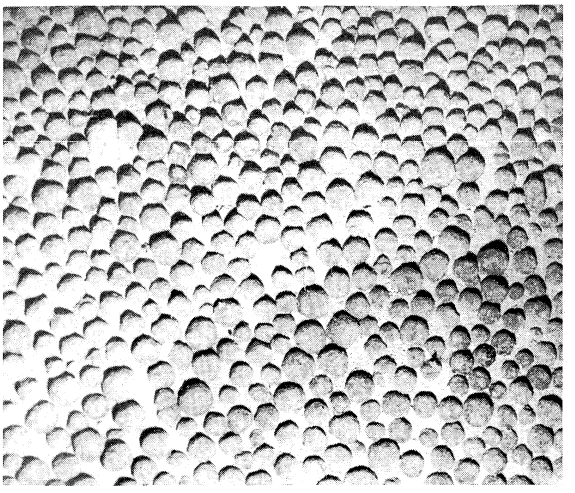


Figure 1: Manganese nodules on the southern Pacific Ocean floor.

By courtesy of the Lamont-Doherty Geological Observatory, Columbia University

*Recycling of minerals.* The use of most metals does not destroy them, although rusting may reduce their quantities by a small amount when they are in use. As commercial products, some metals are found in such large quantities in urban areas that their new concentration may exceed that which existed while they were in the ground. Cities, therefore, may be considered as ore deposits for certain minerals. At present it is cheaper to mine new ores than to recycle used or waste metals (with some exceptions, such as aluminum), but this economic balance does not take into account the cost of disposing of the metallic wastes that accumulate in urban regions. Thus, it is likely that sometime in the future many metals now considered as exhaustible, nonrenewable resources will be treated as recyclable resources.

Not all minerals can be recycled under existing conditions. The concentrated phosphates that are used in fertilizers and detergents, for example, are dispersed widely over the farmlands and waters of the Earth, from which they enter the life cycles of various organisms and eventually, through erosion or in wastes, reach the oceans. Because these phosphates are virtually irretrievable and because the rate at which available reserves are being used probably exceeds the rates at which new reserves are formed, such minerals are considered as nonrenewable and exhaustible resources.

## Management of natural resources

### MANAGING NONLIVING RESOURCES

**Soils.** *Formation of soil.* Soils are the basis of support for most terrestrial life and a source of nutrients for freshwater and marine life. As noted above, soil is formed over time as the result of interaction between the living and the nonliving environment—climate, organisms, and the physical surface of the Earth. Rocks are broken apart by the action of sunlight, wind, rain, snow, sleet, and ice. With the aid of wind and water movements as well as gravity, rock particles from high elevations are deposited on mountain slopes or in valleys, where they are further acted upon by the local climate, by plant and animal life, and by such other environmental factors as fire until they become soil. Nitrogen from the atmosphere, formed into nitrates by the action of lightning and atmospheric water vapour, may enter the soil with rainfall. Other nitrates may be added by the action of such living organisms as soil bacteria and various algae that can convert atmospheric nitrogen into the nitrates required for plant growth. These and other chemicals in the soil eventually become part of the living tissue in plants and animals. The chemicals are returned to the soil as organic wastes and litter that form humus, which is partly decomposed organic material. As humus continues to decompose, the chemicals within it enter the soil for further use by plants and animals.

Soils vary from place to place depending upon the rocks and minerals from which they are derived, the nature of the local climate, and the kinds of organisms that live in or on them, as well as the amount of time that these factors have been operating. Developmental soils—*i.e.,* those still being modified by climate and organisms—reveal the nature of the parent materials from which they are derived; mature soils, those that have achieved a balance among the various forces operating on them, show in particular the influence of the climate and vegetation in which they develop. Soils also differ greatly in their inherent fertility and in their ability to support life. Those derived from quartz sand, for example, may be naturally deficient in calcium, magnesium, and other elements essential to plant growth. The surface layers of those soils developed in humid, forested regions are often heavily leached, as rainwater containing weak organic acids percolates through them and dissolves the more soluble minerals.

Developmental and mature soils

Because soils are essential for such purposes as growing crops, forage, and timber, it is important that they not be allowed to wash or blow away more rapidly than they can be regenerated, that their mineral fertility not be exhausted, and that their physical structure remain suited to the continued production of desired plant materials. The objective of soil management, therefore, is to keep soil in place and in a state favourable to its highest possible productive capacity.

*Soil erosion.* In the past and, to a considerable degree even now, soils have not been managed effectively. Those exposed through cultivation to the erosive effects of wind have been blown away; those laid bare on sloping ground have been washed downhill by rainfall. Although soil erosion has long been recognized as a major conservation problem, erosion as such—and its converse, the deposition of eroded soil particles—is not a problem but a normal and natural process leading to both soil development and maintenance. Soils exist only because of past erosion and deposition. The conservation problem involved in soil erosion is the accelerated erosion that occurs when soil cover in the form of living or dead plant material is removed. In such cases the soil then erodes at a rate faster than it can be replaced by normal deposition of particles on the soil surface or by the breakdown of rocks and minerals. In severe cases, such erosion leads to the formation of deep gullies that cut into the soil and then spread and grow until all the soil is removed from the sloping ground. Under severe wind action, the finer particles of surface soil are blown away and form drifts and dunes, leaving only the coarser sands and gravels on the soil surface.

Erosion of organic soil cover

Although measures to stop soil erosion are now used in most technologically advanced countries, the problem remains a major one. It is particularly severe in the tropics, where high rainfall and steeply sloping ground favour the rapid loss of any soil exposed by agriculture, and around the edges of the world's deserts, where destruction of natural plant cover by cultivation or livestock grazing causes soil loss through wind action and the spread of desert-like conditions.

To prevent wind erosion, shelter belts of trees have been planted to break the force of the wind. The practice of covering soils with plant litter (mulch) when they are not actually covered with growing plants also helps to hold them in place. Cultivating at right angles to the direction of the wind further serves to prevent wind erosion.

Water erosion on sloping ground may be prevented by terracing on steep slopes or by contour cultivation on gentler slopes. In the latter a slope is plowed along horizontal lines of equal elevation. Strip-cropping, in which a close-growing crop is alternated with one that leaves a considerable amount of exposed ground, is another technique for reducing water erosion; the soil washed from the bare areas is held by the closer growing vegetation. In the tropics maintaining a tree shelter over the ground serves as a means for breaking the force of raindrops, thus reducing their erosive power, and also to screen out direct sunlight. In addition to causing damage to certain crops, sunlight can accelerate the breakdown of organic materials in the soil at a rate that is faster than is desirable.

*Soil fertility.* A conservation problem equally as important as that of soil erosion is the loss of soil fertility.

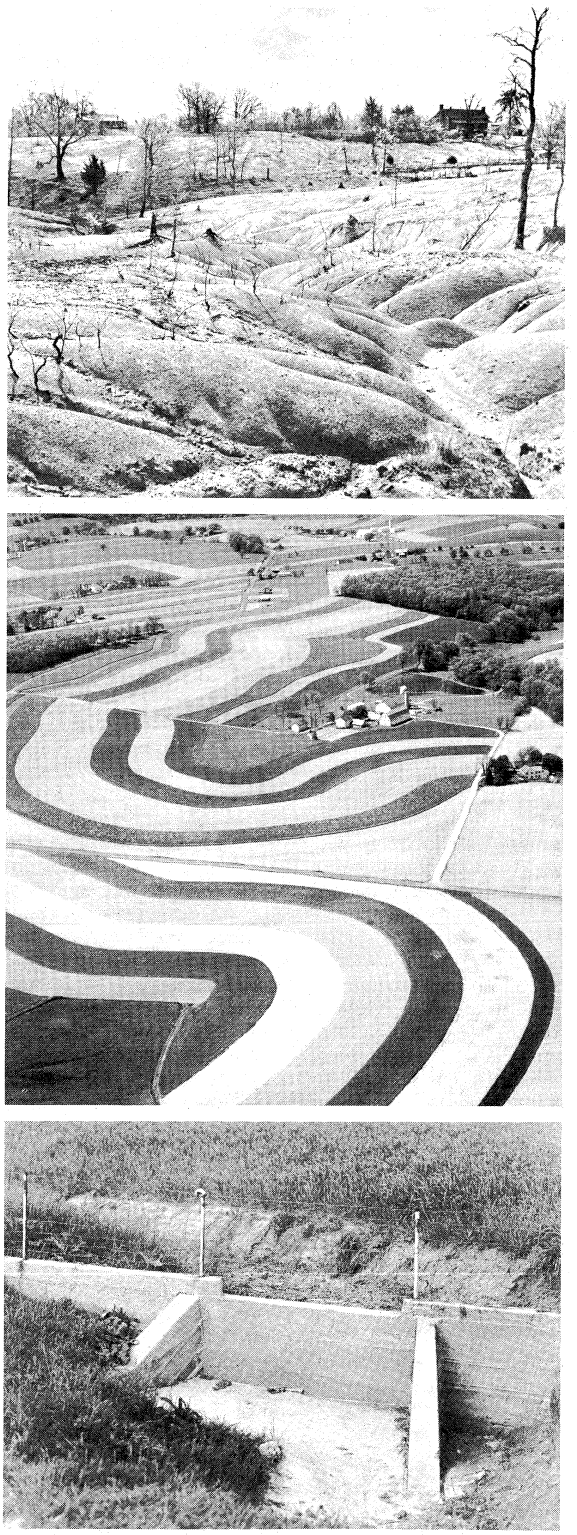


Figure 2: *Soil erosion and ways to prevent it.*
(Top) Eroded farmland in west Tennessee, resulting from tenant–absentee-owner relationship. (Centre) Contour farming in Pennsylvania. (Bottom) Concrete dam built to protect field from erosion at Delavan, Ill.
By courtesy of (top) the U.S. Department of Agriculture; photographs, (centre) Grant Heilman, (bottom) J.C. Allen and Son

Most agriculture was originally supported by the natural fertility of the soil; and, in areas in which soils were deep and rich in minerals, farming could be carried on for many years without the return of any nutrients to the soil other than those supplied through the natural breakdown of plant and animal wastes. In river basins, such as that of the Nile, annual flooding deposited a rich layer of silt over the soil, thus restoring its fertility. In areas of active volcanism, such as Hawaii, soil fertility has been renewed by the periodic deposition of volcanic ash. In other areas, however, natural fertility has been quickly exhausted. This is true of most forest soils, particularly those in the humid tropics. Because continued cropping in such areas caused a rapid decline in fertility and therefore in crop yields, fertility could be restored only by abandoning the areas and allowing the natural forest vegetation to return. Over a period of time the soil surface would be rejuvenated by parent materials, new circulation channels would form deep in the soil, and the deposition of forest debris would restore minerals to the topsoil. Primitive agriculture in such forests was of a shifting nature: areas were cleared of trees and the woody material burned to add ash to the soil; after a few years of farming, the plots would be abandoned and new sites cleared. As long as populations were sparse in relation to the area of forest land, such agricultural methods did little harm. They could not, however, support dense populations or produce large quantities of surplus foods.

Starting with the most easily depleted soils, which were also the easiest to farm, the practice of using various fertilizers was developed. The earliest fertilizers were manures, but later larger yields were obtained by adding balanced combinations of those nutrients (*e.g.*, potassium, nitrogen, phosphorus, and calcium) that crop plants require in greatest quantity. Because high yields are essential, most modern agriculture depends upon the continued addition of chemical fertilizers to the soil. Usually these substances are added in mineral form, but nitrogen is often added as urea, an organic compound.

Fertilizers and crop rotation

Early in agricultural history it was found that growing the same crop year after year in a particular plot of ground not only caused undesirable changes in the physical structure of the soil but also drained the soil of its nutrients. The practice of crop rotation was discovered to be a useful way to maintain the condition of the soil and also to prevent the buildup of those insects and other pests that are attracted to a particular kind of crop. In rotation systems a grain crop is often grown the first year, followed by a leafy-vegetable crop in the second year and a pasture crop in the third. The last usually contains legumes, because such plants can restore nitrogen to the soil through the action of bacteria that live in nodules on their roots.

*Salinization of soil.* In irrigation agriculture, in which water is brought in to supply the needs of crops in an area with insufficient rainfall, a particular soil-management problem that develops is the salinization (concentration of salts) of the surface soil. This most commonly results from inadequate drainage of the irrigated land; because the water cannot flow freely, it evaporates, and the salts dissolved in the water are left on the surface of the soil. Even though the water does not contain a large concentration of dissolved salts, the accumulation over the years can be significant enough to make the soil unsuitable for crop production. Effective drainage solves the problem; in many cases, drainage canals must be constructed and drainage tiles must be laid beneath the surface of the soil. Drainage also requires the availability of an excess of water to flush the salts from the surface soil. In certain heavy soils with poor drainage, this problem can be quite severe; for example, large areas of formerly irrigated land in the Indus basin, in the Tigris–Euphrates region, in the Nile Basin, and in the western United States have been seriously damaged by salinization.

*Watershed soil.* The soils of wildlands in all areas that yield water to streams and rivers are as important as are those in which agricultural crops are raised. If these soils are kept in place and in good condition, they support trees, forage, and wild animal life and yield clear water for human use. If, however, these soils are damaged physically by being compacted or are allowed to erode, they lose not only their capacity to support vegetation but also their capacity to hold and slowly yield useful water to streams or springs. Furthermore, eroded soils cause siltation of waterways; they also accumulate in lakes and reservoirs, filling them and thereby reducing the useful purposes that these bodies of water serve.

In order to maintain soils in watershed areas, their vegetative cover must be retained. This requires avoiding excessive disturbance of forest vegetation and soil cover through logging, avoiding excessive grazing and trampling of pasture lands, and preventing the kinds of wildland fires that destroy plant cover and expose the soil. Often, the success of intensive land and water use downstream depends on the care taken of the soils in the less intensively used forest or range watersheds upstream.

**Water.** Life originated in the oceans, and the chemical composition of body fluids in land animals reflects their primeval origin. The dependence of life on water is complete; it is the major constituent of plant and animal cells. Most of the major groups of animals still live in water; a relatively small number have adapted to life on dry land.

Dependence of life on water

*Uses of water.* Water is required for a variety of purposes; water for drinking is still paramount, and such water must be relatively pure. If it is not supplied in sufficient amounts through precipitation, it must be supplemented by irrigation systems. Irrigation, however, is one of the most wasteful uses of water in areas in which it is scarce, because great quantities are lost through evaporation in both storage areas and transport. In many regions irrigation is, nevertheless, essential for human survival.

Water for transportation has always been important, as indicated by the fact that most major cities are located on the shores of oceans and other large bodies of water or along rivers and other types of navigable waterways. Despite recent advances in ground and air transportation, water transportation has an economic advantage for the movement of goods that have a relatively low value per unit of weight or volume, such as raw mineral ores, fuels, and various types of construction materials. Water for urban use other than drinking serves a multitude of purposes, such as fire fighting, street cleaning, sanitation, and sewage disposal. Steel mills, pulp mills, chemical factories, and most other industrial processes that involve the conversion of raw materials into finished products require water. Next to agriculture, one of the most extravagant uses of water is as a cooling fluid in the generation of power from fossil and nuclear fuels, with the latter consuming far greater volumes. Water has been used directly as a source of power since the time of the first boat and the first waterwheel. A small but important part of the world's electrical supply now is generated by hydropower, in which the force of falling water is used to turn turbines that produce electricity.

*Husbandry of water supplies.* Although water is a renewable resource, the many demands for water of a desired quantity and quality in a particular place require careful husbandry of the supply. After reaching the surface of the Earth as rain, water enters a supply system either by penetrating the ground and moving through subsurface channels, known as aquifers, or through runoff into streams and rivers. As mentioned above, the supply and quality of water depend in part on the management of the vegetation and soil in the watershed areas. Also involved is the control of streamflow or the control of pumping from underground sources. In many parts of the world where rainfall is seasonal, streams run at flood levels during the wet season but are extremely low or completely dry at other times of the year. River-basin-management techniques attempt to equalize this variable supply for human purposes, in part through watershed management and in part through the capture of water by dams and its storage in reservoirs.

Consequences of mismanagement of water

When water is mismanaged, a high percentage is lost through evaporation in watersheds. Moreover, as a result of poor management of watersheds, the seasonality of water flow is more acute: floods that destroy lands in the river basins become more frequent during the wet season, and there is an increase in the frequency of droughts during the dry or low-rainfall season. Soil eroded from watersheds impairs the functioning of dams, reservoirs, and other structures downstream. Furthermore, because of mismanagement water becomes polluted at various stages in its movement from atmosphere to land and thence to the oceans.

Effective water management starts when precipitation first reaches the ground, after which the quality and quantity of water must be protected at every critical point along the hydrologic cycle. Hence, although it may have been practical to use streams, lakes, and the oceans as dumping areas when populations were low and water was abundant, this practice becomes untenable when populations increase and supplies of water decrease relative to human needs. Except in sparsely populated areas of the world, population and technological growth have made necessary the prevention of erosion, the recycling of wastes so that nutrients are restored to the soils and useful minerals are reclaimed for reuse, and the reuse of water to the maximum possible degree. In many dry areas the wasteful use of groundwater, stored in aquifers, has caused a serious lowering of water tables and even a sinking of the land surface. Such cases call for limiting the use of water supplies to more essential and less wasteful purposes.

*"New" sources of water.* Lack of water of proper quality and quantity has been a major factor affecting urban and industrial growth. To overcome this problem, water has been transported great distances—*e.g.,* the channeling of Rocky Mountain water from the Colorado River to Tucson, Ariz. During the 1970s and 1980s the Soviet Union proposed several projects to reverse or divert the waters of northward-flowing rivers of Siberia and the Russian S.F.S.R. to meet the demands of the more heavily populated and water-short regions of the Volga Basin, Central Asia, and Kazakhstan. The predicted environmental and climatic consequences of such undertakings, however, combined with their engineering logistics, prevented the practical application of most of these plans.

The use of the oceans as sources of fresh water is being developed in many areas. Kuwait, a desert nation in Arabia, now receives much of its water supply through the desalinization of seawater, as do a number of small communities and several large urban centres elsewhere in the world. Seawater may be used as a source of fresh water on a more widespread basis if an additional power source—*e.g.,* solar power—can be developed for the desalinization process. Morever, the materials reclaimed from seawater could, if power is available for their separation and concentration, help in meeting many of the world's mineral needs. It seems unlikely, however, at least with foreseeable sources of power, that desalinized ocean water will be extensively pumped to inland regions. Meeting the growing needs of such areas will require the purification of waters polluted by urban or industrial use or of waters that have become salinized through their use in irrigation. The reuse of such waters could go far toward reducing the need for new water by inland communities.

Desalinization of seawater

**Air.** Concern about the quality of air is a relatively recent development, although polluted air has been a problem for urban communities over many centuries. The atmosphere makes possible the existence of life, and in the solar system only the planet Earth has an atmosphere capable of sustaining known forms of life. Seventy-eight percent of the Earth's atmosphere consists of nitrogen, a gas that, combined with other elements, is a key ingredient of plant and animal protein. Most essential to terrestrial life is oxygen, which makes up nearly 21 percent of the atmosphere. The remaining 1 percent includes argon (an inert gas), carbon dioxide, and water. The latter two, despite their small quantities, are vital to life. Sunlight acting on water is the driving force for the world's weather. Carbon dioxide is essential to plant growth and thus to animal life.

The supply of oxygen in the atmosphere has been relatively constant during recent centuries, but its presence is believed to depend on the activities of living organisms. Green plants, through photosynthesis, and photosynthetic microorganisms built up the present level of atmospheric oxygen and provide for its continued maintenance.

The atmosphere extends above the surface of the Earth to a distance of many thousands of miles, but 95 percent of its total mass is to be found within 12 miles of the surface. The layer of atmosphere closest to the surface (extending upward to heights of four to 11 miles) is known as the troposphere. It is only within this zone that life is supported. It is also where water vapour, and hence storms and precipitation, occurs. Above the troposphere

is the stratosphere, extending up to 31 miles above the Earth. Most important to life is the ozone layer within the stratosphere. Ozone ($O_3$) is a form of oxygen that blocks out the shorter wavelengths of solar radiation and thus protects life on the Earth's surface from the damaging effects of ultraviolet light.

There is not much that can be done to manage the atmosphere. What must be done, however, is to manage those activities that affect its composition (see below *The pollution of natural resources: Air pollution*).

**Subsurface deposits.** Several types of conservation activities are associated with the use of fossil fuels and minerals. First are those that involve making the available fuel and mineral reserves serve the most worthwhile purposes for the longest period of time. Second are the problems associated with extracting fuels and minerals from the ground and their subsequent transport, processing, and manufacture, all of which can directly affect the quantity or quality of other resources or the general environment. Third are the side effects, which, because they involve pollution, are considered below.

*Conserving exhaustible fuel and mineral resources.* As has already been noted, fossil fuels and minerals fall into two categories: those that are destroyed by use and those that retain their physical and chemical characteristics during use and are capable of being reclaimed for other uses. The outlook is most bleak with respect to the conservation of those fuels and minerals that are destroyed by use. If petroleum continues to be used at its present rate, the supply will be exhausted or reduced to a level at which further use may be restricted by the high cost of the remaining resource. Because the situation is generally recognized, conservation need involve only allocating the resource to those uses for which it is best suited and restricting its use in those cases in which other materials, less likely to be exhausted, can be equally well substituted. Thereafter, it is necessary to maximize the available resource. This can be done by avoiding waste in its extraction, transportation, and processing and by utilizing fully all that can be made available. Thus, several oil companies no longer pump competitively from the same oil field, leaving only unobtainable oil in the ground. Instead, fields have been unitized; companies now cooperate in bringing out oil by using only strategically located pumping units. The practice of burning the natural gas from oil fields has been replaced by tapping and piping the gas for use as a fuel. Unrestrained gushers are a thing of the past; more important are the methods of drilling and of injecting gas or water under pressure to force out oil that previously would have been left in the ground. Improved methods for refining petroleum have eliminated much of the waste that once occurred in this process. Yet accidental losses still do occur. The leaking well that caused such extensive damage to wildlife and recreational resources in the Santa Barbara, Calif., channel in 1969 resulted in part from failure to use available safety devices. Similar failures were involved in an oil well off the Louisiana coast in 1970 and in the blowout of the Ixtoc 1 oil well in the Bay of Campeche, which spilled oil on the coasts of Mexico and Texas in 1979. Great quantities of oil are still wasted and waters polluted because of improper navigation of oil tankers and because some such tankers are so poorly constructed that excessive quantities of oil can escape as a result of relatively minor damage to the hull.

Maximizing available resources

Beneficiation, the concentrating of relatively dispersed or low-grade resources into a form in which they can be handled economically, is another means of extending the supply of exhaustible petroleum resources. There are, for example, great quantities of petroleum available in oil shales and tar sands in various parts of the world. As long as petroleum is plentiful in a more concentrated form in oil fields, it is not economical to extract the more dispersed petroleum in such shales and sands. With the prospect of an eventual shortage of petroleum, however, techniques have been devised for concentrating these lower grade supplies. The technologies of extracting oil from petroliferous sand or shale differ, but both involve the heating of the extracted oil-bearing medium. Most oil sand has been extracted by surface mining techniques; the extracted sand is then heated or coked, and the distillates are transformed into synthetic crude oil by high-pressure hydrogenation. Oil shale has been generally mined, crushed, and heated or retorted at high temperatures to produce crude oil.

Conservation in the case of recyclable minerals involves reuse. For this to be accomplished, incentives and methods may be necessary to encourage the gathering of used materials for reprocessing. Provisions have been made for the collection of junked automobiles, waste cans, bottles, and other containers and for the reuse of building materials from demolished structures. Although much of this is done more to prevent pollution than to reclaim the materials, it does serve both purposes. Another conservation measure is waste-processing technology and the growth of industries concerned with the recycling of wastes.

Recycling

Just as applicable to the conservation of minerals as they are to the conservation of petroleum are such other techniques as the allocation of scarce resources to their most essential uses, the substitution of other resources for those that have become scarce, and the maximization of supplies. Moreover, there has been great progress toward the development of composite materials, in which relatively small amounts of metal are used in combination with plastics and ceramics. Metallic alloys that minimize the need for scarce materials (as in the substitution for copper or silver in coins of various cheaper metallic alloys and the substitution of aluminum alloys for copper or steel wire, for roofing materials, or for tin in cans) are in many cases more effective than those used for the same purposes. Beneficiation has been used increasingly; thus, low-grade taconite and jasperite ores are now being mined and pulverized, and their iron components separated and pelleted before being shipped to smelters. Such techniques make possible the use of iron deposits that previously would not have been considered economical sources of ores.

Because of the many ways for making better use and reuse of the available supply of metals and other minerals and because the extent of new discoveries cannot be foretold, it is difficult to predict the likelihood of depleting any particular mineral resource; for example, the wasteful use of metals and minerals in most affluent societies creates a drain upon supplies that is entirely unnecessary. Predictions of resource use based on the continuation of wasteful practices are likely to prove unrealistic if these practices are changed voluntarily or forced to change by a scarcity of materials.

Predictions of depletion of a resource

*Conservation problems caused by mining.* Some of the most serious conservation problems are associated not with the use of minerals or fuels but with the methods used to extract these resources. Mining for coal has created widespread devastation in the Appalachian Mountains of the United States, in Bohemia (Czechoslovakia), and, in the past, in England and Germany because deposits in these regions are found near the surface. In the removal of the coal, soil and all living resources have been destroyed, leaving behind a barren, denuded, and eroded wasteland. Because mining wastes often contain much sulfur, their watery runoff contains sulfuric acid, which destroys aquatic life in streams. Similar problems are associated with the surface mining of nickel in such places as New Caledonia and Australia. Dredging for tin in Malaysia has also created widespread conservation problems. The mining of phosphates has brought destruction to many Pacific islands, in particular to the island of Nauru. Titanium mining in the sands along the Australian coast has destroyed natural vegetation on beaches, sand islands, and dunes and opened pathways for dune movement that destroys still greater areas of vegetation. Dredging for gold has damaged many streambeds and riverbeds because it is destructive to aquatic life and water quality.

The means for preventing damage by surface mining do not exist, but damage can be controlled and minimized, in part by the creation of erosion-control structures and the prevention of water pollution. The rehabilitation of mined-over land may involve, among other things, the removal and safe storage of topsoil before mining begins and its restoration after mining operations have been completed. It may also involve the shaping of mining wastes into

landforms that can be covered with soil, replanted, and revegetated. Of equal importance, however, is the decision not to mine areas that have high surface-resource values—*e.g.,* highly productive forest or farmlands, certain urban areas, and areas with natural qualities of scenery, plant life, or wildlife that make them suitable to be maintained as parks or reserves. The ores in such areas are not mined unless the need for their minerals is most urgent.

## MANAGING LIVING RESOURCES

Any area of land and water not yet modified by mankind can be managed in a number of ways. Certain choices must be exercised early, however, because even minor deterioration of an area through unplanned use may make it unusable for its intended purposes. Other choices, particularly those involving modification of the natural features of the landscape, may be exercised at virtually any time; any previous changes in the character of the living resources in a region to be modified have no effect on the ultimate structural modification of that region. In planning for the use of any undeveloped area, therefore, those purposes that have the most exacting requirements must be considered first; they often depend upon the continuance of relatively undisturbed conditions as well as upon the maintenance of the full variety of wild species and the natural environment within the area.

**Natural communities.** The idea that biologic communities should be protected for their own intrinsic value is of relatively recent origin. Although natural communities have been protected since ancient times, the reasons for doing so have not been related to the value of the community per se but to some special feature that was of value to people. Thus, hunting preserves were protected in ancient Mesopotamia, in China, and in England, where the New Forest was set aside by William the Conqueror. While such preserves protected natural areas, their major purpose was to provide a setting for royal hunting. Temple gardens have been preserved over the centuries in China and Japan; the cedars of Lebanon were maintained around holy places. But, again, such preservation was fortuitous rather than intentional.

The idea of preserving wild areas for their own value had its origin in the United States with Catlin, Thoreau, Muir, and others of similar mind. Only more slowly has this concept been accepted in other countries. Appreciation of wild nature is acquired along with scientific knowledge, particularly ecological knowledge. It is a sophisticated taste not usually to be found among those who earn their livelihood in close contact with the wild. Nevertheless, the concept of preserving wild nature for its own value has become widely accepted, although the means for implementing it are not necessarily available in every country.

Value of natural communities

Natural communities, little affected by human activities, are thought to be worth preserving for a variety of reasons. First, perhaps, is the scientific benefit to be derived from studying them, particularly concerning the functioning of the biosphere. From studies of undisturbed ecosystems much can be learned about the behaviour of those systems modified for the production of useful materials. Also, the value of wild species has been little explored; in their totality they are known to be essential to the function of the biosphere, but the importance of individual species is little understood.

Past experience has demonstrated that wild species of little apparent value may prove to be of major importance to medical research and human health. Sea urchins, for example, are used in studies of embryology; nonhuman primates (*e.g.,* monkeys and apes) are used for many studies of human functions and diseases; armadillos are used to study leprosy; and a great variety of wild plants are used as sources of drugs and medicines. Much of the knowledge of population growth and social behaviour under various conditions of crowding has come from the study of wild mammals.

Furthermore, it is known that more or less undisturbed natural communities are important to the continued operation of those systems that people have created. Watershed forests are protected so as to maintain streamflow and to avoid siltation of reservoirs; estuaries are protected so as to guarantee the continued production of forms of marine life important for food or other purposes. Finally, there are aesthetic and recreational values attached to wild areas and wildlife. It would appear that outdoor activities in a natural setting or contact with plants and animals in a wild state are important to psychological well-being, because people of all races and cultures seek such experiences when they achieve the affluence that enables them to do so.

Protection of natural communities

Biologic communities can be protected in a variety of ways, depending upon the desired objectives. The most difficult and exacting task is the protection of unmodified natural communities, with their full array of wild species, for use in scientific research. Because such communities are becoming increasingly rare, the major efforts to protect them are undertaken at an international level, as well as at the national level. The International Biological Program, a worldwide research effort, has focused attention on the many kinds of natural communities that require protection. The International Union for Conservation of Nature and Natural Resources, a semigovernmental international agency, devotes an important part of its activities to the establishment of reserves and parks for the protection of natural communities. The United Nations, through its Food and Agricultural Organization (FAO), UNEP, and UNESCO, has contributed to the establishment of many parks and reserves in developing nations. Yet, despite such activities, certain kinds of natural communities will be irrevocably lost unless there is greater effort toward their conservation.

The danger to natural communities and wild species comes from many causes. One long-standing problem derives from the exploitation of wild species that have commercial value, as, for example, the widespread removal of

By courtesy of (right) the U.S. Department of Agriculture, Soil Conservation Service; photograph (left) Grant Heilman

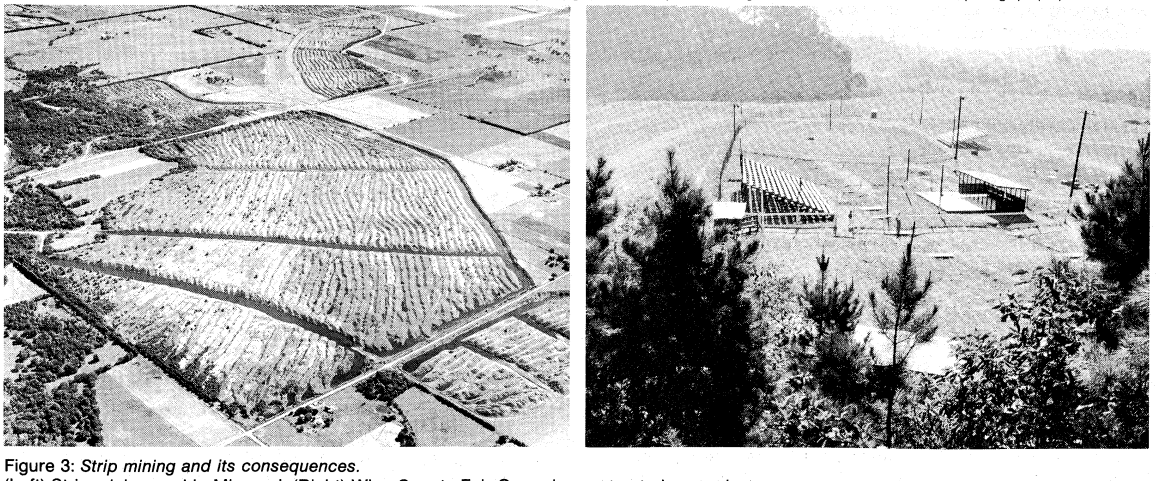


Figure 3: *Strip mining and its consequences.*
(Left) Strip-mining coal in Missouri. (Right) Wise County Fair Grounds constructed on coal strip mine spoil in Virginia.

mahogany trees and other valuable timber from the forests around the Caribbean area. The uncontrolled hunting of whales and other sea mammals has brought some species to the point of near-extinction, and at least one, Steller's sea cow, has been exterminated. The demand for such high-fashion products as shoes and handbags made from crocodile skins or coats and wraps made from the skins of tigers and leopards has endangered the continuing existence of these animals.

Control of those species considered to be detrimental has led to an unreasonable warfare against predatory animals. As a result, the wolf, cougar, lion, lynx, eagles and hawks of various species, and other carnivorous animals have been eliminated from the vicinity of human settlements or from pastoral lands, even though the evidence that some of these species do any appreciable damage is not well substantiated. But the greatest single cause for the depletion of natural communities and wild species has been the desire to use land for more productive purposes. This has led to extensive clearing of forests and woodlands, burning of vegetation, and the cultivation of previously undisturbed land for crop production. Many of the lands that are cleared eventually prove to be poorly suited for the purposes for which they are intended and are ultimately abandoned; however, it is then no longer possible to protect the natural communities that previously existed in these areas.

Programs for the protection of natural communities must involve, first, rational planning for the use of land and control over its exploitation by agencies charged with such responsibilities. Adaptation of land use for commodity production by using only those sites best suited for such purposes is a first step toward protecting other lands that are now being cleared unwisely. Although the establishment and proper management of parks and reserves permits the survival of certain species in certain areas, a more general program of rational management and use of all lands and species is essential to the long-term survival of wild nature throughout the world.

*Strict nature reserves.* The decision to maintain an area in a more or less unmodified condition usually is determined by its overall scientific and aesthetic value. It may also be determined by the contributions the protected area can make to the region as a whole, as, for example, the regulation of water yield and streamflow or perhaps as a reservoir of wild species that subsequently can move or be moved into other modified regions. The most restrictive category of land use is that of the scientific reserve, which is also known as a strict nature reserve. Such areas may be selected because of their unique geologic or biologic features. Or they may be selected because they are representative of widespread biologic communities that will be transformed elsewhere for the needs of commodity production, recreational use, or other more intensive purposes. Retaining representative areas in an unmodified condition provides standards by which the health and productivity of the modified sites can be tested.

Strict nature reserves and other types of scientific reserves often occupy only a relatively small area of land. The amount of land occupied, however, depends to a large extent on the biologic requirements of the species to be protected and the interactions of the area with the surrounding region. Thus, in the rolling taiga country (swampy coniferous forests) of northern Europe, Asia, and North America, the decision to maintain a bog in a natural state may involve the protection of only a small area—the bog and those immediately adjacent slopes that drain into it. On the other hand, the decision to protect the natural habitat and population of migratory caribou in the same region could involve an area several hundred miles in length and more than 100 miles in width—a sizeable area that might include taiga and tundra. The decision to protect the natural habitat of a species of migratory waterfowl would be even more far-reaching. In addition to protecting the tundra breeding ground of such birds, protection of their resting and wintering grounds, which could lie in a different hemisphere, might also be involved. It is always easier to protect rooted plants than mobile animals. And the more mobile the species, the more difficult the task; for many migratory species, international action is required.

Difficulty in protecting migratory species

Any decision to protect an area in an undisturbed condition, however, must take ecological reality into account. The experience of Everglades National Park in the United States, which protects only the lower end of an extensive watershed, is illustrative. The stormy history of this park has involved efforts to control water and land use in areas outside the park because its future existence depends upon the flow and quality of water from those areas. By contrast, high mountain reserves, such as those surrounding Mount Kinabalu, on the island of Borneo; Mount Kenya, in East Africa; or Glacier National Park, in Montana, offer no such difficulties.

The designation of an area as a strict nature reserve is often proclaimed by law and recognized only by the appropriate governmental authorities, with no cognizance of the action by others. Such areas must have boundaries that are clearly demarcated and identifiable to prevent accidental intrusion and modification.

Because a strict nature reserve is set aside for scientific purposes, its use for recreation or any other purpose may disturb its natural integrity. Even scientific use itself can be a disturbing factor; hence, the use of such reserves is regulated by scientists. Usually a scientific advisory committee must rule on the appropriateness of any proposed research to the long-term future of the area.

The decision to protect a natural area for scientific purposes is not a simple one. In the case of most of the less affluent countries, in which neither money nor technical expertise is available, it is virtually impossible to proclaim a strict reserve and have it maintained as such except in the remotest areas. Usually international assistance is required, in money or manpower, if such reserves are to be established.

*National parks.* The establishment of national parks in the United States represented one of the first national efforts to protect wild nature. Yet, in establishing Yellowstone National Park, Congress made clear that it was viewed as "a pleasuring ground" for people and not as an area intended only to safeguard communities of plants and animals. It was not until the formation of the U.S. National Park Service in 1916 that the concept of managing parks so as to maintain their natural qualities was accepted. Nevertheless, the practice of killing predatory animals as "undesirable" elements of wild nature continued in U.S. national parks into the 1930s and lasted in some African national parks as late as the 1960s.

Unlike a strict nature reserve, a national park may be made available for various purposes but usually only for those forms of recreational use that do not create great changes in or require significant modifications of the natural environment. National parks usually are selected on the basis of their unique qualities, outstanding natural beauty, unusual geologic formations, or remarkable array of wild animal or plant life. They may also be selected, however, to protect areas of anthropological, archaeological, or historical importance along with the natural or artificially modified landscapes that surround them. In the United States, national parks are dedicated solely to recreational activity. National parks in England may protect cultural as well as natural landscapes, in that some may be dedicated to the preservation of traditional forms of land use that are disappearing elsewhere. Some national parks, such as in Peru, protect ethnic groups along with their hunting and gathering grounds.

Thus, exactly what constitutes a national park varies according to the nations and people involved. The dedication of an area as a national park is everywhere a highly restrictive form of land use, in which all incompatible activities are prohibited. Hunting, logging, mining, commercial fishing, agriculture, and livestock grazing are excluded from most such parks, as are urban and industrial uses not directly related to recreation. There is much debate as to whether tourist facilities should be within or outside national parks; because of their disruptive effects, the trend is to locate such facilities outside.

Facilities needed in national parks

National parks, at a minimum, require equally extensive boundary demarcation and perhaps policing and pa-

trolling as are necessary for strict nature reserves. They also require the careful planning of trails, roads, and other means of human access in order to channel the activities of visitors in ways that will not disrupt the resources or landscapes. Not only must certain fragile areas be set aside and protected from visitors, but visitor use must be concentrated in those places in which human activities will do a minimum of harm. The trend has been to divide national parks into zones that range from areas of intensive public use at one extreme to the remotest wilderness or strict nature reserves at the other.

Usually a considerable amount of money and energy must be invested in the planning and management of a national park. This is often beyond the resources of the less affluent countries unless international assistance can be provided. Because of their attraction as sites for outdoor recreation and their appeal to tourists, however, national parks often more than pay for themselves in a short period of time. In East Africa, for example, national parks are a major source of foreign exchange of the countries in which they are located because of their unique wild animal life. As a result of the body of expertise that has developed in the planning and management of national parks and because of their growing economic importance, expert direction in the establishment and maintenance of national parks is now available throughout the world whenever it is requested. Within the United Nations such assistance is offered to developing countries by UNESCO, UNEP, and FAO; outside the United Nations, assistance is available from the International Union for Conservation of Nature and Natural Resources and from the World Wildlife Fund.

*Refuges and sanctuaries.* A less restrictive use of an area of wildland is that directed toward the protection of certain species or groups of wild animals or plants. This type of land use includes wildlife refuges and sanctuaries as well as various kinds of botanical reservations. In such areas, use of the land that could interfere with the well-being of the protected species is excluded. Other forms of land use may be encouraged, however, if they are not in conflict or if they actually assist with the creation of a suitable environment for the protected forms of life.

Refuges and sanctuaries often are established for the preservation of endangered species of wildlife or plants, particularly those whose numbers and distribution have been seriously curtailed; examples include the Umfolozi Game Reserve, in South Africa, in which the southern species of the white rhinoceros is protected, and the Mountain Zebra National Park, in the same country. The refuges for the California condor and the Torrey pine in California are of a similar nature. Refuges and sanctuaries may also be provided for more abundant species that require protection at certain periods of their life cycle or in certain areas where they gather in order to reproduce. Many wildlife refuges and sanctuaries in the United States and Europe are of this type; they provide protected sites for the resting, breeding, or wintering of wildlife species (particularly waterfowl) that otherwise are hunted outside of the refuge. In such places management measures are necessary to enhance the habitat for the protected species and to remove any competitors or predators that might interfere with the breeding and survival of the young. Such measures would be inappropriate to parks or strict nature reserves, because their purpose is to protect a total biologic community without favouring a particular species.

efuges as rotected ites for bundant ildlife pecies

The various measures employed for the protection and management of living resources on land also apply to water areas. Marine and other aquatic parks and reserves have been established in many parts of the world to protect various forms of saltwater and freshwater plant and animal life. Australia, for example, has reserves that protect important areas of the Great Barrier Reef; Kenya and Tanzania have marine parks and reserves on their coasts. Lake Baikal, in the Soviet Union, is now included in a major national park designed to protect not only its unique freshwater life but also its watershed areas.

**Wildlife and fisheries management.** The protection of most wild animal life cannot be accomplished solely through parks and reserves. Consequently, wildlife conservation and management represent an activity that extends into other areas. As has been mentioned previously, one form of wildlife conservation with a long history of laws, regulations, and proclamations dating back to ancient centres of civilization was the protection of royal game. Wildlife conservation is also closely related to the management of fisheries, because both are directed toward the preservation of wild species and their habitats and toward increasing the productivity and yield of these species. Increased productivity in wildlife or fish may contribute to a commercial harvest, or it may make contributions to human well-being through sport hunting or fishing, wildlife viewing, or some other form of recreation.

The basis for wildlife and fishery management includes research into the ecological requirements and breeding potentials of the species involved. Protective laws and regulations are necessary to control the allowable commercial or sport take in order to guarantee that it remain within the sustained yield. Means for enforcing such laws should include the employment and deployment of adequate protective forces, preferably specially trained wildlife and fisheries wardens who are capable of identifying species likely to be killed or captured and who are familiar with the ways of hunters and fishermen. Management activities intended to maintain or improve the habitat for the wild animals concerned are also important. Special refuges of the kind already described may be essential both to control hunting and fishing pressure and to allow animals to reproduce, rest, or winter without being disturbed. In addition to these activities, wildlife conservation also involves the management of the requirements of wildlife in the areas in which the land is used for other purposes.

Commercial value of fish and wildlife

Commercial fishing, which is of great economic importance in countries with extensive inland waters or with access to the seas and oceans, contributes a significant share of world food, particularly protein. The problems of fisheries conservation are many, however. It is essential first to locate the fish and derive some estimate of their abundance. Much of the recent increase in yield by world fisheries has come from the discovery of previously little known and unexploited sources of fish. It is also necessary to determine the maximum sustained yield of the fish population. Finally, the most difficult task is the supervision and control of the fishery. Even in national waters, fisheries have been overexploited and depleted to a level such that the species concerned no longer has any commercial value—the sardine fishery of the California coast is an example. In international water it is much more difficult to regulate and control the activities of fishermen.

The commercial use of terrestrial wildlife, though not in the same economic category as fisheries, is nevertheless significant to some countries. Wild animals are killed for their meat, hides, furs, fats, oils, bones, ivory, antlers, and other by-products. Live animals are captured for zoos, for medical research, and for the pet trade. The exceedingly high prices paid for certain species, such as various primates used for medical research or those that are valued for their furs, has been of great importance in causing the reduction of some species to a dangerously low level. Despite the high commercial value of some wildlife, however, in many parts of the world their recreational or sporting value has an even greater economic importance. In the United States, for example, the commercial hunting and sale of game animals is generally prohibited, as is the commercial harvesting of certain game fish, because such wildlife is reserved for recreational use.

In technologically advanced countries it has proved practical and economically desirable to manage wildlife and fisheries so that no species becomes severely depleted. But, in those countries in which a high percentage of the people are illiterate or live near the poverty level, the protection and management of wildlife and fisheries is much more difficult to achieve. In such nations protective laws are not easy to enforce because trained wardens are not widely available and because the laws are neither understood nor accepted by the general population. Thus, the depletion of wildlife and fisheries through hunting and fishing continues in these countries, and growing numbers of species are now endangered.

**Multiple-use management.** A wildlife conservation area may be used for additional purposes. Throughout the world many areas of public wildland are managed on a policy of multiple use; that is, they can provide either at the same time or sequentially a variety of wildland products or services. National forests, for example, are used for recreation and to produce timber and range forage for domestic livestock, wildlife, and fish; and at the same time they stabilize watersheds and yield sustaining water supplies to natural bodies of water or to man-made reservoirs.

Conflicts in multiple use

The decision to manage an area for multiple use, however, necessitates minimization of conflicts among the various forms of use. Thus, removing the timber from an area involves the construction of logging roads, landings, and other facilities; and, while the area is being logged, most other uses are restricted. Furthermore, during and following logging, care is taken not only to provide a site and seed trees suitable for growing the next crop of trees but also to protect the natural reproduction of trees, young seedlings, and saplings. In cases involving more intensive management, seeds or seedlings are planted and given protection during the period when they are becoming established. Moreover, the value of the site for recreation or for other purposes may be greatly reduced during this process, particularly when certain systems are used for cutting timber; thus, clear-cutting, which involves the total removal of all large trees from an area, creates an appearance of greater devastation than is seen in selective cutting, in which only a few trees of a particular species and age class are removed. Obviously, clear-cutting then favours wildlife species that prefer open areas at the expense of those that require dense forest cover. Generally speaking, however, logging of timber involves the exclusive use of an area for this purpose over a period of time; other uses may not be fully restored until the vegetation or animal life has recovered from the effects of timber removal.

Development of an area for grazing by domestic livestock often includes fencing; the demarcation of livestock trails for the transfer of animals from one area to another; the provision of watering points, salt grounds, and other items essential for the welfare of the animals; and sometimes the construction of special corrals or handling chutes. When livestock are in an area, they cannot be disturbed too frequently. Livestock owners also demand the removal of any predatory animals that are likely to attack their stock. Such needs obviously restrict the development of the same site for purposes that could conflict with its use as a livestock range.

Factors involved in development for water power

Intensive development of an area for water power (*e.g.*, the construction of dams, power stations, pipelines, canals, and power lines) entails complete change of the area affected and removes much of its wild or natural quality. Although such a location is obviously less suited for wilderness or wild-country recreation, it often is enhanced for certain forms of mass recreation, such as fishing, boating, and water sports in reservoirs. Recreational development on any intensive scale also involves considerable modification of an area. The construction of tourist and visitor facilities of various kinds—trails, ski runs, campsites, roads and parking areas, lodges, and picnic grounds—necessarily restricts the use of such an area for other purposes.

The concept of multiple use thus applies mostly to the management of large areas of wildland, although within such areas various special uses may from place to place and from time to time be emphasized. Nevertheless, the governing policy of multiple use is valuable because it prohibits the exclusive use of a particular area for the benefit of any one segment of society and, instead, forces accommodation to the greatest variety of uses that best suit a particular site.

**Intensive wildland uses.** *Timber production.* Loss of the natural character of an area usually accompanies the decision to manage it for maximum production of such crops as timber or range forage. Thus, if a forested area is to be managed for maximum timber production, establishment of fast-growing species of trees that will make the most efficient use of the soils, water, and climate is necessary. Crops of timber or other wood products are cut and removed regularly and other trees planted in their place, usually from nursery stock. Roads and other facilities are constructed to facilitate the removal or protection of the forest products. The ultimate objective is usually a high-yielding plantation or tree farm, often using exotic species of trees or specially developed varieties that combine high-yielding capabilities with resistance to insect pests or diseases.

Seldom do such forests have any value for other purposes, except possibly for soil protection and water yield. Aesthetically, such a forest may even lose its wild aspect and have all the regularity and artificiality of an orchard or cornfield. Recreational uses often are discouraged; even

(Left) Ray Atkeson, (right) Josef Muench

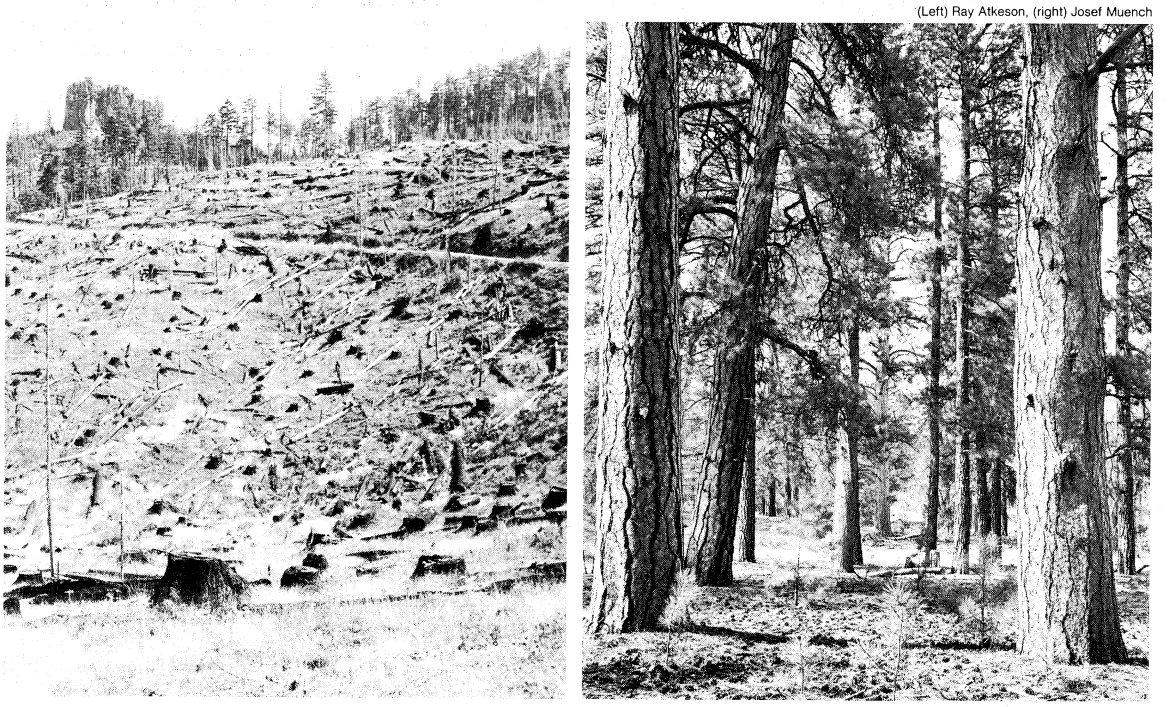


Figure 4: *The cutting of timber.*
(Left) Clear-cut timber in Oregon. (Right) Selective cutting of timber in the Chuska Mountains, New Mexico.

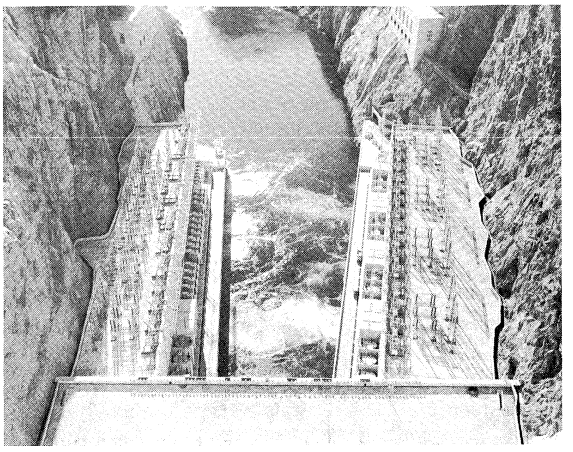


Figure 5: Hoover Dam, Colorado River, on the border of Arizona and Nevada, used for irrigation and hydroelectric power.
Harvey Caplin

if they are not, the recreational appeal of such areas usually is slight. Few if any of the larger forms of wild animal life can find a suitable habitat in a forest plantation, and those that do become established there often are regarded as pests if they interfere in any way with tree production. Such forests, however, produce far more timber, pulp, and other kinds of wood products than can be obtained from an equivalent area of wild forest. Moreover, intensive development of production forests in suitable locations can spare areas of natural forest for other purposes. Privately owned forests in North America are increasingly managed in this way, a practice that increases with the demand for forest products; government-owned production forests and forest plantations in other countries receive similar treatment. Such intensive development, however, is economically feasible only at sites with soils and other habitat conditions favourable to high yields.

*Livestock production.* Management of range and pasture lands for livestock production can become almost a single-purpose use of the land. A range, any area of wildland that is covered with natural vegetation, is usually located in the more arid mountainous or colder regions of a country; such land is suitable for the production of the larger grazing or browsing animals. Pastures, on the other hand, are grazing areas with better growing conditions than ranges because they are located either in regions of higher rainfall or in areas suitable for irrigation. Because of the nature of pasture, more intensive management is economically practical, including regular cultivation and planting to replace the natural vegetation with higher yielding domestic, exotic, or artificially bred varieties of forage plants. Livestock production on pastures is a single-purpose use of the land, although pastures are often included as part of a rotation system in the raising of farm crops. The intensive use and management of pastures effectively removes such areas from the category of wildland.

Range and pasture as single-purpose use of land

Management of the forage on rangelands requires skillful manipulation of natural processes of plant succession and the balancing of livestock numbers to the carrying capacity of the vegetation; that is, the number of animals the vegetation can support without deteriorating and endangering its future productivity. Efforts are made to avoid overgrazing, which reduces the vegetation to a less productive state, and excessive trampling by animals, which can compact the soil and interfere with its water relationships and productivity. A flexible method of livestock handling is essential to success on ranges: animals are not allowed to concentrate excessively in any one area, and their use of vegetation must be consistent with continued production of the more nutritious and palatable forage plants. Moreover, during dry periods, when there is low forage production, the number of livestock using a range must be reduced in order to avoid damage to the range; otherwise, feed lots are provided or the animals are moved to irrigated pastures.

As a primary use of rangeland, livestock production requires the construction of various facilities—fences, water holes, corrals, roads, and troughs for salt or other forage supplements. On productive rangelands it may be economically profitable to provide fertilizers, usually by spraying or dusting from aircraft, to encourage higher forage yields. Undesirable vegetation is sometimes burned, crushed, or sprayed with herbicide to provide more space for the better forage plants; ranges also may be seeded with higher yielding or more disease-resistant forage plants. Such activities necessarily restrict use of the land for other purposes. Nevertheless, except in cases of intensive management, rangelands can provide recreation and often are highly suited to various forms of wildlife management. There is considerable evidence from many parts of the world that some combination of domestic animals and wildlife can provide greater yields and profits from drier, colder, or steeper rangelands than can be obtained from such lands when used exclusively for domestic-livestock production.

**Agricultural management.** The most intensive forms of rural land use for agricultural purposes are those concerned with the raising of harvestable crops or with the production of animal products. Unlike primitive agriculture, which involved only the temporary removal of natural vegetation and depended for a short period of time on natural soil fertility, conventional agriculture today uses large inputs of chemicals, energy, and technical skills to produce increased yields of crops or animals. In the technologically advanced countries food production is often greater than population growth, and it is possible to retire former farmlands from use and to produce crops according to demand without approaching the maximum yields obtainable. The so-called Green Revolution has been based on the spread of such farming methods to less developed nations of the world. It has been made possible by the breeding of high-yielding forms of grain specifically adapted to the ecological conditions of the countries involved.

Green Revolution

(Left) Ray Atkeson, (right) J.C. Allen and Son

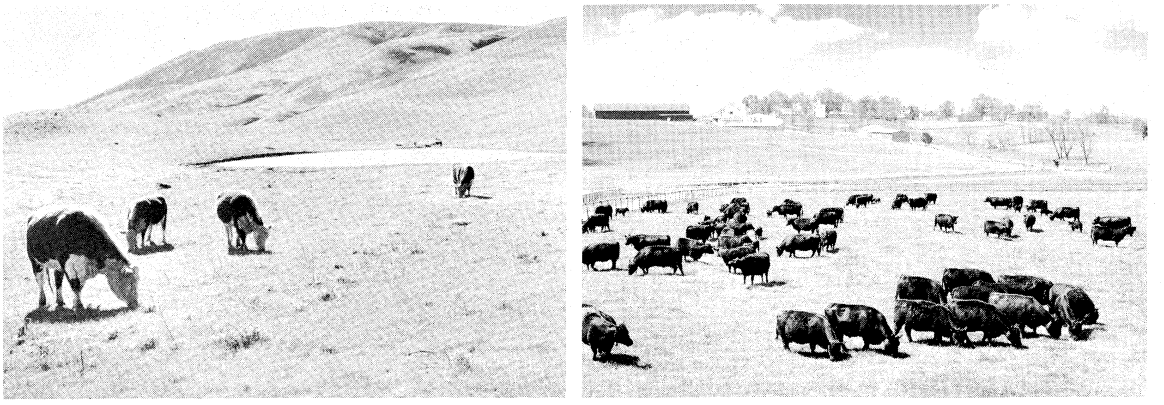


Figure 6: *Livestock production on range and pastureland.*
(Left) Rangeland in southwestern Washington. (Right) Angus cattle on bluegrass pastureland in Kentucky.

The decision to use an area of land for high-yield agriculture essentially rules out its use for other purposes. The intensive production of farm crops in an agricultural region may also have undesirable side effects; as has been previously noted, these may include the pollution of other areas when the pesticides, herbicides, or other agricultural chemicals blown or washed from farmlands affect vegetation and animal life elsewhere. Nevertheless, committing an area to intensive agricultural production does not rule out its future restoration for other uses. As long as the soils are well cared for, such areas can be converted quickly to other purposes if it is not necessary to keep them in farm production. Abandoned farmlands in the southern United States, for example, are now highly productive forest areas, and former farming lands elsewhere are being used to support wildlife and outdoor recreation. In general planning for conservation of natural resources, intensive use and high production in those areas best suited for farming must be encouraged—provided, of course, that the polluting effects of these activities on the general environment are avoided. Such concentration can spare the destruction of other resources through attempts to use inadequate lands for marginal farming activities.

## INTERNATIONAL PROBLEMS OF RESOURCE MANAGEMENT

The management of living resources requires a high degree of international cooperation and a willingness on the part of nations to agree to some forms of international control. This is particularly true for the management of those aquatic animals that occupy international waters; it is equally true for migratory animals that move from one country to another. Furthermore, the animals and plants that are commodities in international trade must also be protected through international agreements. Finally, when an international agency or an agency from a particular country works with the resources of a less developed nation in an effort to help that nation improve its economy, care is taken to avoid any adverse effects on the conservation of that nation's environment.

Despite general recognition of these problems, few nations have shown a willingness to forgo any of their sovereign rights or to cede authority over their affairs to international bodies. International laws and regulations remain, for the most part, gentlemen's agreements among nations. A country can and often does ignore the rules when it is economically advantageous for it to do so. Only international public opinion or, more rarely, the threat of force by one nation or by a group of nations can serve as an effective deterrent in compelling a country to stop certain activities that are endangering a resource. Yet, despite the generally chaotic state of international rules governing the management of environmental resources, there are some excellent international agreements concerning this matter. Such agreements provide hope that the general field of conservation may serve as an area in which nations can learn to work together more effectively for their mutual benefit.

Problems in enforcing international conservation regulations

**Effective international agreements.** *The Antarctic Treaty.* The Antarctic continent remained unknown and unexplored during the period when most other parts of the world were being claimed and colonized by European powers. Uninhabited and, until recently, of little or no value to any country, Antarctica never became an area for international dispute. Although segments of the continent and sub-Antarctic islands have been claimed by various nations, including the United States, the Soviet Union, Great Britain, Australia, New Zealand, Chile, and Argentina, the validity of these claims has never been tested. It was not until the latter half of the 20th century that the potential future importance of Antarctica and the knowledge to be gained from scientific research in and exploration of the continent were recognized. As a consequence, nations with an interest in Antarctica signed the Antarctic Treaty in 1961. This could be considered de facto establishment of the Antarctic continent as the world's largest strict nature reserve. The treaty provided for joint scientific research to be conducted by the signatory nations, with the results to be shared by all. It also prohibited the exploitation by any one nation of living Antarctic resources as well as any activities that would cause deterioration of those resources. The treaty has been respected, Antarctic research has gone forward, and the wild animals and sparse plant life of the continent have been protected.

*Whaling agreement.* Unfortunately, most agreements covering the Antarctic have not included the oceans surrounding the continent. The value of the animal resources in these waters has been known since the 18th century, and exploitation of them was begun during the 19th century. Of particular importance were the huge herds of whales that congregated in the Antarctic seas to feed upon the abundant plankton in these nutrient-rich waters. One component of this plankton is the krill, a shrimplike crustacean; the Antarctic Treaty nations established a protective zone around the ocean waters of the continent in 1982 to preserve the krill and other marine organisms.

The early whaling ships concentrated their efforts on the larger whales, which could yield the greatest quantities of whale oil, whalebone, and other useful products. The most sought after of these creatures, the blue whale, is the world's largest mammal. It is estimated that when whaling began there were 200,000 blue whales in the Antarctic; by 1965, however, the population of blues had been reduced to about 2,000. Preceding the period of concentration on blue whales, whalers had pursued the various species of right whales in the world's oceans and had reduced their numbers to a similar low level. The gray whale of the Pacific was also brought to a point of near-extinction. The efforts of Antarctic whalers then shifted to smaller species; the humpback, once abundant, was reduced to an estimated 1,000 in 1962. Fin whales and sperm whales also bore the brunt of whaling pressure for a time; when they became scarce, the still smaller sei and Minke whales were killed.

Slaughter of blues and other whales

Despite the efforts of conservationists, attempts to restrict the activities of whalers met with little initial success. As whale populations dwindled and the danger of extinction became evident, however, the principal whaling nations agreed to sign the International Whaling Convention in 1946. This led to the establishment of the International Whaling Commission, which was authorized to sponsor scientific studies of whale populations and to recommend to the whaling nations limitations on harvest that were necessary to perpetuate whale populations and the whaling industry. Because the nations involved did not give the commission any firm power or enforcement authority, any nation could dispute the recommendations of the commission's scientific advisers and insist upon higher quotas than had been recommended. Everything depended on the willingness of the nations to obey the rules and to report honestly the number of each species of whales taken during the whaling season. Although there is no evidence that the nations involved deliberately disobeyed the recommendations finally agreed upon by the commission or that they failed to provide the commission with anything but accurate figures, the number of whales continued to decline. Sustained pressure from conservation interests, however, finally accomplished results. After 1967 the endangered whales—the blue, right, gray, and humpback—were given complete protection. Starting in the 1970s, quotas for other species were reduced to limits that enabled the whales to maintain their populations at certain levels. At its 1983 meeting the commission, over the objections of Japan, Norway, and the Soviet Union, agreed to end commercial whaling entirely by 1986; nevertheless, whaling continued after the deadline—though at a reduced level. Thus, the record of international cooperation provided by the International Whaling Convention is a mixed one. Although it did not accomplish all of the desired results, neither did it fail entirely. Indeed, the record of whale conservation would be much worse had it not been for this international agreement.

*Protection of fur seals and migratory birds.* Other activities intended to control the international exploitation of environmental resources have had some successes and many failures. Among the most successful treaties has been one to protect the northern fur seal, a species that breeds in the Pribilof Islands, in the Bering Sea. During

The fur seal treaties

the 19th century fur seals were reduced to a dangerously low level as a result of the heavy slaughter at their breeding grounds to obtain skins for the manufacture of fur coats and various other sealskin products. In 1911 Canada, Japan, Russia, and the United States signed a treaty to limit the annual harvest of seals to a quantity that would not only sustain the population but also increase it annually. In addition, the profits from sealing were to be divided proportionately among the signatory nations. The treaty was adhered to; the seals have increased and now stock the available breeding grounds.

A good example of an international agreement governing the conservation of species that migrate between two or more nations is the International Migratory Bird Treaty. Established in 1918 between the United States and Canada, this treaty was subsequently extended to include Mexico. The treaty, which has been adhered to by the nations concerned, limits the kill of migratory waterfowl. It also provides for the protection of migratory species in their breeding grounds, along their migration routes, and at their wintering areas.

**Territorial limits and marine resources.** For the most part, international control of species occupying habitats in international and national waters remained a serious problem in the 20th century. The traditional three-mile agreement was loosely agreed upon by most coastal countries for military reasons: it evolved in the days when that distance was beyond the range of shells fired from guns aboard vessels offshore. The advent of large naval guns capable of firing shells three miles or more and the dominant role of naval air power during World War II made the three-mile limit useless for defense. Disputes over fishing rights led to a more general acceptance of a 12-mile (22-kilometre) limit to the seaward boundaries of national territories, but even this was not acceptable to many countries with limited coastlines or narrow continental shelves. Many Latin-American countries insisted on jurisdiction over waters to a distance of 200 nautical miles (370 kilometres) from their shores.

These disputes over national jurisdictions, combined with the desire by many countries to further exploit the marine and mineral resources of the continental shelves and deep-ocean floors, led to the first United Nations Law of the Sea Conference in 1958 to negotiate the partitioning of the oceans and their resources. A second conference was convened in 1960, but neither meeting succeeded in reaching an agreement. The Third Law of the Sea Conference was convened in 1973, and it met in numerous sessions during the 1970s and early 1980s. Finally in 1982 the Convention of the Law of the Sea was promulgated by a UN conference. The convention was quickly signed by 117 nations, but the United States and some 20 other, chiefly industrial, countries did not sign the agreement. The nonadopting nations, however, do agree that each coastal state has sovereign rights over a 12-mile territorial sea and sovereign rights over natural resources and certain economic activities within a 200-mile exclusive economic zone (EEZ). Ships are allowed innocent passage through the territorial sea, and ships and aircraft are allowed transit passage through straits used for international navigation. Coastal states also have sovereign rights regarding exploration and exploitation of the continental shelf (200 miles from shore and in some cases farther). Provision is made for delimitation of the territorial sea, EEZ, and continental shelf of archipelagic states and for access to the sea for landlocked nations. A large number of countries, including the United States, have proclaimed their sovereignty over the EEZ and the marine and mineral resources within it.

**International trade in animals and plants.** Special problems are involved in the regulation of trade and commerce in living animals and plants as well as in the products derived from them. Demands by wealthy nations for certain animal and plant products create particularly severe problems in less affluent countries. As mentioned previously, the trade in endangered species of wildlife is illustrative. The demand for furs and skins of rare animal species is artificially created in the fashion centres of the world. Prices paid by wealthy people for these items in affluent countries exceed the lifetime income of most people in the countries from which the leopards, crocodiles, tigers, and other wild species come. Poachers go to great lengths to obtain these animals wherever they can be found, including inside national parks and reserves. Because effective policing is virtually impossible, legal and illegal trade in wildlife begin to overlap, and both become firmly established. Exporters of wild animals and their products are the end links of profitable business chains that include far greater numbers of hunters and trappers in remote areas. Furthermore, for each animal or skin that reaches a foreign market, many more are destroyed in hunting, trapping, and transporting.

urs and
ins for
ie fashion
iarket

The purchasing countries can most effectively control the illegal trade in wildlife. This can be done by directing fashion away from the use of wild furs and by restricting the purchase of live animals and controlling their clearance through customs offices of international airports and seaports. In 1973 representatives of 80 countries signed the Convention on International Trade in Endangered Species of Wild Fauna and Flora, which prohibited commercial trade in 375 endangered species of wild animals. The treaty, which has been ratified by nearly all of the 80 governments, forbids trade in products derived from the animals (*e.g.,* hides) as well as in living animals. In addition, trade in 239 species was allowable only on the granting of permits by both the importing and the exporting countries. The provisions also included endangered plants (such as rare orchids that are now being removed from even the remotest tropical forests).

Perhaps even more serious than the trade in wild animal species is the international exploitation of other living resources, particularly the tropical forests of the world. These forests, which contain many hundreds of species of trees growing in diverse mixtures, were spared from exploitation in earlier decades because of their inaccessibility, the relatively low value of most of the trees for timber purposes, and the limited world demand. Heavily exploited for special uses were a few species of high value, such as teak, ebony, sandalwood, mahogany, and other furniture woods. Most tropical forests were not greatly disturbed, however. This situation has changed, and a wide variety of woods previously considered worthless are used for pulp, chipboard, and fibreboard or as cellulose for plastics production. With new machines and better transportation, it has become profitable to remove trees from previously remote areas and to ship logs, bolts, wood chips, or other partially processed materials to foreign markets. Faced with a high demand for their forest products, most developing nations have been willing to sign over timber rights to foreign companies, hoping thereby to increase their national incomes and to advance the general material welfare of their people. Unfortunately, most of these timber contracts contain few or no provisions for conservation. Forest industries that have excellent management and conservation records in their home countries behave differently in other lands. Great areas of tropical forest have been laid waste, soils bared to erosion, and the wildlife within them destroyed. Because no laws are violated in either the exploited or the home country, there is no effective redress. General international agreements governing the conservation of such living resources would provide an answer to this problem, but they are unlikely to be implemented in time to prevent the devastation of large areas of the tropical world.

Unintended side effects of international assistance

Equally if not more unfortunate have been the side effects of some well-intentioned international development projects. These are sometimes sponsored by international agencies concerned with such affairs and sometimes by the foreign-assistance departments of individual donor nations. Usually the projects are intended to benefit one segment of the economy of the recipient nation; but, because ecological advice generally is not sought and because of the broad effect of the proposed development on other resources or on the total environment, the side effects of some of these activities often far outweigh any benefits that are derived. An example is the Aswān High Dam of Egypt, where the need to increase the supply of water for irrigation and power was considered paramount. The environmental side effects, however, have been enormous

and include the spread of the disease schistosomiasis by snails that live in the irrigation channels, loss of land in the delta of the Nile River from erosion once the former sediment load of the river was no longer available for land building, and a variety of other consequences. Furthermore, water-development projects in other semiarid lands of Africa, sponsored by international agencies and funded by various donor organizations, have quite frequently been far more destructive than constructive. Failure to take ecological and social factors into consideration has resulted repeatedly in overgrazing and creation of deserts in an area intended to be improved by the expenditure of money. The responsibility of agencies concerned with international development to seek the best environmental advice is now generally accepted, but implementation of this responsibility has been slow.

**Extraterrestrial conservation.** Space exploration has begun at a time when the willingness to manage the Earth's resources is in doubt. Experiences gained from early space efforts, however, have helped promote an awareness of planetary realities among a growing number of people. The need to provide the requirements of life to people confined for long periods of time in spaceships and the problems of waste disposal in such vehicles have cast new light on similar problems within communities on Earth. The view of Earth from outer space has engendered the concept that the planet is, in effect, a large spaceship with a limited capacity to support life and that it is highly vulnerable to damage from poorly planned human activities.

Spaceship Earth

Exploration of other planets has been undertaken in full realization of the dangers involved, both in the pollution of a satellite or planet being visited by earthlings and in the potential danger to life on Earth by materials brought back from other celestial bodies. Although there is no reason to expect the existence of life on those planets nearest the Earth and every reason to doubt its existence on planets farther away in the solar system, the most rigid precautions are still necessary. Experience with the rapid spread of diseases in new locations has demonstrated the dangers to other forms of life if other planets were to be contaminated with organisms brought from Earth or if Earth were to be exposed to agents from other planets capable of contaminating life on this one.

If planets capable of supporting life are reached someday, they will offer an opportunity to science never before available. The study of life that has evolved under different conditions in a different world could add greatly to an understanding of the Earth's biologic systems as well as provide a great intellectual challenge to mankind.

## THE POLLUTION OF NATURAL RESOURCES

Although various problems related to pollution of the environment have already been mentioned, the following is a more general discussion of pollution as a phenomenon.

Definition of pollution

Pollution may be defined as the addition of any substance or form of energy (*e.g.*, heat, sound, radioactivity) to the environment at a rate faster than the environment can accommodate it by dispersion, breakdown, recycling, or storage in some harmless form. A pollutant need not be harmful in itself. Carbon dioxide, for example, is a normal component of the atmosphere and a by-product of respiration that is found in all animal tissues; yet in a concentrated form it can kill animals. Human sewage can be a useful fertilizer, but when concentrated too highly it becomes a serious pollutant, menacing health and causing the depletion of oxygen in bodies of water. By contrast, radioactivity in any quantity is harmful to life, despite the fact that it occurs normally in the environment as so-called background radiation.

Pollution has accompanied mankind ever since groups of people first congregated and remained for a long time in any one place. Primitive human settlements can be recognized by their pollutants—shell mounds and rubble heaps. But pollution was not a serious problem as long as there was enough space available for each individual or group. With the establishment of permanent human settlements by great numbers of people, however, pollution became a problem and has remained one ever since. Cities of ancient times were often noxious places, fouled by human wastes and debris. In the Middle Ages, unsanitary urban conditions favoured the outbreak of population-decimating epidemics. During the 19th century, water and air pollution and the accumulation of solid wastes were largely the problems of only a few large cities. But, with the rise of advanced technology and with the rapid spread of industrialization and the concomitant increase in human populations to unprecedented levels, pollution has become a universal problem.

The various kinds of pollution are most conveniently considered under three headings: air, water, and land.

**Air pollution.** Air pollution involves the release into the atmosphere of gases, finely divided solids, or finely dispersed liquid aerosols at rates that exceed the capacity of the atmosphere to dissipate them or to dispose of them through incorporation into solid or liquid layers of the biosphere. Air pollution results from a variety of causes, not all of which are within human control. Dust storms in desert areas and smoke from forest and grass fires contribute to chemical and particulate pollution of the air. Forest fires that swept the state of Victoria, in Australia, in 1939 caused observable air pollution in Queensland, more than 2,000 miles (3,000 kilometres) away. Dust blown from the Sahara has been detected in West Indian islands. The discovery of pesticides in Antarctica, where they have never been used, suggests the extent to which aerial transport can carry pollutants from one place to another. Probably the most important natural source of air pollution is volcanic activity, which at times pours great amounts of ash and toxic fumes into the atmosphere. The eruptions of such volcanoes as Krakatoa, in the East Indies, Mt. St. Helens, in Washington, and Katmai, in Alaska, have been related to measurable climatic changes.

Air pollution may affect humans directly, causing a smarting of the eyes or coughing. More indirectly, the effects of air pollution are experienced at considerable distances from the source, as, for example, the fallout of tetraethyl lead from urban automobile exhausts, which has been observed in the oceans and on the Greenland ice sheet. Still less directly experienced are the possible effects of air pollution on global climates.

*Urban air pollution.* It is the immediate effect of air pollution on urban atmospheres that is most noticeable and causes the strongest public reaction. The city of Los Angeles has been noted for both the extent of its air pollution and the actions undertaken for control. Los Angeles lies in a coastal plain, surrounded by mountains that restrict the inward sweep of air and that separate a desert from the coastal climate. Fog moving in from the ocean is normal to the city. Temperature inversions characterized by the establishment of a layer of warm air on top of a layer of cooler air prevent the air near the ground from rising and thus effectively trap pollutants that have accumulated in the lower layer of air. In the 1940s, the air in Los Angeles became noticeably polluted, interfering with visibility and causing human discomfort. Attempts to control pollution, initiated during the 1950s, resulted in the successful elimination of such sources of pollution as industrial effluents and the outdoor burning of trash and debris. Nevertheless, pollution continued to increase as a result of the increased number of motor vehicles. Exhaust fumes from the engines of automobiles contain a number of polluting substances, including carbon monoxide and a variety of complex hydrocarbons, nitrogen oxides, and other compounds. When acted upon by sunlight, these substances undergo a change in composition producing the brown, photochemical smog for which Los Angeles is well known. Efforts to reduce pollution from automobile engines and to develop pollution-free engines may eventually eliminate the more serious air pollution problems. In the meantime, however, air pollution has driven many forms of agriculture from the Los Angeles basin, has had a serious effect upon the pine forests in nearby mountains, and has caused respiratory distress, particularly in children, elderly people, and those suffering from respiratory diseases.

The air pollution problem of Los Angeles

Los Angeles is neither a unique nor the worst example of polluted air. Tokyo has such a serious air-pollution problem that oxygen is supplied to policemen who direct

traffic at busy intersections. Milan, Ankara, Mexico City, and Buenos Aires face similar problems. Although New York City produces greater quantities of pollutants than Los Angeles, it has been spared from an air-pollution disaster only because of favourable climatic circumstances.

The task of cleaning up air pollution, though difficult, is not believed to be insurmountable. Use of fuels that are low in pollutants, such as low-sulfur forms of petroleum; more complete burning of fossil fuels, at best to carbon dioxide and water; the scrubbing of industrial smokestacks or precipitation of pollutants from them, often in combination with a recycling of the pollutants; and the shift to less polluting forms of power generation, such as solar energy in place of fossil fuels—all are methods that can be used for controlling pollution. The example of London, as well as of other cities, has shown that major improvements in air quality can be achieved in 10 years or less.

*Climatic effects of polluted air.* Less obvious than local concentrations of pollution but potentially more important are the climatic effects of air pollutants. Thus, as a result of the growing worldwide consumption of fossil fuels, atmospheric carbon dioxide levels have increased steadily since 1900, and the rate of increase is accelerating. The output of carbon dioxide is believed by some to have reached a point such that it may exceed both the capacity of plant life to remove it from the atmosphere and the rate at which it goes into solution in the oceans. In the atmosphere carbon dioxide creates a "greenhouse effect." Like glass in a greenhouse, it allows light rays from the Sun to pass through, but it does not allow the escape of the heat rays generated when sunlight is absorbed by the surface of the ground. An increase in carbon dioxide, therefore, can cause an increase in the temperature of the lower atmosphere. If allowed to continue, this could cause melting of the polar ice caps, raising of the sea level, and flooding of the coastal areas of the world. There is every reason to fear that such a climatic change may take place.

The 'greenhouse effect"

Counterbalancing the effect of carbon dioxide is the increase of particulate matter in the air, a result of the output of smoke, dust, and other solids associated with human activity. Such an increase might, in turn, increase the reflectance, or albedo, of the atmosphere, causing a higher percentage of solar radiation to be reflected back into space. This, in time, could cause a lowering of the Earth's surface temperature and, potentially, a new ice age. At present, however, the greater danger appears to lie in the steady increase in carbon dioxide, with its associated atmospheric warming.

Scientists also fear that the ozonosphere (or ozone layer of the atmosphere) is being depleted by the chemical action of chlorofluorocarbons emitted from aerosol cans and refrigerators and by pollutants from rockets and supersonic aircraft. Depletion of the ozone layer, which absorbs ultraviolet radiation from the Sun, would have serious effects on living organisms on the Earth's surface, including increasing frequency of skin cancer among humans.

Acid rain

Another climatic effect of pollution is acid rain. The phenomenon occurs when sulfur dioxide and nitrogen oxides from the burning of fossil fuels combine with water vapour in the atmosphere. The resulting precipitation is damaging to water, forest, and soil resources. It is blamed for the disappearance of fish from many lakes in the Adirondacks, for the widespread death of forests in European mountains, and for damaging tree growth in the United States and Canada. Reports also indicate that it can corrode buildings and be hazardous to human health. Because the contaminants are carried long distances, the sources of acid rain are difficult to pinpoint and hence difficult to control. Acid rain has been reported in areas as far apart as Sweden and Canada, and in parts of the United States from New England to Texas. The drifting of pollutants causing acid rain across international boundaries has created disagreements between Canada and the United States and among European countries over the causes and solutions of the precipitation. The international scope of the problem has led to the signing of international agreements on the limitation of sulfur and nitrogen oxide emissions.

*Radioactive contamination of the atmosphere.* During the 1950s the effects of atmospheric testing of atomic and hydrogen bombs became a source of major concern. The danger of radioactive pollution of the air and the fallout of radioactive particles to the surface of the Earth stimulated serious investigation, resulting in the discovery of potentially dangerous conditions. It was observed, for example, that radioactive materials of many kinds, such as radioactive iodine and strontium, are concentrated in living tissue and can cause damage even when the general level of environmental contamination is low. Atmospheric testing of nuclear bombs was stopped in the United States and the Soviet Union, and radioactive fallout from this source has declined. Concern continues, however, over the dangers resulting from massive releases of radioactive materials from nuclear weapons, which, if used on a major scale, could seriously endanger all of humanity.

Another concern is accidents at nuclear power plants. In 1978 the Three Mile Island nuclear power plant in Pennsylvania suffered a severe accident leading to partial meltdown of its radioactive core. Although most of the radiation was contained within the plant structure, the prospects of massive contamination of nearby cities and towns resulted in plans for the evacuation of hundreds of thousands of people. In 1986 the Chernobyl nuclear power plant near Kiev, in the Ukrainian S.S.R., suffered a fire and partial meltdown, resulting in a major release of radioactive particles. Much of northern and eastern Europe experienced heavy nuclear fallout, and the towns and farmlands around the power plant were no longer safe for human occupancy.

**Water pollution.** Water pollution involves the release into lakes, streams, rivers, and oceans of substances that become dissolved or suspended in the water or deposited upon the bottom and accumulate to the extent that they interfere with the functioning of aquatic ecosystems. It may also include the release of energy in the form of radioactivity or heat, as in the case of thermal pollution. Any body of water has the capacity to absorb, break down, or recycle introduced materials. Under normal circumstances, inorganic substances are widely dispersed and have little or no effect on life within the bodies of water into which they are released; organic materials are broken down by bacteria or other organisms and converted into a form in which they are useful to aquatic life. But, if the capacity of a body of water to dissolve, disperse, or recycle is exceeded, all additional substances or forms of energy become pollutants. Thus, thermal pollution, which is usually caused by the discharge of water that has been used as a coolant in fossil-fueled or nuclear-power plants, can favour a diversity of aquatic life in waters that would otherwise be too cold. In a warmer body of water, however, the addition of heat changes its characteristics and may make it less suited to species that are considered desirable.

Thermal pollution of water

Pollution may begin as water moves through the air, if the air is polluted. Soil erosion adds silt as a pollutant. The use of chemical fertilizers, pesticides, or other materials on watershed lands is an additional factor contributing to water pollution. The runoff from septic tanks and the outflow of manures from livestock feedlots along the watershed are sources of organic pollutants. Industries located along waterways downstream contribute a number of chemical pollutants, some of which are toxic if present in any concentration. Finally, cities and towns contribute their loads of sewage and other urban wastes. Thus, a community far upstream in a watershed may receive relatively clean water, whereas one farther downstream receives a partly diluted mixture of urban, industrial, and rural wastes. The cost of cleaning and purifying this water for community use may be high, and the process may be only partially effective. To add to the problem, the cities and towns in the lower, or downstream, regions of the river basin contribute additional wastes that flow into estuaries, creating new pollution problems.

The output of industries, agriculture, and urban communities generally exceeds the biologic capacities of aquatic systems, causing waters to become choked with an excess of organic substances and organisms to be poisoned by toxic materials. When organic matter exceeds the capacity of those microorganisms in water that break it down and recycle it, the excess of nutrients in such matter encour-

ages rapid growth, or blooms, of algae. When they die, the remains of the dead algae add further to the organic wastes already in the water; eventually, the water becomes deficient in oxygen. Anaerobic organisms (those that do not require oxygen to live) then attack the organic wastes, releasing gases such as methane and hydrogen sulfide, which are harmful to the oxygen-requiring (aerobic) forms of life. The result is a foul-smelling, waste-filled body of water, a situation that has already occurred in such places as Lake Erie and the Baltic Sea and is a growing problem in freshwater lakes of Europe and North America.

Eutrophication of bodies of water

The process by which a lake or any other body of water changes from a clean, clear condition—with a relatively low concentration of dissolved nutrients and a balanced aquatic community—to a nutrient-rich, algae-filled body and thence to an oxygen-deficient, waste-filled condition is known as accelerated eutrophication.

**Land pollution.** Land pollution involves the deposition on land of solid wastes—*e.g.,* used cars, cans, bottles, plastic containers, paper—that cannot be broken down quickly or, in some instances, cannot be broken down at all by the action of organic or inorganic forces. (The term biodegradable is used to describe those materials that can be decomposed and recycled by biologic action.) When such materials become concentrated within any one area, they interfere with organic life and create unsightly accumulations of trash. Methods of disposal other than recycling include ocean dumping, which creates water pollution and destroys marine habitats; landfill, which often requires the availability of low-lying ground and frequently involves the destruction of marshland or swamps that have high biologic value; or burning, which increases air pollution. Obviously, none of these methods is entirely satisfactory, although using landfill to create artificial landscapes, which then are covered with soil and planted with various kinds of vegetation, is a possibility that remains to be fully developed. It is the great quantities of debris produced by urban communities, more so than a shortage of raw materials, that forces the development of more effective means for recycling wastes. Land pollution also involves the accumulation on land of substances in a dispersed solid or liquid form that are injurious to life. This has been particularly noticeable with those chemicals (*e.g.,* DDT) that are spread for the purpose of exterminating pests but then accumulate to the extent that they can do damage to many other forms of life.

**Noise pollution.** One form of pollution that is characteristic of industrial societies is noise. The intensity of sound is measured in logarithmic units known as decibels; a change from a level of 10 decibels to one of 20 decibels actually represents a 100-fold increase in the sound level. At a level of 80 decibels, sound is annoying; permanent loss of hearing can be caused by steady exposure to noises in excess of 90 decibels, a level that is frequently exceeded by many common urban sounds, such as jackhammers, jet planes, and excessively loud music. In addition to causing loss of hearing, there is some evidence that noise can produce other deleterious effects on human health and on work performance.

Many large cities have taken measures to decrease the level of urban noise; the problem has received much attention with the advent of supersonic jet airplanes. These aircraft, which travel at speeds faster than the speed of sound, create sound waves (sonic booms) equivalent to those of major explosions and capable of damaging structures. The extent to which continuous exposure to sonic booms affects human health and functioning has yet to be determined. Nevertheless, in 1971 the U.S. Congress voted down appropriations to support the development of supersonic transport planes; Great Britain, France, and the Soviet Union, however, have developed such planes.

Sonic booms

**Chemical pollutants.** Among the most serious chemical pollutants are the chlorinated hydrocarbon pesticides, such as DDT, aldrin, and dieldrin; the polychlorinated biphenyls (PCBs), which are used in a variety of industrial processes and in the manufacture of many kinds of materials; and such metals as mercury, lead, cadmium, arsenic, and beryllium. All of these substances persist in the environment, being slowly, if at all, degraded by natural processes; in addition, all are toxic to life if they accumulate in any appreciable quantity.

The persistent pesticides have created serious ecological problems. As they move through successively higher organisms in food chains, they accumulate in increasingly concentrated forms at each level, causing damaging effects to the predators at the end of the chains—*i.e.,* they are present in low quantities in simple organisms but become more concentrated as these organisms are consumed by more complex ones, which are themselves consumed by predators. Among the species known to be adversely affected are such meat-eating birds as falcons, hawks, and eagles and such fish-eating birds as pelicans, petrels, cormorants, and egrets. The reproduction capacity of all of these birds has been affected by an accumulation of DDT or a similar compound in their tissues. This is manifested by an impairment in the ability of the females to form eggshells properly. As a result, some species lay soft-shelled or shell-less eggs that cannot be hatched, and there has been a general decline in the numbers of these birds in Europe, Japan, and North America. Although the effects of the same chemicals on mammals is less obvious and still a matter for investigation, some studies suggest that DDT can reduce the productivity of plant plankton, upon which all other marine life depends.

As the widespread biologic effects of pesticides become better known, there is growing evidence that they lose the ability to control the pests they were designed to kill. Many insect species have developed immunity to a wide range of synthetic pesticides, and the resistance is inherited by their offspring. Furthermore, it has been observed that repeated use of such chemicals creates pest populations in areas in which none previously existed. This happens because the pesticides destroy populations of carnivorous, predatory insects that previously kept the plant-eating insects in check.

Among other materials that are harmful to most forms of life are such metals as mercury, lead, and arsenic. The increasing release of these substances into the biosphere by industrial processes has created conditions that are now generally viewed as harmful to human welfare. Studies have been conducted on metallic pollutants to determine the normal environmental levels, the levels that are toxic to humans, and the extent to which industrial processes are responsible for the problem.

The ultimate control of pollution will presumably involve the decision not to allow the escape into the environment of the substances that are harmful to life, the decision to contain and recycle those substances that could be harmful if released into the environment in excessive quantities, and the decision not to release into the environment substances that persist and are toxic to living things. Essentially, therefore, pollution control does not mean an abandonment of existing productive human activities but their reordering so as to guarantee that their side effects do not outweigh their advantages.

## CONSERVATION AND GROWTH

The Earth has supported human civilization for more than 6,000 years and agriculture for twice that period. Before that, reaching back an unmeasured number of years into the past, human or near-human groups occupied various parts of the Earth, modifying them to some degree in the course of hunting, fishing, and food gathering. Patterns of land use were determined over many centuries of trial-and-error experimentation by people equipped with primitive tools who depended on the biologic communities of the Earth for their energy supplies. Today, however, with abundant fossil fuels, growing amounts of nuclear energy, and sophisticated tools and machines, it is possible to quickly modify entire landscapes, changing long-established natural patterns into new patterns with new purposes. The opportunity to enhance the material welfare and general well-being of great numbers of people is now available, as is the opportunity to cause great damage and to impair the capacity of the Earth to support life. The outcome will depend on changes in attitudes toward the use and conservation of the Earth's living and nonliving resources.

Changing patterns of land use

**The role of population, industry, and technology.** Uses of lands and resources are being modified in the expectation of continued population growth, industrial expansion, and accelerating technological change. Yet it is possible that in the future uses of lands and resources will take place in times of population stability, little industrial expansion, and a technology directed toward a reorganization and a rearrangement of activities to achieve a better environmental relationship. Although certain countries of the world have already reached some degree of population stability—*e.g.*, Ireland, Hungary, East Germany, France, Sweden, Switzerland, and Japan—industrial expansion and rapid technological change continue in these countries, in part because of the demands made by other expanding nations. The existing expansionist phase of technological civilization cannot, however, be expected to continue indefinitely. The ecological limitations on growth in a limited space with limited resources lead to predictions of an inevitable end to this expansion, even if mankind fails to voluntarily limit its own growth.

*Ecological considerations and advance planning.* Current decisions about land and resource use have important consequences for the future. If extensive areas of the Earth are badly damaged or their productivity destroyed by the expansion of technological civilization, they will be difficult, if not impossible, to restore. If a species becomes extinct, for example, it cannot be brought back. It is essential, therefore, that care be exercised in further modifying the planet to suit human purposes. Yet in many developing regions of the world, those where the greatest changes may be expected, little attention is being given to planning for and carefully controlling the use of land and resources. Thus, important tropical, semiarid, and subpolar regions of the Earth—the three principal climatic belts that have not yet undergone major technological development—are now being changed drastically without much consideration for their environments. In many parts of the world, ecologically trained experts are not available; in others, because of strong economic pressures toward development, ecological advice either is not sought or is ignored.

limatic
elts
urrently
eing
xploited

Comparatively speaking, the failure to apply current ecological knowledge to the changing land and resource use taking place in tropical, semiarid, and subarctic lands is equivalent to the modification of the more temperate lands that took place centuries ago, when ecological knowledge was not available. In earlier centuries, however, the capacity to do irreparable damage was restricted by the lack of machinery, industry, and fuel energy. Today, capabilities are such that major destruction can be accomplished quickly. There is, therefore, a need to call upon environmental expertise during the process of economic development in any area of the world if natural resources are to be conserved and the future welfare of humanity is to receive due consideration.

Some ecologists now believe that, although it was once possible to allow the development of a region to proceed more or less at random, based on individual wants, aspirations, and decisions about the use of lands and resources, such a process now holds too much risk for the well-being of society as a whole and for the future of the resources on which that society depends. Planning, they argue, must precede development, and regional planning is required if the use of major areas of land and its resources is to be brought into accord with environmental necessities and with the long-term needs of society.

In densely populated and technologically advanced nations, such as those of western and northern Europe, most of the land-use decisions that would affect large areas have already been made. Although changes do occur, mostly in relation to growing urbanization and increasing material wealth, it seems likely that the remaining woodlands and fields will continue to be devoted to their present uses.

ontrol
f land
se and
opulation
istribution

In England the interest of the central government in the planning and control of land use and population distribution was marked by the passage of the Town and Country Planning Act shortly after the end of World War II. This legislation led to decisions to limit the growth of London and to develop a pattern of new towns outside a greenbelt of agricultural and recreational land that surrounds the metropolis. In France, where there are still large areas of open space, a system for regional planning and control of land use and development (*aménagement du territoire*) has been formulated. It has already resulted in the establishment of new cities and recreational sites in previously undeveloped areas along the Mediterranean coast.

It is in the sparsely populated areas in the underdeveloped countries of Africa, Asia, and Latin America as well as in such technologically advanced countries as Canada, Australia, and the Soviet Union that the greatest range of options and choices for the future are available. Because these areas have yet to undergo drastic environmental change, the need for local and regional environmentally oriented planning for resource and land use is most urgent.

*Areas of promise.* Some of the current vexing conservation problems may be solved by technological developments. A highly technological society obviously requires an abundant and reliable source of energy. Research on nuclear fusion as a source of power indicates that this process could replace nuclear fission as a power source in some areas, but it appears to be too technologically demanding for widespread application. Solar energy, in its various modified forms, may be a more universally available source of power.

Apart from the development of major new sources of power, the greatest promise for the future of mineral resources and for the prevention of pollution of the environment lies in new technologies involving the recycling and reclamation of what are now considered waste products. Demands for new minerals will be greatly reduced when those already available in population centres can be reused more readily. Reclamation of sewage and other organic wastes and restoration of these materials to soils can help to arrest losses in soil fertility and structure and to reduce the need for new supplies of chemical nutrients for soil fertilization. If development of technologies for recycling and reutilization continues, many of the existing problems of environmental pollution will be solved.

In addition to the breeding of new strains of crop plants, the development of techniques for the control of diseases and pests of crop plants that do not involve the release of persistent, poisonous chemicals into the environment holds promise for the production of greatly increased quantities of food and fibre from smaller areas of the Earth's surface. Two such techniques are mixed cropping, in which different crops are planted within an area to contain the spread of pests, and integrated pest management, in which as many pest-control methods as possible are used in an ecologically harmonious manner to keep infestation within manageable limits. Much more intensive development of aquaculture (cultivation of the natural produce of water), perhaps utilizing coolant water from nuclear-power plants, can also produce much higher food yields from smaller areas than are now usually obtainable. As a result of these advances in intensive food production, agriculturally marginal lands and the wilder aquatic areas would be spared for the continued support of wild species as well as for the adventure and recreation of mankind, thus helping to solve one of the most troublesome of all conservation problems, the conservation of wild nature.

Possibility of aquaculture for food production

**Problems.** Considering the potential of new technology and the accompanying advances in science, it is possible to foresee a world in which a relatively stable human population can live at a high level of material affluence, with wild nature continuing to exist in abundance and relatively undisturbed lands available for human enjoyment. But this scientific and technological optimism is not supported by existing world conditions. Because knowledge now available is more than adequate to solve most of the world's major environmental problems, the problems are not those of science and technology but of the arrangements and functioning of human institutions and of the attitudes of individuals. Thus, while research in forestry science continues in all the forestry schools of the world, tropical forests are being devastated in ways that suggest that forestry science is still unknown. Although the techniques for managing livestock on natural ranges and pasturelands have reached a high level of sophistication, overgrazing continues around most of the world's major

deserts, and animals die of hunger, people suffer from deprivation, and the deserts spread. Obviously, the knowledge available does not reach or influence the behaviour of most of the pastoral people on Earth.

*Population growth.* Demographic predictions indicate that the population of the world will not be stabilized, even under the best of conditions, before it attains much higher levels. These predictions assume, of course, that there will be no major catastrophes—outbreaks of war, famine, or disease—that would cause drastic reductions in human numbers. There is little doubt that rapid population growth interferes with orderly economic development, leads to a deterioration of the human environment, places a severe strain on human institutions, and constitutes a growing threat to the survival of wild animal and plant life.

Necessity for curbing birthrates

Although techniques for birth control are effective and well known, they are unknown, unavailable, or unacceptable to those people having the most rapid rate of population growth—the ones who also live in the most precarious balance with their environment. This does not mean that the prospects for controlling population increase are poor; actually, they are better than at any time in the past. But more education is needed to encourage people to limit the size of families, and the prospects for material and economic advancement for those who have fewer children must be made more obvious.

*Pollution control.* Next to the widespread and growing loss of biological diversity through the destruction of biotic communities, the conservation problem of greatest magnitude is the control of pollution; it might even be argued that it is more urgent and important. The knowledge and technology needed to control pollution effectively are now available: pollution-free engines can be built, pollution-free factories have been put into operation, and techniques for controlling agricultural insect pests with a minimum use of persistent pesticides have been developed. For economic reasons, none of these measures, however, is being applied universally, and political and social pressures have not yet forced their application. Moreover, developing nations have expressed fear that excessive concern over pollution could impede their economic development. Indeed, some of these countries have become sanctuaries for industries that find it less expensive to operate in areas with more lax standards. It is apparent that pollution control, regardless of the state of its technology, will become a reality only when people demand it and only when nations are willing to agree on appropriate international standards.

*Control over use of resources.* Important also to the future of world conservation is the failure of most societies to exercise adequate controls over land, water, and other resource use. Effective means for controlling land use do not exist in most countries; laws and regulations that permit governments to exercise such control, when existent, often cannot be enforced because of the danger of strong public resentment and resistance. Although it is essential that lands and all other resources be used with a view to preserving their future productivity, this view all too often conflicts with present needs or demands of the resource users. The solution to this conflict is not within the scope of science or technology; instead, it is a question of attitudes and values, and these are less amenable to sudden change than laws or regulations. It appears that economic security and social stability are essential for people to look beyond immediate survival to the well-being of humanity and the future of life on this planet.

**BIBLIOGRAPHY**

*General works:* *World Resources 1986* (1986), an assessment of the resource base that supports the global economy, prepared by the World Resources Institute and the International Institute for Environment and Development and including data for 146 countries; ROBERT REPETTO (ed.), *The Global Possible: Resources, Development, and the New Century* (1985), a collection of papers on current problems and prospects for the 21st century; JOHN R. HOLLUM, *Topics and Terms in Environmental Problems* (1977), a basic reference source for understanding the terminology and general subjects related to resource conservation and environmental concerns; RAYMOND F. DASMANN, *Environmental Conservation,* 5th ed. (1984), a general textbook on the conservation of natural resources, stressing their interrelationships and use; FRANK FRASER DARLING, *Wilderness and Plenty* (1970), a series of lectures, presented by the British Broadcasting Company in 1969, that had a great effect on environmental thinking in Europe; FAIRFIELD OSBORN, *Our Plundered Planet* (1948), a classic in conservation, one of the first world views of the impact of human populations and land use on the natural resources of the Earth; WILLIAM L. THOMAS (ed.), *Man's Role in Changing the Face of the Earth* (1956, reprinted 1971), a comprehensive presentation of mankind's past, present, and probable future roles on the Earth, a classic in conservation; MARTIN W. HOLDGATE, MOHAMMED KASSAS, and GILBERT F. WHITE (eds.), *The World Environment 1972–1982: A Report by the United Nations Environment Programme* (1982); *Ocean Yearbook* (annual), a collection of papers dealing with resource management of the world's oceans, fisheries, and minerals, as well as the law of the sea; JOSEPH M. MORAN, MICHAEL D. MORGAN, and JAMES H. WIERSMA, *Introduction to Environmental Science,* 2nd ed. (1986), a source of information on ecology and environmental concerns; and EDWARD O. WILSON, *Biophilia* (1984), a study of interrelationships in life and lifelike processes.

*History of conservation:* RODERICK NASH, *Wilderness and the American Mind,* 3rd ed. (1982), a history of changes in American attitudes toward wilderness, beginning with the Old World roots and leading to the existing system of protected wilderness areas; ALFRED W. CROSBY, *Ecological Imperialism: The Biological Expansion of Europe, 900–1900* (1986), a historical account of the impact of European trade and colonialism upon traditional peoples and the natural environment. CLARENCE J. GLACKEN, *Traces on the Rhodian Shore: Nature and Culture in Western Thought from Ancient Times to the End of the Eighteenth Century* (1967, reprinted 1976), an account of the philosophies and ideas of nature; GEORGE P. MARSH, *Man and Nature: or, Physical Geography as Modified by Human Action* (1864, reissued 1965), a classic of conservation and the first general view of how people change the Earth and affect its living resources; RODERICK NASH (ed.), *The American Environment: Readings in the History of Conservation,* 2nd ed. (1976), a review of the evolution of conservation in the United States, with readings from the works of those who contributed to its development; HENRY D. THOREAU, *Walden: or, Life in the Woods* (1854, with many later editions available), a classic philosophical exploration of humanity's relationship with nature and the values to be found in living apart from civilization and its artifacts, by one of the first advocates of wilderness preservation; MICHAEL P. COHEN, *The Pathless Way: John Muir and American Wilderness* (1984), an analysis of preservationist thought on human beings in the environment; and STEWART L. UDALL, *The Quiet Crisis* (1963, reissued 1971), a review of the history of conservation and the growing crisis in conservation, with emphasis on public lands and living resources in the United States.

*Ecology:* CHARLES ELTON, *Animal Ecology* (1927, reissued 1966), a classic work in ecology and the forerunner to the application of animal ecology in the management of wildlife; NORMAN MYERS, *The Sinking Ark: A New Look at the Problem of Disappearing Species* (1979, reprinted 1983), an argument for the preservation of a balance of animal species; EUGENE P. ODUM, *Fundamentals of Ecology,* 3rd ed. (1971), a standard textbook on general ecology, which emphasizes the nature and functions of ecosystems and includes a discussion of the relationships between ecology and environmental problems; J.E. LOVELOCK, *Gaia, a New Look at Life on Earth* (1979), an investigation of the hypothesis that the self-regulating biosphere forms a complex system with the capacity to keep the Earth a fit place for life; NORMAN MYERS, *Gaia, an Atlas of Planet Management* (1984), and *The Primary Source: Tropical Forests and Our Future* (1984), studies by an environmentalist; DONALD WORSTER, *Nature's Economy: A History of Ecological Ideas,* new ed. (1985), an exploration of changing ideas about ecology from the 18th century to the present and their effect on environmental management; MICHAEL E. SOULÉ (ed.), *Conservation Biology: The Science of Scarcity and Diversity* (1986), a collection of papers concerning the application of ecological knowledge and population genetics to problems of conserving threatened species and maintaining the diversity of life on Earth; and FRANCIS R. THIBODEAU and HERMANN H. FIELD (eds.), *Sustaining Tomorrow: A Strategy for World Conservation and Development* (1984), a collection of essays on topics of preservation and conservation of ecological balance.

*Pollution:* RACHEL CARSON, *Silent Spring* (1962, reissued 1982), a popular best-seller that first alerted the general public to the dangers inherent in the widespread use of persistent pesticides; GINO G. MARCO, ROBERT M. HOLLINGWORTH, and WILLIAM DURHAM (eds.), *Silent Spring Revisited* (1987), a collection of essays on the topics posed in Rachel Carson's book; ROBERT L. RUDD, *Pesticides and the Living Landscape* (1964, reprinted 1970), a scholarly review of the ways in which pesticides enter into the ecological networks in natural and

man-made environments, with a discussion of the dangers involved in their continued use; *Environmental Quality* (annual), a regular report of the Council on Environmental Quality, which reviews the state of the U.S. environment with particular emphasis on problems of pollution; AMERICAN CHEMICAL SOCIETY, *Cleaning Our Environment: A Chemical Perspective,* 2nd ed. (1978); a report by the Council on Environmental Improvement on the technological state of the art of pollution control; BARRY COMMONER, *Science and Survival* (1966), a discussion of how science and technology can cause unanticipated changes in the environment, with particular reference to pollution from nuclear energy; DONALD L. BARLETT and JAMES B. STEELE, *Forevermore, Nuclear Waste in America* (1985), an investigative report on the environmental hazard; and SCIENTIFIC COMMITTEE ON PROBLEMS OF THE ENVIRONMENT (SCOPE), *Environmental Consequences of Nuclear War,* 2 vol. (1985–86), an international scientific review of the likely effects of nuclear war, prepared by a committee of the International Council of Scientific Unions.

*Population, resources, and technological development:* PAUL R. EHRLICH, *The Population Bomb,* rev. ed. (1971, reprinted 1978), a popular book that touched off the population debates of the late 1960s and early 1970s; PAUL R. EHRLICH, ANNE H. EHRLICH, and JOHN P. HOLDREN, *Ecoscience: Population, Resources, Environment* (1977), a study providing coverage of the major environmental and resource concerns and the scientific basis for their evaluation; GERALD O. BARNEY (ed.), *The Global 2000 Report to the President: Entering the Twenty-First Century,* 3 vol. (1980–81), an analysis of natural resource and environmental problems, with a forecast of conditions in the year 2000 considering existing trends and likely changes, prepared by the Council on Environmental Quality; AMORY B. LOVINS, *Soft Energy Paths: Toward a Durable Peace* (1977), an examination of the future prospects for fossil fuels, nuclear energy, and renewable energy resources; GEORG BORGSTROM, *The Hungry Planet: The Modern World at the Edge of Famine,* 2nd rev. ed. (1972), a review of human food needs and the prospects for meeting them, with particular attention to the use of ocean resources; M. TAGHI FARVAR and JOHN P. MILTON (eds.), *The Careless Technology: Ecology and International Development: The Record* (1972), a collection of case histories and related discussions concerning the unexpected side effects of economic and technological development; and WILLIAM VOGT, *Road to Survival* (1948), a classic work on the relationship between populations and resources, written from an emotional viewpoint in an attempt to draw attention to the coming population crisis and emphasizing the growing threat of world hunger, as well as the impact of humans on soils and renewable resources.

*Management of living resources:* WENDELL BERRY, *The Unsettling of America: Culture & Agriculture* (1977, reprinted 1986), an examination of the past, present, and future of agricultural land use; NORMAN MYERS, *A Wealth of Wild Species: Storehouse for Human Welfare* (1983), an exposition of concerns for the necessity of maintaining genetic and biologic diversity; RAYMOND F. DASMANN, *Wildlife Biology,* 2nd ed. (1981), an introductory college textbook reviewing the principles of wildlife biology and their application to the conservation and management of wild animals; STEPHEN HADEN-GUEST, JOHN K. WRIGHT, and EILEEN M. TECLAFF (eds.), *A World Geography of Forest Resources* (1956), a survey of the forest resources of the world, the extent to which they are being used, and the problems associated with their management; G.V. JACKS and R.O. WHYTE, *Vanishing Lands: A World Survey of Soil Erosion* (1939; U.K. title, *The Rape of the Earth;* reprinted 1972), a classic work that reviews the nature of world soils and describes their destruction by erosion as a result of misuse; ALDO LEOPOLD, *Game Management* (1933, reprinted 1986), the first textbook of wildlife management, a conservation classic, and his *Sand County Almanac and Sketches Here and There* (1949, reprinted 1981), a study of the ethics and aesthetics of conservation, with emphasis on land, wildlife, and wilderness; LAURENCE A. STODDART, ARTHUR D. SMITH, and THADIS W. BOX, *Range Management,* 3rd ed. (1975), a textbook on the principles and practices affecting the use of range and pasturelands by grazing animals.

(R.F.D.)

# Constantine the Great

Constantine I, the first Roman emperor to profess Christianity, not only initiated the evolution of the empire into a Christian state but also provided the impulse for a distinctively Christian culture that prepared the way for the growth of Byzantine and Western medieval culture. He was born on February 27, probably in the later AD 280s, at Naissus (modern Niš, Yugos.). A typical product of the military governing class of the later 3rd century, he was the son of Flavius Valerius Constantius, an army officer, and his wife (or concubine) Helena; his full Latin name was Flavius Valerius Constantinus. In AD 293 his father was raised to the rank of Caesar, or deputy emperor (as Constantius I Chlorus), and was sent to serve under Augustus (emperor) Maximian in the West. In 289 Constantius had separated from Helena in order to marry a stepdaughter of Maximian, and Constantine was brought up in the Eastern Empire at the court of the senior emperor Diocletian at Nicomedia (modern İzmit, Tur.). Constantine was seen as a youth by his future panegyrist, Eusebius, bishop of Caesarea, passing with Diocletian through Palestine on the way to a war in Egypt.



Constantine the Great, marble bust (restored), 4th century AD. In the Uffizi, Florence.
By courtesy of the Galleria degli Uffizi, Florence

**Career and conversion.** Constantine's experience as a member of the imperial court—a Latin-speaking institution—in the Eastern provinces left a lasting imprint on him. Educated to less than the highest literary standards of the day, he was always more at home in Latin than in Greek: later in life he had the habit of delivering edifying sermons, which he would compose in Latin and pronounce in Greek from professional translations. Christianity he encountered in court circles as well as in the cities of the East; and from 303, during the great persecution of the Christians that began at the court of Diocletian at Nicomedia and was enforced with particular intensity in the eastern parts of the empire, Christianity was a major issue of public policy. It is even possible that members of Constantine's family were Christians.

In 305 the two emperors, Diocletian and Maximian, abdicated, to be succeeded by their respective deputy emperors, Galerius and Constantius. The latter were replaced by Galerius Valerius Maximinus in the East and Flavius Valerius Severus in the West, Constantine being passed over. Constantius requested his son's presence from Galerius, and Constantine made his way through the territories of the hostile Severus to join his father at Gesoriacum (modern Boulogne, Fr.). They crossed together to Britain and fought a campaign in the north before Constantius' death at Eboracum (modern York) in 306. Immediately acclaimed emperor by the army, Constantine then threw himself into a complex series of civil wars in which Maxentius, the son of Maximian, rebelled at Rome; with his father's help, Maxentius suppressed Severus, who had been proclaimed Western emperor by Galerius and who was then replaced by Licinius. When Maximian was rejected by his son, he joined Constantine in Gaul, only to betray Constantine and to be murdered or forced to commit suicide (310). Constantine, who in 307 had married Maximian's daughter Fausta as his second wife, invaded Italy in 312 and after a lightning campaign defeated his brother-in-law Maxentius at the Milvian Bridge near Rome. He then confirmed an alliance that he had already entered into with Licinius (Galerius having died in 311): Constantine became Western emperor and Licinius shared the East with his rival Maximinus. Licinius defeated Maximinus and became the sole Eastern emperor but lost territory in the Balkans to Constantine in 316. After a further period of tension, Constantine attacked Licinius in 324, routing him at Adrianople and Chrysopolis (respectively, modern Edirne and Üsküdar, Tur.) and becoming sole emperor of East and West.

Acclaimed emperor by troops

Throughout his life, Constantine ascribed his success to his conversion to Christianity and the support of the Christian God. The triumphal arch erected in his honour at Rome after the defeat of Maxentius ascribed the victory to the "inspiration of the Divinity" as well as to Constantine's own genius. A statue set up at the same time showed Constantine himself holding aloft a cross and the legend "By this saving sign I have delivered your city from the tyrant and restored liberty to the Senate and people of Rome." After his victory over Licinius in 324, Constantine wrote that he had come from the farthest shores of Britain as God's chosen instrument for the suppression of impiety, and in a letter to the Persian king Shāpūr II he proclaimed that, aided by the divine power of God, he had come to bring peace and prosperity to all lands.

Constantine's adherence to Christianity was closely associated with his rise to power. He fought the Battle of the Milvian Bridge in the name of the Christian God, having received instructions in a dream to paint the Christian monogram (☧) on his troops' shields. This is the account given by the Christian apologist Lactantius; a somewhat different version, offered by Eusebius, tells of a vision seen by Constantine during the campaign against Maxentius, in which the Christian sign appeared in the sky with the legend "In this sign, conquer." Despite the Emperor's own authority for the account, given late in life to Eusebius, it is in general more problematic than the other; but a religious experience on the march from Gaul is suggested also by a pagan orator, who in a speech of 310 referred to a vision of Apollo received by Constantine at a shrine in Gaul.

"In this sign, conquer"

Yet to suggest that Constantine's conversion was "politically motivated" means little in an age in which every Greek or Roman expected that political success followed from religious piety. The civil war itself fostered religious competition, each side enlisting its divine support, and it would be thought in no way unusual that Constantine should have sought divine help for his claim for power and divine justification for his acquisition of it. What is remarkable is Constantine's subsequent development of his new religious allegiance to a strong personal commitment.

**Commitment to Christianity.** Shortly after the defeat of Maxentius, Constantine met Licinius at Mediolanum (modern Milan) to confirm a number of political and dynastic arrangements. A product of this meeting has

become known as the Edict of Milan, which extended toleration to the Christians and restored any personal and corporate property that had been confiscated during the persecution. The extant copies of this decree are actually those posted by Licinius in the eastern parts of the empire. But Constantine went far beyond the joint policy agreed upon at Mediolanum. By 313 he had already donated to the Bishop of Rome the imperial property of the Lateran, where a new cathedral, the Basilica Constantiniana (now S. Giovanni in Laterano), soon rose. The Church of St. Sebastian was also probably begun at this time, and it was in these early years of his reign that Constantine began issuing laws conveying upon the church and its clergy fiscal and legal privileges and immunities from civic burdens. As he said in a letter of 313 to the proconsul of Africa, the Christian clergy should not be distracted by secular offices from their religious duties ". . . for when they are free to render supreme service to the Divinity, it is evident that they confer great benefit upon the affairs of state." In another such letter, directed to the Bishop of Carthage, Constantine mentioned the Spanish bishop Hosius, who was important later in the reign as his adviser and possibly—since he may well have been with Constantine in Gaul before the campaign against Maxentius—instrumental in the conversion of the Emperor.

'ublic and ersonal theology"

Constantine's personal "theology" emerges with particular clarity from a remarkable series of letters, extending from 313 to the early 320s, concerning the Donatist schism in North Africa. The Donatists maintained that those priests and bishops who had once lapsed from the Christian faith could not be readmitted to the church. Constantine's chief concern was that a divided church would offend the Christian God and so bring divine vengeance upon the Roman Empire and Constantine himself. Schism, in Constantine's view, was inspired by Satan. Its partisans were acting in defiance of the clemency of Christ, for which they might expect eternal damnation at the Last Judgment. Meanwhile, it was for the righteous members of the Christian community to show patience and long-suffering. In so doing they would be imitating Christ, and their patience would be rewarded in lieu of martyrdom—for actual martyrdom was no longer open to Christians in a time of peace for the church. Throughout, Constantine had no doubt that to remove error and to propagate the true religion were both his personal duty and a proper use of the imperial position. His claim to be "bishop of those outside the church" may be construed in this light. Other such pronouncements, expressed in letters to imperial officials and to Christian clergy, demonstrate that Constantine's commitment to Christianity was firmer and less ambiguous than some have suggested. Eusebius confirmed what Constantine himself believed: that he had a special and personal relationship with the Christian God.

Constantine's second involvement in an ecclesiastical issue followed the defeat of Licinius; but the Arian heresy, with its intricate explorations of the precise nature of the Trinity that were couched in difficult Greek, was as remote from Constantine's educational background as it was from his impatient, urgent temperament. The Council of Nicaea, which opened in the early summer of 325 with an address by the Emperor, had already been preceded by a letter to the chief protagonist, Arius of Alexandria, in which Constantine stated his opinion that the dispute was fostered only by excessive leisure and academic contention, that the point at issue was trivial and could be resolved without difficulty. His optimism was not justified: neither this letter nor the Council of Nicaea itself nor the second letter, in which Constantine urged acceptance of its conclusions, was adequate to solve a dispute in which the participants were as intransigent as the theological issues were subtle. Indeed, for more than 40 years after the death of Constantine, Arianism was actually the official orthodoxy of the Eastern Empire.

Arianism and the Council of Nicaea

The Council of Nicaea coincided almost exactly with the celebrations of the 20th anniversary of the reign of Constantine, at which, returning the compliment paid by the Emperor's attendance at their council, the bishops were honoured participants. But Constantine's visit to the West in 326, to repeat the celebrations at Rome, brought the greatest political crisis of the reign. During his absence from the East, and for reasons that remain obscure, Constantine had his eldest son, the deputy emperor Crispus, and his own wife Fausta, Crispus' stepmother, slain. Nor was the visit to Rome a success. Constantine's refusal to take part in a pagan procession offended the Romans; and when he left after a short visit, it was never to return.

**Final years.** These events set the course of the last phase of the reign of Constantine. After his defeat of Licinius he had renamed Byzantium as Constantinople: immediately upon his return from the West he began to rebuild the city on a greatly enlarged pattern, as his permanent capital and the "second Rome." The dedication of Constantinople (May 330) confirmed the divorce, which had been in the making for more than a century, between the emperors and Rome. Rome had long been unsuited to the strategic needs of the empire: it was now to be left in splendid isolation, as an enormously wealthy and prestigious city—still the emotional focus of the empire—but of limited political importance.

Founding of Constantinople

It was perhaps in some sense to atone for the family catastrophe of 326 that Constantine's mother, Helena, embarked on a pilgrimage to the Holy Land. Her journey was attended by almsgiving and pious works and was distinguished by her church foundations at Jerusalem and at Bethlehem. By the initiative of Eutropia, Constantine's mother-in-law, a church was also built at Mamre, where, according to an interpretation of Genesis shared by Constantine and Eusebius, Christ had first shown himself to men in God's appearance to Abraham; but the most famous of these foundations followed the sensational discovery of the Holy Sepulchre at Jerusalem. The discovery was taken up with enthusiasm by Constantine, who instigated the building of a great new basilica at the spot, offering unlimited help with labour and materials and suggestions as to design and decoration.

Constantine's interest in church building was expressed also at Constantinople, particularly in churches of the Holy Wisdom (the original Hagia Sophia) and of the Apostles. At Rome, the great church of St. Peter was begun in the later 320s and lavishly endowed by Constantine with plate and property. Meanwhile, churches at Trier, Aquileia, Cirta in Numidia, Nicomedia, Antioch, Gaza, Alexandria, and elsewhere owed their development, directly or indirectly, to Constantine's interest.

The Emperor was an earnest student of his religion. Even before the defeat of Licinius he had summoned to Trier the theologian and polemicist Lactantius, to be the tutor of Crispus. In later years, he commissioned new copies of the Bible for the growing congregations at Constantinople. He composed a special prayer for his troops and went on campaigns with a mobile chapel in a tent. He issued numerous laws relating to Christian practice and susceptibilities: for instance, abolishing the penalty of crucifixion and the practice of branding certain criminals; enjoining the observance of Sunday and saints' days; and extending privileges to the clergy while suppressing at least some offensive pagan practices.

Constantine had hoped to be baptized in the Jordan River, but perhaps because of the lack of opportunity to do so—together possibly with the reflection that his office necessarily involved responsibility for actions hardly compatible with the baptized state—he delayed the ceremony until the end of his life. It was while preparing for a campaign against Persia that he fell ill at Helenopolis. When treatment failed, he made to return to Constantinople but was forced to take to his bed near Nicomedia. There, Constantine received baptism, putting off the imperial purple for the white robes of a neophyte; and he died on May 22, 337. He was buried at Constantinople in his Church of the Apostles, whose memorials, six on each side, flanked his tomb. Yet this was less an expression of religious megalomania than of Constantine's literal conviction that he was the successor of the evangelists, having devoted his life and office to the spreading of Christianity.

Death

**Assessment.** The reign of Constantine must be interpreted against the background of his personal commitment to Christianity. His public actions and policies, however, were not entirely without ambiguity. Roman opinion ex-

pected of its emperors not innovation but the preservation of traditional ways; Roman propaganda and political communication were conditioned, by statement, allusion, and symbol, to express these expectations. It is significant, for instance, not that the pagan gods and their legends survived for a few years on Constantine's coinage but that they disappeared so quickly: the last of them, the relatively inoffensive "Unconquered Sun," was eliminated just over a decade after the defeat of Maxentius.

Some of the ambiguities in Constantine's public policies were therefore exacted by the respect due to established practice and by the difficulties of expressing, as well as of making, total changes suddenly. The suppression of paganism, by law and by the sporadic destruction of pagan shrines, is balanced by particular acts of deference. A town in Asia Minor mentioned the unanimous Christianity of its inhabitants in support of a petition to the Emperor; while, on the other hand, one in Italy was allowed to hold a local festival incorporating gladiatorial games and to found a shrine of the imperial dynasty—although direct religious observance there was firmly forbidden. In an early law of Constantine, priests and public soothsayers of Rome were prohibited entry to private houses; but another law, of 320 or 321, calls for their recital of prayer "in the manner of ancient observance" if the imperial palace or any other public building were struck by lightning. Traditional country magic was tolerated by Constantine. Classical culture and education, which were intimately linked with paganism, continued to enjoy enormous prestige and influence; provincial priesthoods, which were as intimately linked with civic life, long survived the reign of Constantine. Constantinople itself was predominantly a Christian city, its dedication celebrated by Christian services; yet its foundation was also attended by a well-known pagan seer, Sopatros.

Secular achievements

An objective assessment of Constantine's secular achievements is not easy—partly because of the predominantly religious significance with which the Emperor himself invested his reign, partly because the restlessly innovatory character that dissenting contemporaries saw in his religious policy was also applied by them to the interpretation of his secular achievement. Some of Constantine's contributions can, in fact, be argued to have been already implicit in the trends of the last half century. So may be judged the further development, taking place in his reign, of the administrative court hierarchy and an increasing reliance upon a mobile field army, to what was considered the detriment of frontier garrisons. The establishment by Constantine of a new gold coin, the solidus, which was to survive for centuries as the basic unit of Byzantine currency, could hardly have been achieved without the work of his predecessors in restoring political and military stability after the anarchy of the 3rd century. Perhaps more directly linked with Constantine's own political and dynastic policies was the emergence of regional praetorian prefectures with supreme authority over civil financial administration but with no direct control over military affairs; this they yielded to new *magistri,* or "masters," of the cavalry and infantry forces. The reduction of the prefects' powers was seen by some as excessively innovatory, but the principle of the division of military and civil power had already been established by Diocletian. A real innovation, from which Constantine could expect little popularity, was his institution of a new tax, the *collatio lustralis.* It was levied every five years upon trade and business and seems to have become genuinely oppressive.

A lavish spender, Constantine was notoriously open-handed to his supporters and was accused of promoting beyond their deserts men of inferior social status. More to the point is the accusation that his generosity was only made possible by his looting of the treasures of the pagan temples as well as by his confiscations and new taxes; and there is no doubt that some of his more prominent supporters owed their success, at least partly, to their timely adoption of the Emperor's religion.

The foundation of Constantinople, an act of crucial long-term importance, was Constantine's personal achievement. Yet it, too, had been foreshadowed; Diocletian enhanced Nicomedia to an extent that was considered to challenge Rome. The city itself exemplified the "religious rapacity" of the Emperor, being filled with the artistic spoils of the Greek temples, while some of its public buildings and some of the mansions erected for Constantine's supporters soon showed signs of their hasty construction. Its Senate, created to match that of Rome, long lacked the aristocratic pedigree and prestige of its counterpart.

In military policy Constantine enjoyed unbroken success, with triumphs over the Franks, Sarmatians, and Goths to add to his victories in the civil wars; the latter, in particular, show a bold and imaginative mastery of strategy. Constantine was totally ruthless toward his political enemies, while his legislation, apart from its concessions to Christianity, is notable mainly for a brutality that became characteristic of late Roman enforcement of law. Politically, Constantine's main contribution was perhaps that, in leaving the empire to his three sons, he reestablished a dynastic succession, but it was secured only by a sequence of political murders after his death.

Above all, Constantine's achievement was perhaps greatest in social and cultural history. It was the development, after his example, of a Christianized imperial governing class that, together with his dynastic success, most firmly entrenched the privileged position of Christianity; and it was this movement of fashion, rather than the enforcement of any program of legislation, that was the basis of the Christianization of the Roman Empire. Emerging from it in the course of the 4th century were two developments that contributed fundamentally to the nature of Byzantine and Western medieval culture: the growth of a specifically Christian, biblical culture that took its place beside the traditional Classical culture of the upper classes; and the extension of new forms of religious patronage, between the secular governing classes and bishops, Christian intellectuals and holy men. Constantine left much for his successors to do, but it was his personal choice made in 312 that determined the emergence of the Roman Empire as a Christian state. It is not hard to see why Eusebius regarded Constantine's reign as the fulfillment of divine providence—nor to concede the force of Constantine's assessment of his own role as that of the 13th Apostle.

**BIBLIOGRAPHY**

*Biographies and commentaries:* Two accounts remain classic: ch. 14–18 of EDWARD GIBBON, *The History of the Decline and Fall of the Roman Empire*, edited by J.B. BURY, vol. 1 and 2 (1896), available also in many later editions; and JACOB BURCKHARDT, *The Age of Constantine the Great* (1949, reissued 1983; originally published in German, 1880). NORMAN H. BAYNES, *Constantine the Great and the Christian Church* (1930, reprinted 1975), is still a fundamental study, emphasizing the authenticity of Constantine's own writings. See also A.H.M. JONES, *Constantine and the Conversion of Europe* (1948, reissued 1978), and *The Later Roman Empire, 284–602,* 2 vol. (1964, reprinted 1986); ANDREW ALFÖLDI, *The Conversion of Constantine and Pagan Rome*, trans. from German (1948, reprinted 1969); JOSEPH VOGT, *Constantin der Grosse und sein Jahrhundert*, 2nd rev. ed. (1960); and RAMSAY MACMULLEN, *Constantine* (1969, reissued 1971). TIMOTHY D. BARNES, *Constantine and Eusebius* (1981), and *The New Empire of Diocletian and Constantine* (1982), are basic reappraisals of the political and religious background of Constantine's career. On the foundation of Constantinople, see GILBERT DAGRON, *Naissance d'une capitale: Constantinople et ses institutions de 330 á 451* (1974). On Constantine's church building, see RICHARD KRAUTHEIMER, *Early Christian and Byzantine Architecture*, 3rd ed. rev. (1981).

*Sources:* For Constantine's letters, see especially the modern translations of EUSEBIUS, *Church History,* Book X, and *Life of Constantine,* both in *A Select Library of Nicene and Post-Nicene Fathers of the Christian Church, Second Series,* vol. 1 (1961); and LACTANTIUS, *De Mortibus Persecutorum*, edited and translated by J.L. CREED (1984), part of the "Oxford Early Christian Texts" series. Eusebius' panegyrics of Constantine are translated and discussed in H.A. DRAKE, *In Praise of Constantine: A Historical Study and New Translation of Eusebius' Tricennial Orations* (1976). The ancient secular accounts are scanty; the fullest account, although with a hostile bias, is ZOSIMUS, *Historia nova*, available in a modern translation by RONALD T. RIDLEY, *New History* (1982). The bulk of Constantine's surviving legislation is in the *Codex Theodosianus*, in an edition translated by CLYDE PHARR, *The Theodosian Code and Novels, and the Sirmondian Constitutions* (1952, reissued 1969).

(J.F.Ma./D.M.N.)

# Constitution and Constitutional Government

The general idea of a constitution and of constitutionalism originated with the ancient Greeks and especially in the systematic, theoretical, normative, and descriptive writings of Aristotle. In his *Politics, Nicomachean Ethics, Constitution of Athens,* and other works, Aristotle used the Greek word for constitution (*politeia*) in several different senses. The simplest and most neutral of these was "the arrangement of the offices in a *polis*" (state). In this purely descriptive sense of the word, every state has a constitution, no matter how badly or erratically governed it may be.

This article deals with the theories and classical conceptions of constitutions as well as the features and practice of constitutional government throughout the world. It is divided into the following sections:

Theories about constitutions 690
Influence of the church 690
The social contract 690
Features of constitutional government 691
Constitutionality 692
Constitutional change 692
Constitutional stability 692
The practice of constitutional government 693
Great Britain 693
United States 693
Europe 694
Latin America, Africa, and Asia 694
Bibliography 694

## THEORIES ABOUT CONSTITUTIONS

Aristotle's classification of the "forms of government" was intended as a classification of constitutions, both good and bad. Under good constitutions—monarchy, aristocracy, and the mixed kind to which Aristotle applied the same term *politeia*—one person, a few individuals, or the many rule in the interest of the whole *polis.* Under the bad constitutions—tyranny, oligarchy, and democracy—the tyrant, the rich oligarchs, or the poor *dēmos,* or people, rule in their own interest alone.

Aristotle regarded the mixed constitution as the best arrangement of offices in the *polis.* Such a *politeia* would contain monarchic, aristocratic, and democratic elements. Its citizens, after learning to obey, were to be given opportunities to participate in ruling. This was a privilege only of citizens, however, since neither noncitizens nor slaves would have been admitted by Aristotle or his contemporaries in the Greek city-states. Aristotle regarded some humans as natural slaves, a point on which later Roman philosophers, especially the Stoics and jurists, disagreed with him. Although slavery was at least as widespread in Rome as in Greece, Roman law generally recognized a basic equality among all humans. This was because, the Stoics argued, all humans are endowed by nature with a spark of reason by means of which they can perceive a universal natural law that governs all the world and can bring their behaviour into harmony with it.

Influence of Roman law

Roman law thus added to Aristotelian notions of constitutionalism the concepts of a generalized equality, a universal regularity, and a hierarchy of types of laws. Aristotle had already drawn a distinction between the constitution (*politeia*), the laws (*nomoi*), and something more ephemeral that corresponds to what could be described as day-to-day policies (*psēphismata*). The latter might be based upon the votes cast by the citizens in their assembly and might be subject to frequent changes, but *nomoi,* or laws, were meant to last longer. The Romans conceived of the all-encompassing rational law of nature as the eternal framework to which constitutions, laws, and policies should conform—the constitution of the universe.

**Influence of the church.** Christianity endowed this universal constitution with a clearly monarchical cast. The Christian God, it came to be argued, was the sole ruler of the universe, and his laws were to be obeyed. Christians were under an obligation to try to constitute their earthly cities on the model of the City of God.

Both the church and the secular authorities with whom the church came into conflict in the course of the Middle Ages needed clearly defined arrangements of offices, functions, and jurisdictions. Medieval constitutions, whether of church or state, were considered legitimate because they were believed to be ordained of God or tradition or both. Confirmation by officers of the Christian Church was regarded as a prerequisite of the legitimacy of secular rulers. Coronation ceremonies were incomplete without a bishop's participation. The Holy Roman emperor travelled to Rome in order to receive his crown from the pope. Oaths, including the coronation oaths of rulers, could be sworn only in the presence of the clergy because oaths constituted promises to God and invoked divine punishment for violations. Even in an imposition of a new constitutional order, novelty could always be legitimized by reference to an alleged return to a more or less fictitious "ancient constitution." It was only in Italy during the Renaissance and in England after the Reformation that the "great modern fallacy" (as the Swiss historian Jacob Burckhardt called it) was established, according to which citizens could rationally and deliberately adopt a new constitution to meet their needs.

**The social contract.** The theoretical foundations of modern constitutionalism were laid down in the great works on the social contract, especially those of the English philosophers Thomas Hobbes and John Locke in the 17th century and the French philosopher Jean-Jacques Rousseau in the 18th.

Impact of the Reformation

As a result of the Reformation the basis of divinely sanctioned contractual relations was broken up. The Holy Roman Empire was torn apart by the wars of the Reformation. Henry VIII made the Church of England independent of Rome. In these circumstances, it became necessary to search for a new basis of order and stability, loyalty and obedience. In their search, political theorists—and especially the Protestants among them—turned to the old biblical concept of a covenant or contract, such as the one between God and Abraham and the Israelites of the Old Testament.

In a sense, the secular theorists of the social contract almost reversed the process of choice. Instead of God choosing his people, a people through its representatives was now looked upon as choosing its governors, or its mode of governance, under God, by means of a social contract or constitution. According to modern theories of the social contract, the political unit is nevertheless established as in the biblical model by means of a promise or promises.

Thomas Hobbes's state, or "Leviathan," comes into being when its individual members renounce their powers to execute the laws of nature, each for himself, and promise to turn these powers over to the sovereign—which is created as a result of this act—and to obey thenceforth the laws made by this sovereign. These laws enjoy authority because individual members of society are in effect their co-authors. According to Locke, individuals promise to agree to accept the judgments of a common judge (the legislature) when they accede to the compact that establishes civil society. After this (in one interpretation of Locke's *Second Treatise on Civil Government*), another set

of promises is made—between the members of the civil society, on the one hand, and the government, on the other. The government promises to execute its trust faithfully, leaving to the people the right to rebel in case the government breaks the terms of the contract, or, in other words, violates the constitution. Subsequent generations accept the terms of the compact by accepting the inheritance of private property that is created and protected by the compact. Anyone who rejects the constitution must leave the territory of the political unit and go *in vacuis locis,* or "empty places"—America, in Locke's time. In his *Letters on Toleration,* Locke characteristically excluded atheists from religious toleration because they could be expected either not to take the original contractual oath or not to be bound by the divine sanctions invoked for its violation. For Rousseau, too, the willingness to subject oneself to the "general will" to which only the popular sovereign can give expression is the essential ingredient of the social contract. In taking this position, Rousseau may have been influenced by the experience of his native Geneva. The Swiss Confederation is still referred to officially, in German, as an *Eidgenossenschaft,* a term best translated as "fellowship of the oath."

Hobbes's view on sovereignty

Hobbes's main contribution to constitutionalism lies in his radical rationalism. Individuals, according to Hobbes, come together out of the state of nature, which is a state of disorder and war, because their reason tells them that they can best ensure their self-preservation by giving all power to a sovereign. The sovereign may consist of a single person, an assembly, or the whole body of citizens; but regardless of its form, all the powers of sovereignty have to be combined and concentrated in it. Hobbes held that any division of these powers destroyed the sovereign and thereby returned the members of the commonwealth to the state of nature, in which the condition of man is ". . . solitary, poore, nasty, brutish, and short." Hobbes therefore preferred the singular sovereign since he was less likely than an assembly or than the whole body of citizens to become internally or functionally divided. The individual should retain only his natural rights, which he cannot surrender into the common pool of sovereign powers. These rights include the right against self-incrimination, the right to purchase a substitute for compulsory military service, and the right to act freely in instances in which the laws are silent.

Locke attempted to provide firm assurance of the individual's natural rights, partly by assigning separate though coordinated powers to the monarch and Parliament and partly by reserving the right of revolution against a government that had become unconstitutionally oppressive. Locke did not use the word sovereignty. In this as in other respects, he remained within the English constitutional tradition, which had eschewed the concentration of all powers in a single organ of government. The closest that English constitutionalists came to identifying the centre of sovereign power was in the phrase, used frequently from the 16th century onward, the king (or queen) in Parliament.

Rousseau's theory of the general will

Whereas Hobbes created his unitary sovereign through the mechanism of individual and unilateral promises and whereas Locke prevented excessive concentration of power by requiring the cooperation of different organs of government for the accomplishment of different purposes, Rousseau merged all individual citizens into an all-powerful sovereign whose main purpose was the expression of the general will. By definition. the general will can never be wrong; for when something contrary to the general interest is expressed, it is defined as the mere "will of all" and cannot have emanated from the sovereign. In order to guarantee the legitimacy of government and laws, Rousseau would have enforced universal participation in order to "force men to be free," as he paradoxically phrased it. In common with Hobbes and Locke, Rousseau required the assent of all to the original social contract. He required smaller majorities for the adoption of laws of lesser importance than the constitution itself. His main concern was to provide for legitimacy through universal participation in legislation, whereas Locke and Hobbes were more concerned to provide constitutional stability through consent. As a result, Rousseau's thought appears to be more democratic than that of his English predecessors. He has even been accused of laying the philosophical foundations of "totalitarian democracy," for the state he describes in *The Social Contract* would be subject, at the dictates of its universal and unanimous sovereign, to sudden changes, or even transformations, of its constitution.

Influence of social contract theory

In the political thought of Hobbes, Locke, and Rousseau may be found theoretical consideration of the practical issues that were to confront the authors of the American and French constitutions. The influence of theories of the social contract, especially as they relate to the issues of natural rights and the proper functions of government, pervades the constitution making of the revolutionary era that began with the U.S. War of Independence and is indeed enshrined in the great political manifestos of the time, the American Declaration of Independence and Bill of Rights, and the French Declaration of the Rights of Man and the Citizen.

The constitutional experience of these two countries, and, of course, of England, had great influence on liberal thought in Europe and other parts of the world during the 19th century and found expression in the constitutions that were demanded of the European monarchies. The extent to which the ideal of constitutional democracy has become entwined with the practice of constitutional government will be apparent from the examination in the following section of the main features of constitutional government.

## FEATURES OF CONSTITUTIONAL GOVERNMENT

Virtually all contemporary governments have constitutions, but possession and publication of a constitution does not make a government constitutional. Constitutional government is in fact comprised of the following elements.

*Procedural stability.* Certain fundamental procedures must not be subject to frequent or arbitrary change. Citizens must know the basic rules according to which politics are conducted. Stable procedures of government provide citizens with adequate knowledge of the probable consequences of their actions. By contrast, under many nonconstitutional regimes, such as Hitler's in Germany and Stalin's in the Soviet Union, individuals, including high government officials, never knew from one day to the next whether the whim of the dictator's will would not turn today's hero into tomorrow's public enemy.

*Accountability.* Under constitutional government, those who govern are regularly accountable to at least a portion of the governed. In a constitutional democracy, this accountability is owed to the electorate by all persons in government. Accountability can be enforced through a great variety of regular procedures, including elections, systems of promotion and discipline, fiscal accounting, recall, and referendum. In constitutional democracies, the accountability of government officials *to* the citizenry makes possible the citizens' responsibility *for* the acts of government. The most obvious example of this two-directional flow of responsibility and accountability is the electoral process. A member of the legislature or the head of government is elected by adult citizens and is thereby invested with authority and power in order that he may try to achieve those goals to which he committed himself in his program. At the end of his term of office, the electorate has the opportunity to judge his performance and to reelect him or dismiss him from office. The official has thus rendered his account and has been held accountable.

*Representation.* Those in office must conduct themselves as the representatives of their constituents. To represent means to be present on behalf of someone else who is absent. Elections, of course, are not the only means of securing representation or of ensuring the representativeness of a government. Hereditary medieval kings considered themselves, and were generally considered by their subjects, to be representatives of their societies. Of the social contract theorists only Rousseau denied the feasibility of representation for purposes of legislation. The elected status of officeholders is sometimes considered no guarantee that they will be "existentially representative" of their constituents, unless they share with the latter certain

other vital characteristics such as race, religion, sex, or age. The problems of representation are in fact more closely related to democratic than to constitutionalist criteria of government: a regime that would be considered quite unrepresentative by modern standards could still be regarded as constitutional so long as it provided procedural stability and the accountability of officeholders to some but not all of the governed and so long as the governors were representative of the best or the most important elements in the body politic.

*Division of power.* Constitutional government requires a division of power among several organs of the body politic. Preconstitutionalist governments, such as the absolute monarchies of Europe in the 18th century, frequently concentrated all power in the hands of a single person. The same has been true in modern dictatorships such as Hitler's in Germany. Constitutionalism, on the other hand, by dividing power—between, for example, local and central government and between the legislature, executive, and judiciary—ensures the presence of restraints and "checks and balances" in the political system. Citizens are thus able to influence policy by resort to any of several branches of government.

*Openness and disclosure.* Democracy rests upon popular participation in government, constitutionalism upon disclosure of and openness about the affairs of government. In this sense, constitutionalism is a prerequisite of successful democracy, since the people cannot participate rationally in government unless they are adequately informed of its workings. Originally, because they were concerned with secrets of state, bureaucracies surrounded their activities with a veil of secrecy. The ruler himself always retained full access to administrative secrets and often to the private affairs of his subjects, into which bureaucrats such as tax collectors and the police could legally pry. But when both administrators and rulers were subjected to constitutional restraints, it became necessary that they disclose the content of their official activities to the public to which they owed accountability. This explains the provision contained in most constitutions obliging the legislature to publish a record of its debates.

**Constitutionality.** Written constitutions normally provide the standard by which the legitimacy of governmental actions is judged. In the United States, the practice of the judicial review of congressional legislation for its constitutionality—that is, for its conformity with the U.S. Constitution—though not explicitly provided for by the Constitution, developed in the early years of the republic. More recently, other written constitutions, including the Basic Law of the Federal Republic of Germany and Italy's republican constitution, provided explicitly for judicial review of the constitutionality of parliamentary legislation. This does not necessarily mean that a constitution is regarded as being prior and superior to all law. Although several European countries, including France, Italy, and West Germany, adopted new constitutions after World War II, they kept in force their codes of civil law, which had been legislated in the 19th century; and the U.S. Constitution guarantees citizens certain substantive and procedural rights to which they deemed themselves entitled as subjects of the British crown under the ancient English common law. Despite the greater antiquity of law codes, however, portions of them have been revised from time to time in order to eliminate conflicts between the law and certain constitutional norms that are regarded as superior. Parts of German family law and of the criminal code, for example, were revised in order to bring them into conformity with the constitutional provisions regarding the equality of persons irrespective of sex and with the individual's constitutionally guaranteed right to the free development of his personality.

Constitutional interpretation

Conflicting interests or parties are, of course, likely to place different interpretations on particular provisions of a constitution, and means, therefore, have to be provided for the resolution of such conflicts. The constitution itself may establish an institution, the task of which is to interpret and clarify the terms of that constitution. In the American system, the Supreme Court is generally regarded as the authoritative interpreter of the Constitution. But the Supreme Court cannot be regarded as the "final" interpreter of the meaning of the Constitution for a number of reasons. The court can always reverse itself, as it has done. The president can gradually change the interpretative outlook of the court through the nomination of new justices, and the Congress can exert a more negative influence by refusing to confirm presidential nominations of justices.

Provision was made in the constitution of the Fifth French Republic for the interpretation of certain constitutional matters by a Constitutional Council. Soon after the French electorate, in a referendum in 1958, had voted to accept the Constitution, a controversy erupted in France over the question of whether the president of the republic could submit to popular referendum issues not involving constitutional amendments but on which parliament had taken a position at odds with the president's. The Constitution itself seemed to provide that the Constitutional Council could rule definitively on this question, but Pres. Charles de Gaulle chose to ignore its ruling, which was unfavourable to himself. As a result, the Constitutional Council lost authority as the final interpreter of the meaning of the Constitution of the Fifth Republic.

It may thus be seen that because of the inherent difficulties in assessing the intentions of the authors of a constitution and because of the possibility that the executive or legislative branch of government may be able to ignore, override, or influence its findings, it is difficult to ensure constitutional government merely by setting up an institution whose purpose is constitutional interpretation.

**Constitutional change.** Written constitutions are not only likely to give rise to greater problems of interpretation than unwritten ones, but they are also harder to change. Unwritten constitutions tend to change gradually, continually, and often imperceptibly, in response to changing needs. But when a constitution lays down exact procedures for the election of the president, for relations between the executive and legislative branches, or for defining whether a particular governmental function is to be performed by the federal government or a member state, then the only constitutional way to change these procedures is by means of the procedure provided by the constitution itself for its own amendment. Any attempt to effect change by means of judicial review or interpretation is unconstitutional, unless, of course, the constitution provides that a body (such as the U.S. Supreme Court) may change, rather than interpret, the constitution.

Many constitutional documents make no clear distinction between that which is to be regarded as constitutional, fundamental, and organic, on the one hand, and that which is merely legislative, circumstantial, and more or less transitory, on the other. The constitution of the German Weimar Republic could be amended by as little as four-ninths of the membership of the Reichstag, without any requirement for subsequent ratification by the states, by constitutional conventions, or by referendum. Although Hitler never explicitly abrogated the Weimar Constitution, he was able to replace the procedural and institutional stability that it had sought to establish with a condition of almost total procedural and institutional flux.

A similar situation prevailed in the Soviet Union under the rule of Stalin. But Stalin took great trouble and some pride in having a constitution bearing his name adopted in 1936. The Stalin constitution continued, together with the Rules of the Communist Party of the Soviet Union, to serve as the formal framework of government until the ratification of a new, though rather similar, constitution in 1977. The procedures established by these documents, however, have not been able to provide Soviet citizens and politicians with reliable knowledge of the rules of the political process from one year to the next or with guidance as to which institutions and practices they are to consider fundamental or virtually sacrosanct and which they can safely criticize. As a result, changes in the personnel and policies of the Soviet Union and of similar Communist regimes have rarely been brought about smoothly and have frequently required the use of violence.

**Constitutional stability.** If one distinguishes between stability and stagnation on the one hand and between flexibility and flux on the other, then one can consider

Characteristics of stable constitutions

those constitutional systems most successful that combine procedural stability with substantive flexibility—that is, that preserve the same general rules of political procedure from one generation to the next while at the same time facilitating adaptation to changing circumstances. By reference to such criteria, those written constitutions have achieved the greatest success that are comparatively short; that confine themselves in the main to matters of procedure (including their own amendment) rather than matters of substance; that, to the extent that they contain substantive provisions at all, keep these rather vague and generalized; and that contain procedures that are congruent with popular political experience and know-how. These general characteristics appear to be more important in making for stability than such particular arrangements as the relations between various organs and levels of government or the powers, functions, and terms of tenure of different officers of state.

There is little evidence to support the thesis that a high level of citizen participation necessarily contributes to the stability of constitutional government. On the contrary, the English political economist Walter Bagehot, who in 1867 wrote a classic analysis of the English constitution (*The English Constitution*), stressed the "deferential" character of the English people, who were quite happy to leave government in the hands of the governing class.

Much more important than formal citizen behaviour, such as electoral participation, are informal attitudes and practices and the extent to which they are congruent with the formal prescriptions and proscriptions of the constitution itself. Constitutional government cannot survive effectively in situations in which the constitution prescribes a pattern of behaviour or of conducting affairs that is alien to the customs and way of thinking of the people. When, as happened in many developing countries in the decades after World War II, a new and alien kind of constitutional democracy is imposed or adopted, a gap may soon develop between constitutionally prescribed and actual governmental practice. This in turn renders the government susceptible to attack by opposition groups. Such attack is especially easy to mount in situations in which a constitution has a heavy and detailed substantive content, when, for example, it guarantees the right to gainful employment or the right to a university education for all qualified candidates. In the event of the government being unable to fulfill its commitment, the opposition is able to call the constitution a mere scrap of paper and to demand its improvement or even its complete replacement. Such tactics often have succeeded, but they ignore the dual strategic function of the constitution. It is meant not only to arrange the offices of the state, in Aristotle's sense, but also to state the goals toward which the authors and ratifiers of the constitution want their community to move.

## THE PRACTICE OF CONSTITUTIONAL GOVERNMENT

**Great Britain.** It is accepted constitutional theory that Parliament (the House of Commons and the House of Lords acting with the assent of the monarch) can do anything it wants to, including abolish itself. The interesting aspect of British government is that, despite the absence of restraints such as judicial review, acts that would be considered unconstitutional in the presence of a written constitution are attempted very rarely, certainly less often than in the United States.

The English constitution and the English common law grew up together, very gradually, more as the result of the accretion of custom than through deliberate, rational legislation by some "sovereign" lawgiver. Parliament grew out of the Curia Regis, the King's Council, in which the monarch originally consulted with the great magnates of the realm and later with commoners who represented the boroughs and the shires. Parliament was, and is, a place in which to debate specific issues of disagreement between, initially, the crown, on the one hand, and the Lords and Commons, on the other. The conflicts were settled in Parliament so that its original main function was that of a court—it was in fact known as "the High Court of Parliament" as late as the 16th century.

The locus of power in the English constitution shifted gradually as a result of changes in the groups whose consent the government required in order to be effective. In feudal times, the consent of the great landowning noblemen was needed. Later, the cooperation of commoners willing to grant revenue to the crown—that is, to pay taxes—was sought. The crown itself, meanwhile, was increasingly institutionalized, and the distinction was drawn ever more clearly between the private and public capacities of the king. During the course of the 18th century, effective government passed more and more into the hands of the king's first minister and his cabinet, all of them members of one of the two houses of Parliament. Before this development, the king's ministers depended upon their royal master's confidence to continue in office. Henceforward they depended upon the confidence of the House of Lords and especially the House of Commons, which had to vote the money without which the king's government could not be carried on. In this way the parlay that was originally between the monarch and the houses of Parliament was now struck between the ministry and its supporters, on the one hand, and opposing members of Parliament, on the other. Parliamentary factions were slowly consolidated into parliamentary parties, and these parties reached out for support into the population at large by means of the franchise, which was repeatedly enlarged in the course of the 19th century and eventually extended to women and then to 18-year-olds in the 20th.

Parliamentary opposition

When a prime minister loses a vote of confidence in the House of Commons, he can either resign to let the leader of the Opposition form a new government or ask the monarch to dissolve Parliament and call for new elections. As a result of the strong party discipline that developed in the 20th century, prime ministers generally do not lose votes of confidence any more, and they call for new elections at the politically most favourable moment. According to an act of Parliament, elections must be held at least every five years—but another act of Parliament can change or suspend this apparently "constitutional" provision, as was done during World War II, when the life of the incumbent House of Commons was extended until the defeat of Germany. Similarly, relations between, and the relative powers of, the House of Lords and the House of Commons have been repeatedly redefined to the disadvantage of the House of Lords by acts of Parliament, to such an extent that the Lords retain only a weak suspensory veto. All such fundamental constitutional changes have occurred either informally and without any kind of legislation at all or as a result of the same legislative procedures employed to pass any other ordinary circumstantial bill.

**United States.** The U.S. Constitution is not only replete with phrases taken from the British constitutional vocabulary, but in several respects, it also represents a codification of its authors' understanding of the English constitution, to which they added ingenious federalist inventions and the formal amending procedure itself. Despite the availability of this procedure, however, many if not most of the fundamental changes in American constitutional practice have not been effected by formal amendments. The Constitution still does not mention political parties or the president's cabinet. Nor was the Constitution changed in order to bring about or to sanction the fundamentally altered relations between the executive and the Congress, between the Senate and the House, and between the judiciary, the legislature, and the executive.

The presence of a constitutional document, however, has made American politics more consciously "constitutionalist," at least in the sense that politicians in the United States take more frequent recourse than their British counterparts to legalistic argumentation and to actual constitutional litigation. The United States, moreover, is denied the kind of flexibility illustrated by the postponement of British parliamentary elections during World War II since the Constitution explicitly provides the dates for congressional and presidential elections. It is one of the remarkable facts of American constitutional history that the constitutional timetable for elections has always been observed, even during external war and the Civil War of the 19th century.

Constitutional Government in France

**Europe.** France, Germany, and Italy, as well as most non-European countries influenced by continental concepts of constitutionalism, have no record of unbroken constitutional fidelity similar to that found in Britain and the U.S. Because of the highly substantive and ideological content of most French constitutions, the best way to change them has been to replace them altogether with a new, ideologically different document. Only the constitution of the Third Republic (established in 1870) was exceptional in this respect, since it consisted of very short, highly procedural organic laws, which served France well for 70 years, until the German invasion of 1940.

The main political problem attributed to the constitution of the Third Republic was the instability of cabinets. The negative majorities that voted "no confidence" in a cabinet usually could not stay together for the positive purpose of confirming a new cabinet. The constitution of the Fourth Republic (1946–58) made the overthrow of governments by the National Assembly more difficult. In fact, however, the life of the average cabinet in the Fourth Republic was even shorter than in the Third, and French government became virtually paralyzed when it had to deal with the problems raised by the Algerian independence movement. To avert a military takeover, General de Gaulle was given wide discretion in 1958 in the formulation of a new constitution, which was overwhelmingly accepted in a referendum. The constitution of the Fifth French Republic gives the president of the Republic the power to dissolve Parliament and the means of circumventing a hostile National Assembly through the referendum. Since 1958, French cabinets have been very stable indeed, and the constitution proved resilient during the "revolution of 1968."

Germany, which was unified as a national state only in 1871, established its first democratic constitution in 1919, after its defeat in World War I. Although some of the greatest German jurists and social scientists of the time participated in writing the Weimar Constitution, it has been adjudged a failure. Political parties became highly fragmented, a phenomenon that was explained partly by an extremely democratic electoral law (not a part of the constitution) providing for proportional representation. Some of the parties of the right, such as Hitler's Nazis, and of the left, such as the Communists, were opposed to the constitutional order and used violence in their efforts to overthrow the Republic. To deal with these threats, the President used his constitutional emergency powers under which he could suspend civil rights in member states of the federal system. Several chancellors (the German equivalent of a prime minister) stayed in office after the President had dissolved a Parliament in which the chancellor lacked a supporting majority. They continued to govern with the help of presidential emergency powers and by legislating on the basis of powers previously delegated to them by Parliament.

When a new constitution was drafted for the Western zones of occupation after World War II, every effort was made to correct those constitutional errors to which the failure of the Weimar Republic was attributed. Under the Basic Law of Bonn, Parliament cannot delegate its legislative function to the chancellor, and civil rights cannot be suspended without continuous parliamentary surveillance. The president has been turned into a figurehead on the model of the French presidents of the Third and Fourth Republics, and Parliament cannot overthrow a chancellor and his cabinet unless it first elects a successor with the vote of a majority of its members. Negative majorities cannot paralyze government unless they can agree on alternative policies and personnel. The extreme form of proportional representation used before Hitler came to power was replaced by a mixed electoral system under which half the members of the Bundestag (the lower house of the legislature) are elected from party lists by proportional representation, while the other half are elected in single member constituencies. In order to benefit from proportional representation, a party must obtain at least 5 percent of the votes cast. As a result, the number of parties steadily contracted during the first two decades of the Federal Republic and extremist parties were kept out of Parliament. Cabinets have been very stable, and the provision for the "constructive vote of no confidence" was invoked for the first time only in 1982.

**Latin America, Africa, and Asia.** The experience of constitutional government in continental Europe exerted great influence on the newly independent former colonies of Europe in the Middle East, Asia, and Africa. In the early years of their independence from Spain, most Latin-American countries adopted constitutions similar to that of the United States. But since they lacked the background that produced the American Constitution, including English common law, most of their efforts at constitutional engineering were unsuccessful.

In Asia and Africa and in the Caribbean, many former colonies of Great Britain, such as India, Nigeria, Zambia, and Jamaica, have been comparatively more successful in the operation of constitutional government than former colonies of the continental European countries (*e.g.*, Indonesia, Congo, and Haiti). The British usually left a modified and simplified version of their own constitution upon granting independence to their former subjects, some of whom they had previously trained in the complicated operating procedures of the British constitution. British parliamentary procedure proved sufficiently adaptable to remain in use for some time after the departure of the British themselves. France's former colonies in Africa, because they achieved independence after the founding of the Fifth Republic, modelled their new constitutions upon General de Gaulle's, partly because this enhanced the power of the leaders under whom independence had been achieved.

**BIBLIOGRAPHY.** Current texts of more than 150 national constitutions are available in English translation in *Constitutions of the Countries of the World,* ed. by ALBERT P. BLAUSTEIN and G.H. FLANZ, 14 vol., plus supplements issued yearly (1971– ). Another compendium of constitutions is AMOS J. PEASLEE (ed.), *Constitutions of Nations,* 3rd ed., 4 vol. in 7 (1965–70; 4th ed., vol. 1, 1974).

A comprehensive contemporary book on this subject is CARL J. FRIEDRICH, *Constitutional Government and Democracy: Theory and Practice in Europe and America,* 4th ed. (1968). See also HERBERT J. SPIRO, *Government by Constitution: The Political Systems of Democracy* (1959). The best study of constitutional change is WILLIAM S. LIVINGSTON, *Federalism and Constitutional Change* (1956, reprinted 1974). Aristotle's classic *Politics* is available in translation by SIR ERNEST BARKER (1962). Sir Ernest Barker has also edited *Social Contract: Essays by Locke, Hume, Rousseau* (1947, reprinted 1980). The best history of the origins of English constitutionalism is CHARLES HOWARD MCILWAIN, *The High Court of Parliament and Its Supremacy: An Historical Essay on the Boundaries Between Legislation and Adjudication in England* (1910, reprinted 1979). See also FRANCIS DUNHAM WORMUTH, *The Origins of Modern Constitutionalism* (1949). On the related problems of accountability and responsibility, see HERBERT J. SPIRO, *Responsibility in Government: Theory and Practice* (1969). For an understanding of the origins of American constitutionalism, *The Federalist* by ALEXANDER HAMILTON, JAMES MADISON, and JOHN JAY is indispensable. The classic exposition of *The English Constitution* was provided by WALTER BAGEHOT (1867, reprinted 1978); for a more recent analysis, see SAMUEL H. BEER and ADAM B. ULAM (eds.), *Patterns of Government: The Major Political Systems of Europe,* 3rd ed. (1973). On French constitutionalism, see also STANLEY HOFFMANN *et al., In Search of France* (1963), and *Decline or Renewal? France Since the 1930s* (1974). RALF DAHRENDORF, *Society and Democracy in Germany* (1967, reissued 1979; originally published in German, 1965), is a sociological account, complemented by A.J. HEIDENHEIMER and DONALD P. KOMMERS, *The Governments of Germany,* 4th ed. (1975). For excerpts of constitutional documents of the major European governments, see JOHN J. WUEST and MANFRED C. VERNON, *New Source Book in Major European Governments* (1966). The concept of congruence between political and social patterns of authority is elaborated in a work by HARRY H. ECKSTEIN, *Division and Cohesion in Democracy: A Study of Norway* (1966).

Studies of constitutional development include: HERBERT J. SPIRO (ed.), *Patterns of African Development: Five Comparisons* (1967); B.O. NWABUEZE, *Constitutionism in the Emergent States* (1973); WILLIAM B. SIMONS (ed.), *The Constitutions of the Communist World* (1980); and LAWRENCE WARD BEER (ed.), *Constitutionalism in Asia: Asian Views of the American Influence* (1979).

(H.J.Sp./Ed.)

# Constitutional Law

Constitutional law is the body of rules, doctrines, and practices that govern the operation of political communities. In modern times by far the most important political community has been the national state. Modern constitutional law is the offspring of nationalism as well as of the idea that the state must protect certain fundamental rights of the individual. As national states have multiplied in number, so have constitutions and with them the body of constitutional law. But constitutional law originates today sometimes from non-national sources too, while the protection of individual rights has become the concern also of supranational institutions.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 552.

This article is divided into the following sections:

Constitutions and constitutional law 695
The nature of constitutional law
Characteristics of constitutions
Unitary, federal, and regionalist systems 697
The distinction between unitary, federal, and regionalist states
Classifying states as federal, regionalist, and unitary
International unions of states
Executives and legislatures 698
The constitution and the executive
Unicameral and bicameral legislatures
Judicial review 700
Judicial review in the United States
Judicial review in Europe and elsewhere
Trends of judicial review in the United States and in Europe
Transnational judicial review
Bibliography 703

## Constitutions and constitutional law

### THE NATURE OF CONSTITUTIONAL LAW

In the broadest sense a constitution is a body of rules governing the affairs of an organized group. A parliament, a church congregation, a social club, or a trade union may operate under the terms of a formal written document labeled constitution. This does not mean that all of the rules of the organization are in the constitution, for usually there are many other rules such as bylaws and customs. Invariably, by definition, the rules spelled out in the constitution are considered to be basic, in the sense that, until they are modified according to an appropriate procedure, all other rules must conform with them. Thus the presiding officer of a club is obliged to rule that a proposal is out of order if it is contrary to a provision of its constitution. Implicit in the concept of a constitution is that of a higher law that takes precedence.

Every political community, and thus every national state, has a constitution, at least in the sense that it operates its important institutions according to some fundamental body of rules. In this sense of the term the only conceivable alternative to a constitution is a condition of anarchy. Constitutions may be written or unwritten; they may be complex or simple; they may provide for vastly different patterns of governance. Even if the only rule that matters is the whim of an absolute dictator, that may be said to be the constitution. (G.B./D.Fe.)

The constitution of a political community is therefore composed, in the first place, of the principles determining the agencies to which the task of governing the community is entrusted and their respective powers. In absolute monarchies, such as the Oriental kingdoms and the Roman Empire in antiquity and the French monarchy between the 16th and 18th centuries, all sovereign powers were concentrated in one person, the king or emperor, who exercised them directly or through subordinate agencies that had to act according to his instructions. In ancient republics, such as Athens and Rome, the constitution provided, as do the constitutions of most modern states, for a distribution of powers among distinct agencies. But whether it concentrates or distributes these powers, a constitution always contains at least the rules that define the structures and operations of the government that runs the community.

Necessary and contingent elements of constitutions

The constitution of a political community may contain more, however, than the definition of the authorities endowed with powers to command. It may also include principles that delimit those powers in order to secure against them fundamental rights of persons or groups. The idea that political sovereignty is not unlimited stems from an old tradition in Western philosophy. Well before the advent of Christianity, Greek philosophers thought that positive law—*i.e.,* the law actually enforced in a community—in order to be just must reflect the principles of a superior, ideal law: natural law. Similar conceptions were propagated in Rome by Cicero and by the Stoics. Later the Church Fathers and the Scholastics held that positive law was binding only if it did not conflict with the precepts of divine law. These considerations did not remain abstract speculations of philosophers and theologians; to a measure, they found reception in fundamental rules of positive legal systems. In Europe, for example, the authority of political rulers throughout the Middle Ages did not extend to religious matters, which were strictly reserved to the jurisdiction of the church. The powers of political rulers, moreover, were limited by the rights of at least some classes of subjects. Quarrels and fights over the extent of such rights were not infrequent; and they were sometimes settled through solemn, legal "pacts" among the contenders, the prominent example being Magna Carta (1215). In the modern age, even the powers of an absolute monarch such as the king of France were not truly absolute: acting alone, he could not alter the fundamental laws of the kingdom or disestablish the Roman Catholic Church.

Against this background of already existing legal limitations on the powers of governments, a decisive turn in the history of Western constitutional law occurred when a theory of natural law based on the "inalienable rights" of the individual was developed. John Locke (1632–1704) was the first outstanding champion of the theory. He was followed by others, and in the 18th century the doctrine of the rights of the individual became the banner of the Enlightenment. The theory assumed that there are certain rights belonging to every single human being (religious freedom, freedom of speech, freedom to acquire and possess property, freedom not to be punished on the basis of retroactive laws and of unfair criminal procedures, and so on), which governments cannot "take away" because they were not "created" by governments. The theory further assumed that governments must be organized in such a way as to afford an effective protection of the rights of the individual. For that purpose it was thought that, as a minimal prerequisite, governmental functions must be divided into legislative, executive, and judicial; that executive action must comply with the rules laid down by the legislature; and that remedies, administered by an independent judiciary, must be available against illegal executive action.

The theory of the rights of the individual was a potent factor in reshaping the constitutions of Western states in the 17th, 18th, and 19th centuries. The first step was made by England at the time of the Glorious Revolution (1688). All of these principles concerning the distinction of governmental functions and their appropriate relations were incorporated in constitutional law. England also soon

Modern constitutional law

changed some of its laws so as to give more adequate legal force to the newly discovered individual freedoms. It was in the United States, however, that the theory scored its most complete success. Once the English colonies became independent states (1776), they faced the problem of giving themselves a fresh political organization. They seized the opportunity to spell out in special legal documents, which could be amended only through a special procedure, all the main principles providing for the distribution of governmental functions among distinct state agencies as required by the theory, as well as the main principles concerning the rights of the individual the theory wanted to be respected by all state powers. The federal Constitution (1788) and its Bill of Rights (Amendments I–X, 1791) did the same, shortly thereafter, at the national level. By giving through this device a formal, higher status to rules defining the essential organization of government, as well as the essential limitations of its legislative and executive powers, U.S. constitutionalism put in full evidence the character that belongs, in essence, to all constitutional law: the fact of its being "basic" with respect to all other laws of the legal system. This also made it possible to set up institutional controls over the conformity even of legislation with the group of rules considered, within the system, to be of supreme importance.

The American idea of stating in an orderly, comprehensive document the essentials of the rules that must guide the operations of government became popular very quickly. Since the end of the 18th century scores of states, in Europe and elsewhere, have followed the United States' example. Today, almost all states have constitutional documents describing the fundamental organs of the state, the ways they should operate, and, usually, the rights they must respect and even sometimes the goals they ought to pursue. To be sure, not all of these documents are inspired by the individualistic ideas that permeate modern Western constitutional law. The constitutions of Communist countries, for instance, organize state powers and their relations in ways that differ from Western thought and subordinate individual freedoms to the goal of achieving a classless society. Yet, notwithstanding great differences among themselves, the constitutional charters of contemporary states are all similar at least in one respect: they are meant to express the core of the constitutional law governing their respective countries. (G.B.)

## CHARACTERISTICS OF CONSTITUTIONS

Every state has a constitution, since every state functions on the basis of certain rules and principles. It has often been asserted that the United States has a written constitution but that the constitution of Great Britain is unwritten. This is true, but only in the sense that in the United States there is a formal document called the Constitution, whereas there is no such document in Great Britain. In fact, however, many parts of the British constitution exist in written form, whereas important aspects of the American constitution are wholly unwritten. The British constitution includes the Bill of Rights (1689), the Act of Settlement (1700–01), the Parliament Act of 1911, the successive Representation of the People acts (which extended the suffrage), the statutes dealing with the structure of the courts, the various local government acts, and many others. These are not ordinary statutes, even though they were adopted in the ordinary legislative way, and they are not codified within the structure of a single orderly document. On the other hand, such institutions in the United States as the presidential cabinet and the system of political parties, though not even mentioned in the written constitution, are most certainly of constitutional significance. (D.Fe.)

Written and unwritten constitutions

Written constitutions, indeed, can never exhaust the whole constitutional law of a state. They are always supplemented, to varying degrees, by statutes, judicial doctrines interpreting the constitution, intergovernmental practices, and nongovernmental institutions (such as political parties) and their practices. Without these supplementary elements the overall constitutional framework of the political community would not be what it is.

Whether "long" or "short," written constitutions can concern themselves exclusively or prevalently with the "organization" of government or deal extensively also with the rights of the people and with the goals of governmental action. The U.S. Constitution is a model of brevity (*c.* 7,000 words). Just a little longer are most of the Western countries' constitutions. On the other hand, the constitutions of India and Yugoslavia extend to hundreds of pages. Merely "organizational" constitutions (*i.e.*, documents containing no guarantees for rights) have become very rare.

Written constitutions are said to be "normative" when their binding principles are more or less all observed in the actual operations of the political system. This applies to the constitutions of the United States, of Canada, and of the western European countries. Other constitutions are said to be "nominal," because they are largely or in substantial parts disregarded and do not provide insight into the real functioning of the system. This is often the case with constitutions of rapidly developing countries and of countries ruled by a one-man or a one-party dictatorship.

Constitutions, written or unwritten, must be distinguished according to whether they are "rigid" or "flexible." Rigid are those constitutions at least some part of which cannot be modified in the ordinary legislative way. Flexible are those whose rules can all be modified through the simple procedure by which statutes are enacted. The United States has a rigid constitution, because proposals to amend the constitutional document adopted in 1788 must have a two-thirds majority vote in each house of Congress or be made by a convention called by two-thirds of the states, with subsequent ratification, in either case, by the legislatures or specially elected conventions of three-fourths of the states. Great Britain has a flexible constitution because all of its constitutional institutions and rules can be abrogated or modified by an act of Parliament.

Rigid and flexible constitutions

As almost all states have by now adopted written constitutions and these almost invariably provide for a special, exacting procedure for their amendment, the great majority of today's constitutions are rigid.

The distinction between rigid and flexible constitutions is important because only under rigid constitutions is it possible to establish institutional controls over the conformity of legislation with the principles considered indispensable for the well-being of the community. The importance of the distinction in relation to the issue of the stability and continuity of a country's constitutional law may, however, be easily exaggerated. In the United States it is difficult to activate the amending process. But in other countries with rigid constitutions, amending them is easier. In Switzerland the federal constitution of 1874 can be amended, on the initiative of 100,000 citizens or of the legislature, by a majority vote in a national referendum, with a favourable decision in the majority of cantons. Thus the provisions of the Swiss constitution have been changed repeatedly on many important points. In addition, even if the provisions of a rigid constitution remain unaltered, in the long run they often assume different meaning and scope. The "commerce clause" and the "due process clause" of the U.S. Constitution do not have the same legal implications today as they did 100 years ago. This is because formal constitutional provisions must be interpreted by the courts or by the legislature, the executive, and other institutional subjects. To a certain extent, interpretation inevitably involves adaptation of the letter of the law to the evolving needs and expectations of the community.

On the other hand, the constitutional law of countries with flexible constitutions must not necessarily be unstable and rapidly changing. Great Britain can modify all of its constitutional law by statute (or even in important areas by "conventions" between the supreme institutional powers of the state: the Crown, Parliament, the Cabinet). Nevertheless, statutes and common law principles of constitutional import cannot be changed as easily as other statutes and rules; they have a considerable capacity for permanence. For instance, the "rule of law"—roughly the equivalent of the American "due process" principle—has been an indefectible part of the British constitution approximately since the Glorious Revolution.

In sum, the relative continuity of a country's constitu-

tional law does not depend on the adoption of a rigid constitution, although such a constitution may make changes at times more complicated and difficult. It depends rather on the people's feelings about the fundamental political values the legal system ought to honour. If and when these feelings change, the change will be able to make its way sooner or later into the constitution, whether by using the amending process and the method of interpretation under a rigid constitution, or the easier ordinary legislative procedure or other methods under a flexible constitution (apart, of course, from the further possibility of a change, in both cases, through the extreme means of a violent revolution). Because the political values felt to be supreme by the dominant forces in a community have ultimate controlling influence, some European continental scholars have been prompted to call them the "material constitution," at any given historical moment, of that community. The developments of the material constitution are decisive in determining the retention or the demise, as well as the actual meaning and scope in application, of the principles and rules of the constitution, whether the latter is rigid or flexible.

## Unitary, federal, and regionalist systems

### THE DISTINCTION BETWEEN UNITARY, FEDERAL, AND REGIONALIST STATES

No modern state can govern a country only from a central point. The affairs of municipalities and rural areas must be left to the administration of local governments. Accordingly, in all modern states there are at least two levels of government: the central government and the local governments. But in a number of states between the two levels there exists still a third one consisting of governments that take care of the interests of, and rule over, more or less large regions.

Levels of government

The distribution of powers among different levels of government is an important aspect of the constitutional organization of a state. States with two levels of government can be distinguished on account of the greater or lesser autonomy they grant to the local level. Great Britain's respect for local self-government has always been a characteristic of its constitution. France instead, at least until recently, used to keep under strict central control its local authorities. In states with three levels of government the distribution of powers among the central and the intermediate governments varies. States formed through the union of formerly independent states usually maintain considerable legislative, executive, and judicial power at the intermediate level; the United States and Switzerland fall into this category. However, other states with three levels of government grant few powers to the intermediate level. This often happens in states that have introduced this level as a correction of their previous choice of two levels; this was done, for example, by Italy in its 1948 constitution.

States with two levels of government are called unitary, with three levels of the first category federal and with three levels of the second kind decentralized or "regionalist." These definitions, however, cannot be properly understood unless other elements characterizing the three types are mentioned and a tendency of the types to overlap is kept in mind.

The federal state

The model federal state requires the existence, at the national level, of a written, rigid constitution guaranteeing not only the permanence and independence of the several intermediate governments but also the amplitude of their legislative, executive, and judicial powers. The national constitution must delegate to the central government only enumerated powers; the remaining powers are reserved to the intermediate governments. These (be they called states as in the United States or cantons as in Switzerland) must be represented as such, possibly on an equal footing, in a second chamber of the national legislature. They must also be allowed to participate somehow in the amending process of the national constitution. Such constitutional arrangements are at the same time the hallmarks of the genuine federal state and a guarantee against possible efforts of the central government to enlarge its jurisdiction and so imperil the important political role the intermediate governments must play in this kind of state. Formal constitutional safeguards, however, are not enough to preserve that role. Apart from constitutional amendments, the central government may always broaden its own sphere through the use of constitutional clauses granting "implied powers" or, more simply, through a suitable interpretation of the constitution by its own agencies (to which the final interpretation of the constitution is reserved). As a matter of fact, in the 20th century the balance in all federal systems has shifted toward central governments; intermediate governments have lost much of their previous exclusive rights and political weight. The shift was caused by the growth of governmental intervention in the economy and by the development of welfare measures: neither could take place without substantial involvement of the centre. An increase in the powers of the central government does not by itself impair the federal nature of the state if it keeps within certain boundaries. Beyond these boundaries, however, no matter what the formal constitutional principles seem to indicate, the federal state tends to become in fact a regionalist state.

The regionalist state

Regionalist states are also based, as a rule, on written, rigid constitutions granting some limited legislative and administrative (seldom judicial) powers to the intermediate or regional governments. But because regional governments possess jurisdiction only over enumerated matters, and even there are subject in part to the overriding powers of the national authorities, their actual role and political weight within the system largely depend on the will of the central government to buttress or to restrict their autonomy. Where the powers attributed by the constitution to the regional governments are particularly exiguous, the regionalist state will look in many respects like a unitary state. Where the powers are relatively large and the central government favours their expansion, the state tends to assume federal connotations even if the typical hallmarks of the federal system are not there.

In general, regionalism recommends itself in situations where the need is felt for a strong, all-controlling national government but the territorial extension of the country is great or the country is divided into sections with different ethnic, linguistic, or social characteristics (the last is the case of Yugoslavia and Belgium). Regionalism has also offered itself as a solution to unitary states that have decided to introduce regional governments in order to reinforce their democratic features or to alleviate the evils of an overcentralized bureaucracy (the case of Italy).

### CLASSIFYING STATES AS FEDERAL, REGIONALIST, AND UNITARY

Classifying a particular state as federal, regionalist, or unitary may at times be difficult.

United States, Switzerland, Canada, Australia

The United States and Switzerland are clearly federal states, although the role respectively of states and cantons has shrunk much since World War I: all of the above mentioned characteristics of the federal state are present in their constitutional systems. Canada is also a federal state, despite the fact that at least one of the formal marks of perfect federalism is absent from its 1982 constitution: the provinces' powers, not the central government's, are enumerated. But the provinces' powers are vast, and the guarantees for the provinces' independence and rights anchored in the constitution are particularly strong. For the same reasons Australia too can be considered a federal state. The Federal Republic of Germany is federal in all respects. Yet the legislation of the *Bund,* the central government, extends over so many matters that it is questionable whether West Germany is not in fact a regionalist state. The question is even more open as regards India. The Indian federal constitution spells out a long list of important subjects over which the states and territories that compose the union have exclusive jurisdiction. But the constitution gives the central government the power to legislate on any subject—including the ones reserved to the regional governments—if it deems the matter of national importance. In addition, the central government has direct powers of control over the regional governments: *e.g.,* the national Parliament can dissolve the legislative

West Germany and India

council of any state or territory. Similar remarks can be made with respect to the federal structure of many Latin-American states. The Soviet Union is, by constitution, a federal state; but, apart from the question of the "nominal" value of at least certain parts of its constitution, the constitutional role entrusted to the Communist party has such unifying effects on the whole system that the Soviet state can probably be defined at best as a regionalist state.

Italy and Spain

Italy and Spain are states with regional governments. These, by constitution, are endowed with legislative and administrative powers in certain areas (the courts are all national). Italy, however, is perhaps the best example of how a regionalist state may closely resemble a unitary one. The constitution grants limited powers to the regions. Parliament has extended these limits by devolving additional matters to the purview of regional legislatures. But regional laws must respect general principles laid down in national statutes, and in practice little room is left for really autonomous regional legislation. The regions, moreover, are not financially independent. Thus, all in all, they might almost be considered a branch of the system of local governments, together with communes and provinces, rather than a distinct, third level of government.

Great Britain and France

Great Britain and France are unitary states. Northern Ireland had special autonomy within the United Kingdom until restrictions were introduced to cope with the emergency situation in that region, and Scotland and Wales would have had special autonomy but in referendums their people rejected the devolution plans offered by the British Parliament. Had the plans been accepted, Great Britain would have become in part a regionalist state, though Parliament, under a flexible constitution, might, in theory, have later repealed the grant of autonomy. France in 1982 established by statute elective regional governments (*régions*) as part of a more general reform intended to make the system of local governments less dependent on the centre. *Régions* have fewer powers than the Italian regions, but their role may grow with time.

INTERNATIONAL UNIONS OF STATES

Finally, mention must be made of a growing tendency among national states to allow the direct operation within their constitutional systems of the laws of the international community and of the laws of special international organizations to which they belong.

The constitutions of West Germany (article 25) and of Italy (article 10) provide that the legal system must conform with international customary law. Because both are rigid constitutions, this means that ordinary national statutes conflicting with that law are unconstitutional.

International unions and a new level of government

At times unions are formed among national states that, without unifying the member states into a new political community in the strict sense of the word, nonetheless set up governmental agencies whose laws immediately become part of the national systems. It is as if in the national constitutional frameworks a new level of government were added, from above, to the ones already existing. The most important example is the European Economic Community (EEC). Under the Treaty of Rome (1957) the EEC has its own government consisting of a commission, a council of ministers, a parliament, and a court. It can issue regulations on economic matters indicated by the treaty. Up to now, EEC regulations have been issued only if in fact consented to by each of the executives of the 12 member states. Once in force, however, they must be applied by the national courts with precedence over national legislation. The binding interpretation of the treaty and of EEC regulations belongs to the EEC court, to which individuals can have recourse.

The EEC may be the embryo of a future federal state, if the union develops into an organization whose central government is capable of making decisions independently of the will of member states (which for the moment it is not) and if it assumes functions in the field of foreign and military policy (which at present it does not possess). The community may also never become a federal state. But even as it exists now, it is much more than a simple international alliance of national states that have in common economic interests and assume the obligation to respect the regulations of such interests made by some external agency. The structures of the EEC penetrate deeply into the constitutional structures of the national member states, much in the same way as the structures of the central government do with respect to the member states in a federal system. Some features of federalism are already present in the relations between the EEC and the states belonging to it. Just as in a federal system, community law prevails immediately in the areas assigned to it over national law; it is the "supreme law of the land," which the national judiciaries must prefer to the laws of their own states in case of conflict. Just as in a federal system, the interpretation of EEC law is reserved to a central government agency (the community court), whose decisions are directly binding for the member states and their courts. A member state, however, may always withdraw from the union if it chooses to do so, unlike a state member of a true federal system. But until it takes such a step, its subjection to the law of the EEC is practically the same as that of a member state to federal law in a federal system. (G.B.)

## Executives and legislatures

THE CONSTITUTION AND THE EXECUTIVE

States may be classified as monarchical or republican. From another point of view they may be described as having presidential or parliamentary executives (see below).

Though the institution of monarchy is as old as recorded history, the modern age has been moving steadily in the direction of republican government. Depending on how they are counted, there are today less than 30 monarchies. Many monarchs, as in Great Britain, Japan, the Scandinavian countries, and the Low Countries, are best described as constitutional monarchs. This means that they are mainly titular heads of state and do not in fact possess important powers of government. Most of the executive powers are in the hands of ministers, headed by a prime minister, who are politically responsible to the parliament and not to the monarch. The executive powers of government in Great Britain, for example, are exercised by ministers who hold their offices by virtue of the fact that they command the support of a majority in the popularly elected House of Commons. The monarch can act only on the advice of the ministers and cannot exercise an independent will. The position of the monarchs in Scandinavia and the Low Countries is more or less the same as that of the British: they reign but they do not rule. In countries where no political party has a majority of its own in the parliament, the monarch may exercise some discretion in deciding whom to invite to serve as prime minister and to form a government. Even so, since the monarch must first consult with the various party leaders, he is not likely to have much discretion. In a country with a stable two-party system, all the monarch can do is offer the prime ministership to the leader of the majority party. Since 1975 the Swedish king has not even possessed this formal power; it is the president of the legislative assembly who chooses and appoints the prime minister. A constitutional monarch is the head of the state, not of the government. Standing above party and the active political controversies of the day, the sovereign is a focus of national loyalty and a useful symbol of the nation's unity and its historical past.

In a few monarchies, however—for example, those of Jordan and Saudi Arabia—the king exercises real powers of government. The ministers are chosen by and are responsible only to the king rather than to some elective parliamentary body. Hereditary rulers with this degree of personal power were quite common in the 18th century, but they are rare today.

Presidential and parliamentary systems

Far more significant than the distinction between monarchy and republicanism is the contrast between presidential and parliamentary executives. Since the United States has for long been the world's leading exponent of presidential government and Great Britain the oldest and most successful practitioner of parliamentary government, their systems may be taken as models with which the systems of other countries can be compared.

The U.S. system is based upon a strict concept of sep-

aration of powers: the executive, legislative, and judicial powers of government are vested by the Constitution in three separate branches. The president is not selected by Congress, nor is he a member of Congress. He has a fixed term of office of four years, and he holds it no matter how his legislative program fares in Congress and whether or not his political party controls either or both houses of Congress. The members of the Cabinet are chosen by the president and are politically responsible to him. The Constitution does not permit them to be members of Congress; it provides that "no Person holding any Office under the United States, shall be a Member of either House during his Continuance in Office."

The British Parliament

The parliamentary executive system proceeds upon different assumptions. In Great Britain, whose system many countries have chosen to emulate, the executive officers of the state are not entirely separated from the legislative branch. On the contrary, the British Cabinet may be described as the leading committee of Parliament. Although the prime minister, the head of the government, could at one time hold a seat in either the House of Lords or the House of Commons, the contemporary convention is membership in the House of Commons. The other ministers who make up the Cabinet must be members of one or the other house of Parliament. If the prime minister wants someone who is not in Parliament to serve in the Cabinet, he must either appoint him to the peerage or find a vacancy in the House of Commons to which he can be elected.

Whereas in the U.S. system, with its separation of powers, the chief executive and the Cabinet officers are institutionally apart from the legislature, in Great Britain the ministers of the crown hold their powerful executive positions only so long as they enjoy the support of a majority of the House of Commons. A Cabinet that loses that support must either dissolve the House and call a new election, thus in effect putting the issue to the voters, or resign and permit others to form a government. In the 20th century most changes in power have occurred as a result of the outcome of a general election.

It follows that in the British Parliament the prime minister and the Cabinet are fully in charge. They are responsible, as the guiding committee of Parliament, for the preparation and enactment of most legislation and of the budget. There can be no permanent or serious conflict between the House of Commons and the Cabinet, for responsibility means that the government of the day must either prevail or give way to another government. Thus the deadlocks between the chief executive and the Congress that are a frequent occurrence in the United States cannot occur under the parliamentary system.

A system may appear to be parliamentary or presidential without actually being either. In Latin-American and African states many presidential systems have been converted into military dictatorships. The Communist governments of central and eastern Europe are parliamentary in form, but power is effectively in the hands of a party leader. The fact that the head of state is called a president does not mean that the system is presidential. The head of the Italian state, for example, is a president elected for a seven-year term by a joint session of Parliament, but the executive power is in fact exercised by a prime minister and other Cabinet ministers who are responsible to Parliament. In former monarchies that now have parliamentary systems the abolition of monarchy invariably led to the substitution of a president for the hereditary ruler.

Hybrid forms

There are some hybrid forms of government that combine features of both presidential and parliamentary systems. France's Fifth Republic (1958) is a good example. According to the terms of a constitutional amendment adopted in 1962, the president of the republic is elected by direct vote of the people for a seven-year term. This gives the president an enormous moral power derived from the fact that he is a product of universal suffrage. Although in the exercise of some powers the president needs the signature of the prime minister or of some other minister, he has great substantive powers of his own: he appoints the prime minister; he dominates the management of foreign relations; he may dissolve the National Assembly, though not more often than once a year; and he possesses vast emergency powers. The cabinet, called the Council of Ministers, is presided over by the president. Members of the council cannot be members of Parliament, but they have access to both chambers; and they may speak there, though they do not vote. The council is responsible to the National Assembly and can be defeated by censure motion. Thus the French system of government is neither presidential nor parliamentary in form; it combines elements of both in a unique fashion.

The governmental system of the Federal Republic of Germany is mainly parliamentary, but with some interesting variations. The head of state is the president, elected for a five-year term by a body known as the Federal Assembly, which consists of members of the lower house of Parliament (the Bundestag), plus an equal number of representatives elected by the *Länder* (the regional subdivisions of the national state). The head of government is the chancellor, who must have majority support in the Bundestag; but only he, and not the Cabinet, is responsible. The Bundestag can express its lack of confidence in the chancellor only by electing a successor by majority vote. In this fashion it was hoped to eliminate some of the weaknesses of the Weimar constitution of the post-World War I period. The chancellor may seek a vote of confidence, and if he fails he may ask the president to dissolve the Bundestag and call a new election.

The governmental system of India is more typically parliamentary in form. The president is elected for a five-year term by an electoral college consisting of all members of the national Parliament and of the state legislative assemblies. The president can act only on the advice of ministers who are responsible to Parliament. The lower house, the House of the People, is subject to dissolution.

The Swiss executive is unique. The Federal Council consists of seven members elected for four-year terms by the parliament (the Federal Assembly). They are elected as individuals, and they are never forced to resign; in fact, they are almost always reelected, some serving for as long as 25 to 30 years. A disagreement with the Federal Assembly leads neither to resignation of the Federal Council nor to a dissolution of the parliament; the ministers simply adjust their positions to conform with the wishes of the parliamentary majority. This does not mean that the Federal Council is not an important body; as a group, it originates most new legislation, and its members, as individuals, head up the great departments of government. Each year the parliament appoints a member of the Federal Council to serve as president of the confederation. The president is chairman of the Federal Council and titular head of state. Although this system has not been adopted in other countries, it has worked well for the Swiss.

## UNICAMERAL AND BICAMERAL LEGISLATURES

A central feature of any constitution is the legislature. It may be a unicameral body with one chamber or a bicameral body with two chambers.

Unicameral legislatures are to be found in small states with unitary systems of government, among them Denmark, Sweden, Finland, Israel, and New Zealand, or in very tiny states such as Andorra, Luxembourg, and Liechtenstein. They are also found in newly independent states undergoing rapid social and political change, such as some Third World states, and in Socialist countries, where unicameralism is a feature deriving from the Communist concept of the appropriate organization of political power.

Bicameralism and federalism

Federal states, whether large or small, have bicameral legislatures, one house of which represents the main territorial subdivisions. The classic example is the Congress of the United States, which consists of a House of Representatives of 435 members elected for two-year terms from single-member districts of approximately equal population and a Senate consisting of two persons of each state elected by the voters of the state. The fact that all states are represented equally in the Senate, regardless of their size, stems from the federalistic character of the American union. The U.S. Senate enjoys special powers not shared by the House of Representatives: it must authorize by a two-thirds majority vote the ratification of international

treaties concluded by the president and must confirm the appointments of the most important federal officers made by the president.

The federal character of the Swiss constitution is likewise reflected in the makeup of the nation's central legislature, which is bicameral. One house, the National Council, consists of 200 members apportioned among the cantons according to population; the other house, the Council of States, consists of 46 members elected by direct ballot—two from each canton. The Canadian Parliament is also bicameral, just as is the Australian.

Bicameralism is also characteristic of governmental systems that are best described as regionalist. Here, too, bicameralism is expressive of the territorial subdivisions that are joined together to form the national state. The parliament of the Federal Republic of Germany includes a Bundestag elected by general suffrage and a Bundesrat appointed by the 10 *Länder*. Since most members of the Bundesrat are ministers in the *Land* governments, the chamber is a significant link between the *Länder* and the central government. The Parliament of India includes a Council of States and a House of the People, the former elected by the state legislative assemblies and the latter elected directly from territorial constituencies.

Bicameralism and unitary systems

A unitary governmental system does not imply unicameralism in the legislature. Most legislatures of unitary states are, in fact, bicameral, though one chamber is usually more powerful than the other. This is true for the world's oldest and most successful parliament, that of Great Britain, which consists of the House of Lords and the House of Commons. The House of Commons has become by far the more powerful of the two chambers, and the Cabinet is politically responsible only to it. The House of Lords has no control over finances and with respect to other legislation only a modest suspensory veto, which can be easily overcome in the House of Commons by a second vote at an early date. The parliaments of Italy, Japan, and France are also bicameral. In the United States all of the 50 states except Nebraska have bicameral legislatures, even though their governmental systems are unitary. In all states the two houses have equal legislative authority, but the so-called upper houses, usually called senates, have the special function of confirming the governors' appointments. (G.B./D.Fe.)

## Judicial review

From the United States came the first examples of written, rigid constitutions. The United States gave the world another institution that has become a fundamental feature of many contemporary constitutional systems: judicial review.

Under rigid constitutions it is possible to have special state agencies control the conformity of ordinary legislation with the rules of the constitution and, in case of conflict, set the former aside. Under a flexible constitution this cannot happen; in Great Britain, statutes, even if contrary to long-established constitutional principles, are binding for everybody and can be set aside only by subsequent statutes repealing them. The power to invalidate legislation conflicting with the provisions of a rigid constitution can be entrusted either to some particular political organ of the state or to the judiciary. A political organ endowed with such power was, for example, the Senate (Sénat Conservateur) under the French constitutions of 1799 and of 1852. Although some precedents existed in the constitutional history of other countries, it was in the United States that the idea of making the judiciary the guardian of the constitution first took definitive shape. Judicial review—the power of courts to rule on the constitutional validity of legislation—has the advantage over political review in securing more impartial judgments, supported by reasons articulated according to traditional, tested rules of legal interpretation.

### JUDICIAL REVIEW IN THE UNITED STATES

Because U.S. judicial review has been a model for other countries, it is appropriate to devote a few words to it and the body of constitutional law it has produced.

Judicial review is not explicitly mentioned in the U.S. Constitution and is itself a product of judicial construction. The Supreme Court argued in *Marbury* v. *Madison* (1803) that because the Constitution is the supreme law of the land (it says as much in Article VI) and because it is the province of the judiciary to uphold the law, it follows that when state laws and even acts of Congress are inconsistent with a provision of the Constitution, the former must yield to the latter and must be declared void by the courts. The same principle holds, of course, with regard to executive action contrary to the Constitution. Supreme Court pronouncements on questions of constitutionality are final and binding for all other courts and governmental authorities, state and federal.

*Marbury* v. *Madison*

In the U.S. system of judicial review, constitutional questions can be raised only in connection with actual "cases and controversies." Advisory opinions to the government are not rendered by the courts. Although the requirements to litigate cases have been relaxed by the Supreme Court, the rule still is that courts will not decide a constitutional question unless it is rooted in a controversy in which the parties have a direct, personal interest; this can sometimes frustrate efforts to obtain pronouncements on disputed issues. In the American system courts, moreover, are the guardians of the Constitution, but they are not bound to consider all of the provisions of the Constitution justiciable. Under the doctrine of "political questions" the Supreme Court has refused at times to apply standards prescribed by or deducible from the Constitution to issues that it believed could be better decided by the political branches of government. For instance, Article IV, 4, provides that the states must have a republican form of government. Since *Luther* v. *Borden* (1849) it is settled that the court will not use the provision to invalidate state laws; it is for Congress and the president to decide whether a particular state government is republican in form. Many military and foreign policy questions, such as the question of the constitutionality of a particular war, have been likewise considered political and therefore nonjusticiable.

Political questions

Judicial review is more impartial than political review. This does not mean that it is entirely immune to political considerations of a general, abstract nature and to the impact of the evolution of the people's needs and of their political attitudes. As a matter of fact, the Supreme Court's reading of the Constitution has itself evolved in the course of two centuries in concomitance with the large transformations that have occurred in American society.

Given the nature of the U.S. constitutional system, the Supreme Court must see to it that the Constitution is respected and applied in three main areas: the relations between the states and the national government, the separation of powers within the national government, and individual rights. In each of these areas the court's conception of the Constitution has undergone substantial changes.

In the first half of the 19th century the Supreme Court succeeded in reinforcing the newly born structures of the federal system by interpreting generously those provisions of the Constitution that concern federal judicial power and those that set limitations on the states' powers to regulate interstate commerce. Later, when Congress began to enact laws intended to regulate some aspects of the nation's economic processes, the court was very cautious in granting the necessary powers. Up to the New Deal years (1932–37), the court, while admitting several federal regulatory interventions, also ruled that some economic relations, such as labour relations, lay almost entirely outside the scope of the commerce clause (the clause that gives Congress the power to regulate "commerce among the states") and were therefore matters exclusively for state legislation. After 1937 the court lifted the obstacles it had previously erected to federal interventions in the economic and social transactions of the nation. Today, under the new interpretation of the commerce clause laid down in such decisions as *Wickard* v. *Filburn* (1942) and under the doctrine of the federal spending power first enunciated in *U.S.* v. *Butler* (1936), Congress can make laws with respect to practically all subjects it deems of national relevance and in need of regulation from the centre.

The federal system

In the area of separation of federal powers the court has

Separation of powers

countenanced in the course of time a transfer of powers to the executive and to administrative agencies that probably was not envisaged by the Founding Fathers. Because all legislative powers are conferred by Article I, 1, of the Constitution upon Congress, the court at first ruled that such powers cannot be delegated by Congress to the executive. This doctrine was much diluted in the 20th century when it became clear that delegated legislation was necessary to govern a system of mixed economy. The court's attitude favourable to a reinforcement of the executive within the constitutional system has manifested itself also in other respects, notably in the field of foreign affairs. But the court has been able also to draw boundaries to an excessive expansion of the presidency. It has ruled that the president cannot, under the pretext of an emergency, disregard rules of conduct prescribed by Congress precisely for the circumstances of the case (*Youngstown Sheet and Tube Co.* v. *Sawyer,* 1952). It has established that the prerogative of the president to keep confidential statements secret must yield to the need of the judiciary to enforce criminal justice if the secret does not relate strictly to military or diplomatic matters (*Nixon* v. *U.S.,* 1973).

Rights of the individual

Until the New Deal the court used the provisions of the Constitution concerning individual rights primarily to protect property and economic liberties. That use helped to preserve a system of laissez-faire economy against state and federal efforts to interfere with the market. In particular, the "due process" clauses of the Fifth and Fourteenth amendments (no person shall be deprived of "life, liberty, or property, without due process of law") were often employed by the court to invalidate social legislation (*Lochner* v. *New York,* 1905; *Adair* v. *U.S.,* 1908). In the second half of the 20th century the posture of the court has changed entirely. The court today seldom concerns itself with economic liberties. It is engaged rather in protecting citizens' noneconomic freedoms as well as their equality before the law, focusing on issues such as civil and political rights, procedural rights in the criminal and administrative processes, or the right to privacy. In the course of developing this new jurisprudence the court has declared unconstitutional segregation in the schools (*Brown* v. *Board of Education of Topeka,* 1954) and malapportionment in electoral districts (*Baker* v. *Carr,* 1962; *Wesberry* v. *Sanders,* 1964); it has defended the rights of the suspect and of the accused (*Mapp* v. *Ohio,* 1961; *Miranda* v. *Arizona,* 1966); it has liberalized voluntary abortion (*Roe* v. *Wade,* 1973).

Viewed in the light of its two-century performance, U.S. judicial review can be assessed as an institution that defends the values of the political ideology prevailing in a given historical period against by and large occasional deviations from them on the part of the political branches of government. During the 19th and early 20th centuries, for example, the ideal of the minimal state and of a self-governing market was dominant with the elites of the Western world, and the Supreme Court did its best to enforce it in the peculiar context of the U.S. political system. At present the court is dedicated to furthering the values of the currently dominant ideal of a democracy: a system in which the equality and the noneconomic freedoms of persons are recognized and the state possesses all the necessary means to regulate the economy. Conflicts between the court and the political powers, state and federal, have occurred, but they have never been sharp except occasionally under particular circumstances: in the difficult years following the establishment of the new federal government and in the years of the Civil War; in the phase of transition from one to the other dominant political ideal (the New Deal years) and in the 1950s and 1960s, when the federal and state governments were seriously lagging behind in reshaping the legal system in accordance with fundamental requirements of the new democratic model.

## JUDICIAL REVIEW IN EUROPE AND ELSEWHERE

In Europe and other parts of the world the idea of making the judiciary guardian of the constitution was adopted much later than in the United States, as a rule only in the 20th century. While adopting it, however, some European countries introduced a variant on the U.S. model. Instead of letting any court pass on the constitutionality of statutes with the Supreme Court having the final say, as in the United States, they chose to set up a special constitutional court to which all questions concerning the validity of legislation must be referred and which alone has the power to declare statutes unconstitutional. In continental Europe judges are usually career functionaries; it was thought that the delicate task of invalidating legislation would be better performed by a court whose members were appointed directly either by parliament or, in part, by the president of the republic, the executive, the highest ordinary courts, and so forth.

European constitutional courts

The U.S. system of judicial review is "decentralized"; whereas the system based on a special constitutional court is "centralized." Austria was the first to inaugurate the centralized system (1920). After World War II it was adopted by Italy (1948) and West Germany (1949). Constitutional courts are at work today also in Spain, Portugal, and Turkey. In European centralized systems constitutional questions concerning the validity of statutory law reach the court chiefly through references by the ordinary courts engaged in deciding cases or through appeals by the losing parties of ordinary courts' decisions depriving them of their constitutional rights. But sometimes the questions can be raised also by such political agencies as the national executive and the regional governments or by a parliamentary minority. These can attack the law before the constitutional court without having to sue in an ordinary court. In certain instances the centralized method probably makes judicial settlement of constitutional issues easier and quicker than in the United States. And the American doctrine of "political questions" does not seem to have found reception, at least as such, in the jurisprudence of European centralized systems. Besides adjudging the validity of statutory law, European constitutional courts usually must also resolve conflicts between state agencies (the legislature, the executive, the president of the republic, and the judiciary) disputing about their respective constitutional prerogatives. In addition, they may sit as judge in trials upon impeachments and dispose of other matters of constitutional import.

Since 1958 France has had a Constitutional Council, which, though not a true court, can set aside unconstitutional statutes before they are promulgated, upon petition by the president of the republic or by the prime minister, the chairman of either of the two legislative assemblies, or a parliamentary minority.

Judicial review in other countries

The U.S. system of judicial review by ordinary courts has been adopted, sometimes with partial modification, by several states both in Europe and elsewhere, including Ireland (1937), Greece (1975), Japan (1946), India (1950), Mexico (1917), Brazil (1946), and Argentina (1949). It has been in operation in Switzerland, with some limitations, since 1874. It was introduced in Canada in 1867 and in Australia in 1900 and is presently an important feature of their constitutions.

Communist countries, with the exception of Yugoslavia, do not admit judicial review in their constitutions. The reason adduced is that it would undermine the unity of political powers and the sovereignty of the people as interpreted by the legislature. Hungary in 1984 and Poland in 1985 set up constitutional councils, which, however, serve in practice only as consultative organs for the legislature.

## TRENDS OF JUDICIAL REVIEW IN THE UNITED STATES AND IN EUROPE

Constitutional courts and supreme courts exercising judicial review outside the United States are often less politically influential than their American counterpart. Nevertheless, judicial review, wherever it is at work, has become an element of at least some weight in the constitutional processes of the state, and it tends in general to advance the same democratic values that have inspired the decisions of the U.S. Supreme Court since the 1930s.

European constitutional courts, in particular, have been able to correct here and there the legal system of their countries by reading the bill of rights present in their respective constitutions in ways that bear comparison with the American experience of judicial review. Sometimes, of

course, the trends of decisions diverge on special points, but the general direction seems to be the same. It may be remarked, however, that European courts have often been more cautious than the U.S. Supreme Court in expanding the freedoms of the individual at the expense of other competing values. A few examples will suffice.

Freedom of expression

In the area of freedom of expression the American doctrine of "clear and present danger," as restated in *Brandenburg* v. *Ohio* (1969), implies that no seditious or subversive speech can be punished unless it constitutes an incitement to immediate unlawful action and the incitement is likely to produce, in the circumstances, such action. The freedom to express unorthodox opinions is clearly recognized by European constitutions and is upheld by the constitutional courts, when they are confronted with laws that curtail it. The doctrine of clear and present danger, however, has not been adopted by them. The Italian constitutional court requires, for the punishment of speech advocating the use of violence, that the speech create, in the circumstances, a "danger," but it does not specify that the danger must be "immediate." The West German constitutional court, judging on the basis of constitutional provisions that forbid speech directed at impairing the liberal-democratic foundations of the state and associations pursuing the same goal, dissolved in the 1950s a neo-Nazi and a Communist party without considering the element of actual "danger" relevant. Later it countenanced, on the same basis, laws excluding from public employment persons holding subversive beliefs. In the United States the law of libel concerning public figures actively protects free speech inasmuch as, under the doctrine of *New York Times* v. *Sullivan* (1964), plaintiffs cannot win unless they prove that the libeler acted with "special malice" (*i.e.*, knowingly asserted the false). No special malice is constitutionally required in Europe to find liability in cases of defamation of public figures.

Abortion

The U.S. Supreme Court found in *Roe* v. *Wade* that a woman is entitled by the Constitution to obtain an abortion freely, after consultation with a doctor in the first trimester of pregnancy and in an authorized clinic in the second trimester. No European constitutional court has gone that far in recognizing freedom of abortion as part of the woman's right to privacy. The Italian court held in 1975 that voluntary abortions cannot be punished if performed in view of saving the life and health, both physical and psychological, of the woman. The Austrian court (1974) and the French Constitutional Council (1975), without tackling the problem of a woman's constitutional right to interrupt pregnancy, have validated statutes that provide in liberal terms for the possibility of voluntary abortions. The West German court, instead, quite alone in Europe, has ruled (1975) that the constitution, by declaring the life of persons inviolable, implicitly wants also the life of fetuses protected and that an adequate protection is afforded by the state only if voluntary abortion is made a crime by law. The court, however, has permitted the law not to punish abortions if they are performed in order to save the life or health of the woman, in the cases of rape and incest, or when the woman acted in conditions of extreme distress.

Church and state

The U.S. Supreme Court considers the teaching of religion and even praying in public schools unconstitutional (*Engel* v. *Vitale,* 1962). Separation of church and state, although contemplated in principle also by European constitutions, is sometimes tempered by constitutional provisions making accords between church and state possible in matters of common interest. No European court has found that accords giving students the opportunity to attend religious courses in public schools violate the principle of religious freedom or the principle of the equality of all citizens before the law.

In other areas of the law European constitutional courts have proved to be as ready as, and sometimes even more ready than, the U.S. court to afford protection of the rights of the individual. In the United States *Mapp* v. *Ohio* established that illegally obtained evidence cannot be produced at a trial to substantiate criminal charges against the defendant. This "exclusionary rule" is in force at least partially also on much of the European continent. The Italian constitutional court has stated that it is required on constitutional grounds. *Miranda* v. *Arizona* ruled that a confession made by an accused under arrest cannot be used as evidence unless the accused has been previously advised of his rights, among which is the right to consult with a lawyer. The Italian constitutional court has declared unconstitutional (1970) a law that excluded the suspect's attorney at the interrogation by the investigatory authorities and at other proceedings intended to secure evidence against him.

Right of action against the administration

Although as a rule courts in the United States can be asked to pass on the lawfulness of administrative action, the Supreme Court is still reluctant to establish as a matter of constitutional due process that citizens are always entitled to sue in court in order to have administrative decisions set aside if contrary to ordinary substantive or procedural rules. The Italian and German constitutions explicitly state the principle without admitting of any exception, and the courts of both countries carefully see to it that the principle is respected and that citizens are not deprived of their day in court even if the other party is the administration.

While applying the principle of equality in cases of sex discrimination and discrimination against illegitimate children, European courts have often displayed an activism superior to that of the U.S. Supreme Court in the same areas. The West German court, for instance, has ruled that husband and wife must have the same rights within the family and that, in particular, parental power over the children belongs equally to both, the law not being permitted to prefer the husband's will in case of divergences of opinions. The Italian court has in many respects reshaped family law in order to ensure the equal rights of the wife and of children born out of wedlock. It has also defended the right of women to a treatment equal to that of men in labour relations. Effective legislative protection against discrimination of non-European immigrant workers and their families is still deficient in EEC countries, and by and large constitutional courts have said little in this area. But they have shown in general remarkable sensitivity when the problem affects local ethnic or linguistic minorities. The U.S. Supreme Court, after having declared segregation unconstitutional, found that programs of "affirmative action" meant to help minorities to rise from their lower position in society did not constitute "reverse discrimination" even if they included some privileges in the access to university education (*University of California* v. *Bakke,* 1978) and in the conservation of jobs. A ruling analogous to this has come from the Italian court with regard to the treatment of the German-speaking population of Alto Adige. The court in the 1970s validated laws that provided for reserved "quotas" in public employment and publicly financed housing with a view of preserving the integrity of that linguistic group in the region they had inhabited for centuries.

It is true that thus far European courts have never openly defied the political powers of the state in the way the U.S. Supreme Court has sometimes done. Even when the European courts have somehow challenged such powers by annulling laws that were supposedly of special interest to them, the conflict has not really been acute. But the greater prudence of European courts is not difficult to explain. It depends on many causes, prominent among which are the facts that their legitimation as independent and active agencies within the political system is recent and that the tradition of judicial review does not yet have in Europe the firm roots it possesses in the United States.

### TRANSNATIONAL JUDICIAL REVIEW

Judicial review has also transcended the boundaries of the national state. All western European states except Finland have ratified a convention on civil and political rights, first signed in 1950. On the basis of the convention a commission was organized in Strasbourg, Fr., to which individuals can appeal against actions of a member state that violate rights protected by the convention. The commission and a member state can ultimately refer the issue to a court, the European Court of Human Rights, for a final decision. The court, besides granting a remedy in

the pending case, may find statutory and other national laws contrary to the provisions of the convention, and the state concerned is under an obligation to adapt its legal rules to the principles stated by the court. Great Britain too has accepted this special jurisdiction of the European Court of Human Rights; this means that in its political system possible legislative encroachments on individual rights can now be rebuked by a judicial agency, though not a national one. In countries having an internal system of judicial review the European protection is added to that provided by the national system.

The decisions of the European Court of Human Rights are important because they may gradually work as a factor that will unify this branch of the law throughout Europe. Some national constitutional courts, notably the Austrian court, already pay particular attention to the jurisprudence of the European court. So does the court of the European Economic Community, which while not having to apply a community bill of rights (the Treaty of Rome contains none) must nevertheless take into account the problems of fundamental rights in developing the economic law issuing from the legislative organs of the EEC.

A convention on human rights similar to the European was signed by several Latin-American states at San José, Costa Rica, in 1969. A court having jurisdiction over individual complaints began functioning in 1982.

Thus the idea of the rights of the individual, after having contributed three centuries ago to the birth of the modern constitutional law of the national state, has now become the mainspring of another incipient, promising experience: judicial review with transnational dimensions. (G.B.)

**BIBLIOGRAPHY.** For the definition of the concept of constitutional law and of the elements, necessary and contingent, that make up a constitution, see HANS KELSEN, *General Theory of Law and State,* trans. from German and French (1945, reissued 1971). See also H.L.A. HART, *The Concept of Law* (1961, reprinted 1981). For the national origins of modern constitutions, see HANS KOHN, *The Idea of Nationalism: A Study in Its Origins and Background* (1944, reprinted 1977). A concise description of the contribution made by the idea of the "inalienable rights" of the individual to the development of modern constitutional law, together with an analysis of the recent expansion of the protection of such rights at the international level, can be found in LOUIS HENKIN, *The Rights of Man Today* (1978). Valuable works on modern constitutionalism include CARL J. FRIEDRICH, *Constitutional Government and Democracy: Theory and Practice in Europe and America,* 4th ed. (1968); and KARL LOEWENSTEIN, *Political Power and the Governmental Process,* 2nd ed. (1965).

The texts of almost all the state constitutions presently in force in the world are available in English translations in A.P. BLAUSTEIN and G.H. FLANZ (eds.), *Constitutions of the Countries of the World: A Series of Updated Texts, Constitutional Chronologies and Annotated Bibliographies,* 17 vol. (1971– ), with quarterly revisions published for loose-leaf update. On federalism as a form of government, see WILLIAM S. LIVINGSTON, *Federalism and Constitutional Change* (1956, reprinted 1974); K.C. WHEARE, *Federal Government,* 4th ed. (1963, reprinted 1980); CARL J. FRIEDRICH, *Trends of Federalism in Theory and Practice* (1968); and MICHAEL BURGESS (ed.), *Federalism and Federation in Western Europe* (1986). For a history of the separation of powers in Europe and America, see M.J.C. VILE, *Constitutionalism and the Separation of Powers* (1967).

For surveys of modern constitutional trends, see JOHN A. HAWGOOD, *Modern Constitutions Since 1787* (1939, reprinted 1987); HERBERT J. SPIRO, *Government by Constitution: The Political Systems of Democracy* (1959); C.F. STRONG, *Modern Political Constitutions: An Introduction to the Comparative Study of Their History and Existing Forms,* 8th rev. and enl. ed. (1972); K.C. WHEARE, *Modern Constitutions,* 2nd rev. ed. (1966, reprinted 1980); and ARNOLD J. ZURCHER (ed.), *Constitutions and Constitutional Trends Since World War II: An Examination of Significant Aspects of Postwar Public Law with Particular Reference to the New Constitutions of Western Europe,* 2nd ed. (1955, reprinted 1975). The most comprehensive, summary study of the forms of government presently existing in the world is probably PAOLO BISCARETTI DI RUFFÍA, *Introduzione al diritto costituzionale comparato,* 6th ed. rev. (1988).

For studies of the constitutions of particular groups of countries, see WILLIAM DALE, *The Modern Commonwealth* (1983); MARTIN C. NEEDLER (ed.), *Political Systems of Latin America,* 2nd ed. (1970); W.F. ABBOUSHI, *Political Systems of the Middle East in the 20th Century* (1970); and H. GORDON SKILLING, *The Governments of Communist East Europe* (1966). Surveys of individual countries include E.C.S. WADE and A.W. BRADLEY, *Constitutional and Administrative Law,* 10th ed. (1985), a classic on Great Britain's constitution; EDWARD MCWHINNEY, *Canada and the Constitution, 1979–1982: Patriation and the Charter of Rights* (1982); PETER HANKS, *Australian Constitutional Law,* 3rd ed. (1985); GEORGE ARTHUR CODDING, JR., *The Federal Government of Switzerland* (1961); KLAUS VON BEYME, *The Political System of the Federal Republic of Germany* (1983); WILLIAM PICKLES, *The French Constitution of October 4th, 1958* (1960); ALAN GLEDHILL, *The Republic of India: The Development of Its Laws and Constitution,* 2nd ed. (1964); ARDATH W. BURKS, *The Government of Japan,* 2nd ed. (1964, reprinted 1982); JERRY F. HOUGH and MERLE FAINSOD, *How the Soviet Union Is Governed* (1979); and ARYEH L. UNGER, *Constitutional Development in the USSR: A Guide to the Soviet Constitutions* (1981, reprinted 1986). For the European Economic Community, see EMILE NOËL, *The European Community: How It Works* (1979).

General works on the U.S. Constitution include EDWARD S. CORWIN, *Edward S. Corwin's The Constitution and What It Means Today,* 14th ed. rev. by HAROLD W. CHASE and CRAIG R. DUCAT (1978); ARTHUR N. HOLCOMBE, *Our More Perfect Union: From Eighteenth-Century Principles to Twentieth-Century Practice* (1950, reprinted 1967); C. HERMAN PRITCHETT, *The American Constitution,* 3rd ed. (1977); and LAURENCE H. TRIBE, *American Constitutional Law,* 2nd ed (1988). LEONARD W. LEVY (ed.), *Encyclopedia of the American Constitution,* 4 vol. (1986), is a comprehensive, multidisciplinary reference work. For history, see MAX FARRAND, *The Framing of the Constitution of the United States* (1913, reprinted 1974); CARL BRENT SWISHER, *American Constitutional Development,* 2nd ed. (1954, reprinted 1978); ALFRED H. KELLY, WINFRED A. HARBISON, and HERMAN BELZ, *The American Constitution: Its Origins and Development,* 6th ed. (1983); and PHILIP B. KURLAND and RALPH LERNER (eds.), *The Founders' Constitution,* 5 vol. (1987), a monumental collection of 17th-, 18th-, and 19th-century documents that bear on all parts of the Constitution.

On judicial review as an institution, and the constitutional law produced by it, see, in general, MAURO CAPPELLETTI, *Judicial Review in the Contemporary World* (1971). On judicial review in the United States, see EDWARD S. CORWIN, *The "Higher Law" Background of American Constitutional Law* (1929, reprinted 1971); ROBERT G. MCCLOSKEY, *The American Supreme Court* (1960, reprinted 1964); and HENRY J. ABRAHAM, *Freedom and the Court: Civil Rights and Liberties in the United States,* 4th ed. (1982).

For a comparative analysis of judicial review in the United States and other countries, see WALTER F. MURPHY and JOSEPH TANENHAUS, *Comparative Constitutional Law: Cases and Commentaries* (1977); and MAURO CAPPELLETTI and WILLIAM COHEN, *Comparative Constitutional Law: Cases and Materials* (1979). See also FRANCIS G. JACOBS, *The European Convention on Human Rights* (1975).

(G.B./D.Fe.)

# Continental Landforms

Continental landforms are the surface features of the largest land areas of the Earth. Such structures are rendered unique by the tectonic mechanisms that generate them and by the climatically controlled denudational systems that modify them through time. The resulting topographic features tend to reflect both the tectonic and the denudational processes involved. The most dramatic expression of tectonism is mountainous topography, which is either generated along continental margins by collisions between the slablike plates that make up the Earth's lithosphere or formed somewhat farther inland by rifting and faulting. Far more subtle tectonic expressions are manifested by the vast continental regions of limited relief and elevation affected by gentle uplift, subsidence, tilting, and warping. The denudational processes act upon the tectonic "stage set" and are able to modify its features in a degree that reflects which forces are dominant through time. Volcanism as a syn-tectonic phenomenon may modify any landscape by fissure-erupted flood basalts capable of creating regional lava plateaus or by vent eruptions that yield individual volcanoes.

The denudational processes, which involve rock weathering and both erosion and deposition of rock debris, are governed in character by climate, whose variations of heat and moisture create vegetated, desert, or glacial expressions. Most regions have been exposed to repeated changes in climate rather than to a single enduring condition. Climates can change very slowly through continental drift and much more rapidly through variations in such factors as solar radiation.

In most instances, a combination of the foregoing factors is responsible for any given landscape. In a few cases, tectonism, some special combination of denudational effects, or volcanism may control the entire landform suite. Where tectonism exists in the form of orogenic uplift, the high-elevation topography depends on the nature of denudation. In humid or glacial environments whose geomorphic agencies can exploit lithologic variations, the rocks are etched into mountainous relief like that of the Alps or the southern Andes. In arid orogenic settings, the effects of aggradation and planation often result in alluviated intermontane basins that merge with high plateaus interrupted or bordered by mountains such as the central Andes or those of Tibet and Colorado in the western United States.

In continental regions where mountainous uplifts are lacking, denudational processes operate on rocks that are only slightly deformed—if they are sedimentary—and only moderately elevated. This produces broad basins, ramps, swells, and plains. These are most thoroughly dissected in rain-and-river environments (sometimes attaining local mountainous relief on uplifts). Elsewhere, they may be broadly alluviated and pedimented where mainly arid, or widely scoured and aggraded where glacial.

Minor denudational landforms are superimposed on the major features already noted. Where aridity has dominated, they include pediments, pans, dune complexes, dry washes, alluvial veneers, bajadas, and fans. Ridge-ravine topography and integrated drainage networks with associated thick soils occur where humid conditions have prevailed. Combinations of these features are widespread wherever arid and humid conditions have alternated, and either category may merge laterally with the complex suite of erosional and depositional landforms generated by continental glaciers at higher latitudes.

This article deals with the major continental landform categories of each type. It describes the main topographic expressions, probable modes of origin, and distribution with respect to continental configurations. In addition, it reviews the significant theories of landform genesis developed during roughly the past two centuries.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 212, 231, and 232.

This article is divided into the following sections:


Theories of landform evolution 705
- General observations 705
  - Basic concepts and considerations
  - Constraints on modern landform theory
- Historical survey 706
  - Landform theories of the 18th and 19th centuries
  - Davis' erosion cycle theory and related concepts
  - Climatic morphogenesis
  - Tectonic geomorphology
  - Consideration of unique landforms and other distinctive topographic features
- Theoretical overview 709
  - Orogenic and epeirogenic morphogenesis
  - A unified landform theory

Tectonic landforms 712
- Mountains, mountain ranges, and mountain belts 713
  - Geomorphic characteristics
  - Tectonic processes that create and destroy mountain belts and their components
  - Major types of mountain belts
  - Major mountain belts of the world
- Plateaus 723
  - Geomorphic characteristics
  - Formative processes
  - Geographic distribution
- Tectonic basins and rift valleys 724
  - Geomorphic characteristics
  - Principal types
- Volcanic and tectonic caves 726
  - Volcanic caves
  - Tectonic caves

Structural landforms 728
- Stream valleys and canyons 728
  - Geomorphic characteristics
  - Origin and evolution
  - Other landforms associated with stream erosion 733
    - Planation surfaces
    - Inselbergs
- Landforms associated with stream deposition 735
  - Alluvial fans and bajadas
  - Playas, pans, and saline flats
- Sand dunes 737
  - Geomorphic characteristics
  - Formation and growth of dunes
  - Dune and sheet patterns
  - Fixed dunes in semiarid regions
- Glacial landforms 739
  - General considerations
  - Erosional landforms
  - Depositional landforms
  - Periglacial landforms
- Caves and karst landscape 745
  - Cave types
  - Evolution and demise of solution caves
  - Geomorphic characteristics of solution caves
  - Solution cave features
  - Karst topography
  - Geographic distribution of karst terrain
- Landforms produced by coastal processes 751
  - Factors and forces in the formation of coastal features
  - Landforms of erosional coasts
  - Landforms of depositional coasts

Other types of landforms 755
- Impact craters 755
  - General characteristics of impact craters on the terrestrial surface
  - Formation of impact craters
  - Population of terrestrial impact craters
- Biogenic landforms 758

Bibliography 758

# THEORIES OF LANDFORM EVOLUTION

## General observations

### BASIC CONCEPTS AND CONSIDERATIONS

Landform evolution is an expression that implies progressive changes in topography from an initial designated morphology toward or to some altered form. The changes can only occur in response to energy available to do work within the geomorphic system in question, and it necessarily follows that the evolution will cease when the energy is consumed or can no longer be effectively utilized to induce further change. The latter steady state, or dynamic equilibrium, situation will then continue with little topographic change until the prevailing conditions cease or are disrupted, so that a new evolutionary sequence can begin.

The English poet Alfred, Lord Tennyson once wrote:

> The hills are shadows, and they flow
> From form to form, and nothing stands;
> They melt like mist, the solid lands,
> Like clouds they shape themselves and go.

Tennyson's verse speaks well of the geomorphic necessities of time and landform change. Even the ancients were well aware of the ongoing effects of gravity, and it has long been realized that, given time and in the absence of opposing forces, gravity would pull the Earth's surface roughness down to form a featureless subaqueous spheroid. Such an evolution would be simplicity in the extreme and may in fact foretell the eventual destiny of terrestrial landforms when internal processes that generate relief cease to operate some billions of years hence in response to growing entropy in the system.

Even now in regions where the uplifting and relief-creating mechanisms have been inoperative for several hundreds of millions of years, the lands have been reduced by denudation to low and often nearly featureless plains. Yet, it is clear that any modern theory of landform evolution must take into account the possibility of a periodic regeneration of continental elevations, particularly of large-scale relief features. For without such regeneration, there would be no continents or mountains even today, given their present rates of erosional destruction.

The history of landscape evolution theory is one of adapting concepts to new evidence of increasing complexity. This situation is quite apparent in the way thinkers and scientists have dealt with the processes within the Earth that oppose gravity and re-create land elevation and roughness. The existence of such processes was implicit in the writings of Xenophanes of Colophon (*c.* 570–*c.* 478 BC), Herodotus (*c.* 484–420? BC), and Leonardo da Vinci (AD 1452–1519). The culmination of ideas of continental renewal and relief genesis is found in the isostatic theory formulated by John Henry Pratt and George Biddel Airy of England during the mid-1800s and in the concepts of plate tectonics put forth by Harry H. Hess and Robert S. Dietz of the United States during the early 1960s. Periodic resurrection of the surface roughness of the Earth is an event that geologists continually plot, widely accept, and increasingly understand.

Continental renewal and relief genesis

Over the years there have been many other ideas that have posed complications for geomorphic theory. Notable among these were notions of continental submergence by seas (proposed by Georges-Louis Leclerc, comte de Buffon, about 1750), which had implications of relative sea-level changes and sedimentary leveling of submersed areas.

Theoretical matters were complicated further by suggestions during the 19th century that iceberg rafting of gravel during Noah's Flood accounted for glacial "drift." Since that time, the Noachian Deluge has lost much of its geomorphic appeal. Yet, sedimentary deposits laid down in ancient inland seas are widely acknowledged to account for much continental bedrock, and they underlie and create vast structural plains in areas such as Australia.

The role of climate in landform evolution

The geomorphic implications of volcanism were already widely appreciated in the 1700s, though they were not well integrated into modern tectonic mechanisms until 1961. Climate, however, is another story. Glacial theory was introduced during the early 1800s and was seen by many to have climatic and geomorphic implications. Nonetheless, the most popular theory of landform evolution of the past century, that proposed by the American geologist and geographer William Morris Davis (*c.* 1899), relegated continental glaciation to accidental status and gave no real consideration to the geomorphic effects of non-glacial climates. Until about 1950 this Davisian view held sway in geomorphology. Since then, research has shown beyond question that a variety of climatic effects can have a profound influence on landscape, that climates change (often with great frequency and intensity), and that virtually none of these events can be termed accidental.

### CONSTRAINTS ON MODERN LANDFORM THEORY

Rather than merely trace the hit-or-miss development of geomorphic ideas from their beginnings roughly two centuries ago, it seems preferable to cite here those conditions firmly determined by intensive research that must serve as constraints for any modern theory of landform evolution. A brief mention of the postulates of earlier theorists will then show immediately what they accomplished, ignored, failed to consider, or were ignorant of. In sum, a modern theory of landform evolution must contend with the following well-established factors:

(1) Continents consist of a craton of crystalline rocks 1,000,000,000 to 3,000,000,000 or more years old, have been periodically submerged by epicontinental seas, and are in most cases locally covered with veneers of nearly flat-lying sedimentary rocks.

(2) Where orogenic events were involved less than 500,000,000 years ago, mountainous elevations and relief containing deformed rocks exist on continents.

(3) Lowering of the land by denudational processes is accompanied by essentially continuous isostatic adjustment by load-compensating uplift.

(4) Mountainous relief of the continent-to-continent collision type (*e.g.,* the Appalachian Mountains of eastern North America) can eventually be eliminated by erosion, whereas trench-type mountains (*e.g.,* the Andes of western South America) probably cannot as long as the associated trench subduction system endures.

(5) Climates on lands vary through time in response to lateral continental drift of 0–12 centimetres (0–5 inches) per year. North America, for example, is moving northwest at a rate of about three centimetres per year. On the other hand, Antarctica is hardly moving and has been in a polar position undergoing glaciation for about 30,000,000 years.

(6) Over most lands, climates also vary with atmospheric, oceanic, and solar factors in cycles lasting thousands of years (the Milankovitch solar radiation cycle, for example, has a duration of ± 25,800 years).

(7) In select hydrographically favoured sites on time scales not influenced by continental drift (*e.g.,* Antarctica), climates on continents or portions thereof can remain essentially constant for periods of millions of years.

(8) Since geomorphic processes under arid, humid, glacial, and possibly other climate conditions can induce particular landforms, areas subject to periodic climate change often show polygenetic landform associations.

(9) Landforms exposed for millions of years to a constant environment may display a climax (steady-state) landform association that is essentially timeless and in which landform evolution through denudation is reduced to mere negative allotropic growth.

(10) Since volcanism is seemingly localized in accordance with mobile heat-dispersal patterns within the Earth, eruptive effects may be imposed on any surficial geomorphic system at any stage of development.

(11) Similarly, mobile tectonic patterns involving rock deformation may be brought to bear on any surficial geomorphic system, with resulting relief, elevation, and topographic changes.

(12) Impacts on planetary surfaces by falling meteoroids,

asteroids, and cometary bodies are periodic but are capable of generating landforms of mountainous proportions. Such impacts are apparently diminishing in frequency and scale with time.

(13) Surficial geomorphic agents of denudation responsible for many, if not most, landforms include mass wasting, running water, glacial ice, and wind. They are not all of equal significance in every climatic setting, however.

(14) The geomorphic agents respond to various climates, changing in character and effect. They also respond in some degree to altered conditions of elevation and relief.

(15) The behaviour of denudational agencies and related geomorphic processes is neither constant nor linear in nature. Rates vary from long-term, imperceptible, and gradual to brief, rapid, and catastrophic.

(16) Changes produced by geomorphic agents vary in magnitude but not directly with time—*i.e.,* the same change involving the same energy expenditure may be either slow or fast. (Studies of river systems, for example, suggest that greater changes in channel morphology occur during brief infrequent floods than during protracted low-flow periods.)

(17) Perturbations in geomorphic processes or environments cause accelerated changes in most landform configurations, soils, and deposits, which eventually slow down as new equilibrium forms develop.

(18) A given landform or deposit is only stable in association with its formative process and environment, and in any subsequent alternative setting it begins to change toward a new equilibrium morphology.

(19) In a denudational setting, slope as an influence over process rate may be subordinated to such factors as runoff volume, soil-moisture content, bedrock coherence, ground-cover type, channel roughness, channel cross section, weathering type, sediment calibre, and sediment quantity.

(20) When climate in a region changes, elimination of relict landforms and deposits causes a disequilibrium phase, which is followed by a dynamic equilibrium phase as new geomorphic equilibria are established. The disequilibrium phase may range from a few score or hundred years for certain organic responses to many thousands of years for soil, hillslope, or drainage adjustments.

(21) Certain landforms can only develop when a particular climatic sequence ensues. For example, major marine deltas form on rivers that drain newly humid regions following glaciation or aridity; streams from regions that have long been humid have no deltas.

(22) Some landforms or deposits, once formed, strongly resist subsequent changes regardless of climatic history—*e.g.,* entrenched meanders such as those that exist in parts of the Appalachians, chert felsenmeers (accumulations of rock blocks) like those in the southern Ozark region of the United States, and duricrusts of the type commonly found in Australia.

(23) No such thing as an "average" terrestrial climate seems to exist, and certainly a climatic "norm" for one continental configuration would differ from that for another—*e.g.,* the supercontinent Pangaea of pre-Cretaceous times (more than 136,000,000 years ago) differed climatically from its subsequent fragments for both the Cretaceous (136,000,000 to 65,000,000 years ago) and the present.

(24) Sea level has been found wanting as a stable limiting datum for erosional processes or as an influence on stream behaviour. Glacioeustatic fluctuations on the order of 130–150 metres appear to have been commonplace during the continental glacial sequences of the Carboniferous (345,000,000 to 286,000,000 years ago), Pleistocene (2,000,000 to 10,000 years ago), and at other times, and periodic dessication of restricted ocean basins has occasionally permitted major rivers to deepen their courses thousands of metres below mean sea level.

(25) The Earth and the solar system as a whole are at least 4,500,000,000 years old. This is long enough for some geomorphic phenomena to occur several times but probably not long enough for others to happen even once. Certainly it is doubtful if more than nine collision-type mountain systems can have been eroded away in one spot, even if it were possible for them to form there.

## Historical survey

Some of the more significant landform theories of the past 200 years or so are considered here, with particular attention to the degree to which they reflect the list of geomorphic constraints cited above. It should be noted that most early theorists operated within the chronological limitations imposed by theologians. During the 17th century, for example, Archbishop James Ussher of Ireland added up the ages of men cited in the Old Testament of the Bible and concluded that the creation had occurred in 4004 BC. John Lightfoot, an English divine and Hebraist, was so stimulated by this revelation that he additionally observed that the exact time was October 26 at 9:00 AM! This meant that all of the Earth's surface features had to have been formed in less than 6,000 years. Given this time frame, geomorphologists could explain the genesis of landforms in only one way—on the basis of catastrophic events. Everything had to occur quickly and therefore violently.

### LANDFORM THEORIES OF THE 18TH AND 19TH CENTURIES

**Catastrophism.** During the late 18th and early 19th century, the leading proponent of this view was the German minerologist Abraham Gottlob Werner. According to Werner, all of the Earth's rocks were formed by rapid chemical precipitation from a "world ocean," which he then summarily disposed of in catastrophic fashion. Though not directed toward the genesis of landforms in any coherent fashion, his catastrophic philosophy of changes of the Earth had two major consequences of geomorphic significance. First, it indirectly led to the formulation of an opposing, less extreme view by the Scottish scientist James Hutton in 1785. Second, it was in some measure correct: catastrophes do occur on the Earth and they do change its landforms. Asteroid impacts, Krakatoa-type volcanic explosions, hurricanes, floods, and tectonic erosion of mountain systems all occur, may be catastrophic, and can create and destroy landforms. Yet, not all change is catastrophic.

**Uniformitarianism.** The Huttonian proposal that the Earth has largely achieved its present form through the past occurrence of processes still in operation has come to be known as the doctrine of uniformitarianism. This is a geologic rather than a simply geomorphic doctrine. It is, however, more nearly aimed at actual surficial changes that pertain to landforms than were Werner's notions. The idea championed by Hutton formed the basis of what is now often referred to as process geomorphology. In this area of study, research emphasis is placed on observing what can be accomplished by a contemporary geologic agency such as running water. Later, the role of moving ice, gravity, and wind in the molding of valleys and hillslopes came to be appreciated by study of these phenomena. Uniformitarianism also became the working principle for a growing number of geologic historians, notably William Smith and Sir Charles Lyell, in the 19th century. This was necessary as Lyell argued increasingly that geologic change was incremental and gradual. He needed a longer time scale if this approach was to work, and geologic historians were finding it for him.

Basis of process geomorphology

**Gradualism.** Lyell's concept of gradualism and accompanying process observation on an expanded time scale resulted in firmly establishing the fact that much could be accomplished by small forces working constantly for long periods. That conclusion is consistent even with present-day thought. Lyell's almost total rejection of any geologic process that was abrupt and suggestive of catastrophe, however, was in itself an extreme posture. Research has shown that both gradual and rapid changes occur.

In the philosophical climate established by Hutton's uniformitarianism and Lyell's gradualism, geomorphologists of the 19th century realized many impressive accomplishments. Most notable among these were the studies of glacial phenomena in Europe by Johann von Charpentier and Louis Agassiz and the investigations of regional denudation in the American West by Grove K. Gilbert and Clarence E. Dutton, which emphasized the work of running water. The findings pertaining to glaciers still stand for the most part, and Gilbert's hydraulic studies laid the

groundwork for modern ideas. Yet, neither he nor Dutton made comprehensive theoretical proposals of terrestrial morphogenesis of a scope that could match those of the aforementioned W.M. Davis.

## DAVIS' EROSION CYCLE THEORY AND RELATED CONCEPTS

**The geographic cycle.** Beginning in 1899, Davis proposed that denudation of the land occurs in what he called "the geographical cycle." According to Davis, this cycle is initiated by an uplift of an area above sea level, followed by a wearing down of the surface through the action of running water and gravity until either the region is worn away (base leveled) or the events are interrupted by renewed uplift (Figure 1). It was further explained that such a cycle of erosion occurs under conditions of a rain-and-rivers environment (what present-day investigators would call a humid climate), which were assumed to reflect the normal climate for the Earth. The fact that Davis dismissed glacial phenomena as accidents of climate and viewed climatic areas as geographically fixed afforded his theory more latitude. Furthermore, Davis proposed the idea of a separate arid geographical cycle in 1905. In all cases, erosive power was presumed to be controlled primarily by slope; hence, the cyclic system was slowed down as the land was leveled and relief and elevation were diminished. The end point of a low-inclination landform was termed a peneplain, and it was said to be locally surmounted by erosionally resistant highs called monadnocks. The peneplain as a whole was presumed to be graded to regional base level (in all likelihood mean sea level) by denudational agencies (*e.g.,* running water), which were supposedly controlled by this datum.

Slope as a denudational influence

The provisions of Davis' erosion cycle run counter to at least half of the 25 constraints on theories of landform evolution listed above. The Davisian erosion cycle theory is hurt by three factors in particular: (1) the presently understood need for continuous isostatic uplift during erosion, (2) the climatic variability displayed by most lands, and (3) the hydraulic behaviour of rivers noted by Gilbert that precludes valley alluviation under normal humid conditions and limits base-level influences over interior slopes.

The notion of an erosion cycle initiated by uplift is still possible within known constraints. Such a cycle is only possible under one particular climatic umbrella, however, and under much more limited geographic and hydrographic circumstances than Davis had assumed. Moreover, the morphological sets of landforms selected by Davis as chronological "mile posts" for his cycle of landform change (*i.e.,* stages of development) have been found to constitute special, generally polygenetic arrays of landscape features that reflect the interplay of several environments and that have little or no sequential time significance.

**Davisian dynamic equilibrium.** Davis' contribution to the theory of landform evolution also includes the idea of process interruption as a means of accelerating change (rejuvenation) and the notion of process slowing in a late stage of process evolution as energy is consumed. The latter idea comes close to the present-day description of dynamic equilibrium, or attainment of a steady-state (climax) environment and parallels modern thinking on entropy relationships.

Davis proposed his scheme of landscape development stages close on the heels of Charles Darwin's theory of organic evolution, and his designations "youth," "maturity," and "old age" (Figure 1) are blatantly anthropomorphic. Thus, it is quite understandable why they had, at the turn of the century, such appeal and acceptance in spite of their actual lack of chronological significance. Their continued use is less comprehensible.

**The geomorphic concepts of Penck and King.** The theoretical groundwork laid by Davis for geomorphic evolution was further developed in a rather special fashion in 1924 by Walther Penck of Germany, and subsequently (1953) championed with variations by Lester C. King of South Africa. Both retained some Davisian devices—peneplain, graded stream, and base-level control of erosion surfaces in Penck's case and the latter two in King's. Each thought that tectonic uplift punctuated the erosion cycle by initiating renewed stream incision, and each utilized the concept of parallel retreat of fluvial-structural escarpments to generate plains. King designated the planation process pedimentation, and his end point "pediplains" were surmounted by inselbergs (isolated hills standing above plains, the name being derived from the German term for "island mountains") rather than monadnocks. Because the resulting stair-stepped landscapes (*Treppen,* the German word for "steps") of scarps and flats (Figure 2) were presumed to reflect tectonics and to be correlatable, the term Tectonic Geomorphic School has been applied to its advocates.

The *Treppen* concept

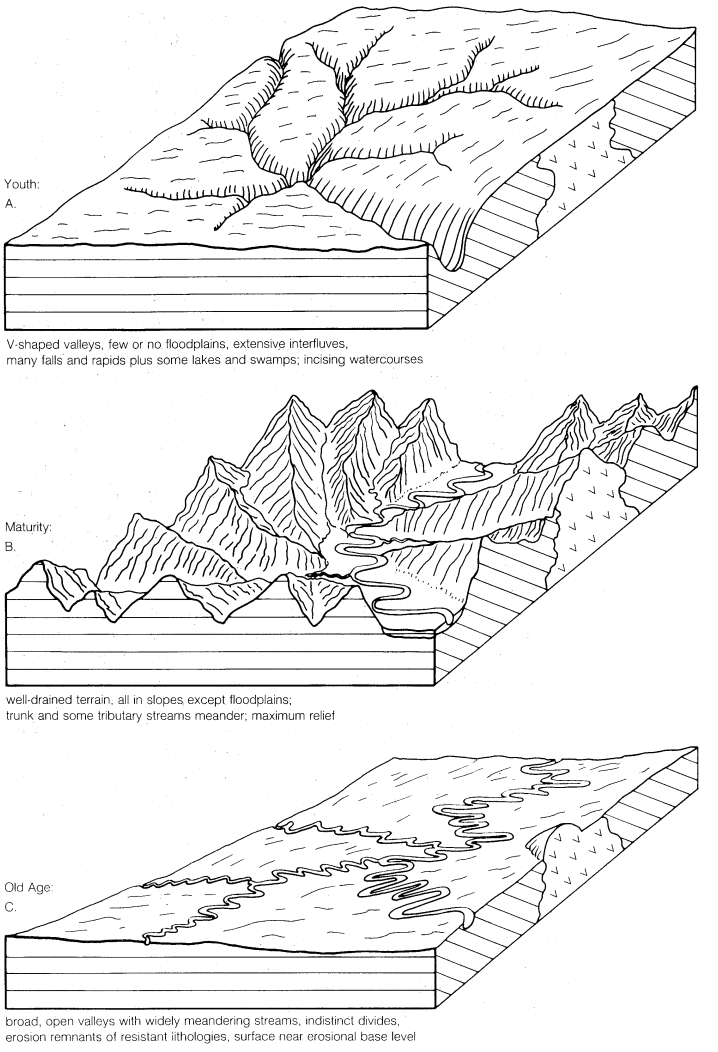



Figure 1: Davis' proposed landscape development states. The morphology shown is not actually time indicative. For example, (A) could be a gully system in soft sediment or a canyon such as the Royal Gorge in Colorado, which is millions of years old. The ridge-ravine topography of (B) would normally develop under humid conditions, but the river meandering on alluvium indicates a prior or extraneous non-humid aggrading mechanism. The riverine plain of (C) implies a complex history of planation and aggradation in a current fluvial mode.

Adapted from H.F. Garner, *The Origin of Landscapes* (1974); Oxford University Press, Inc.

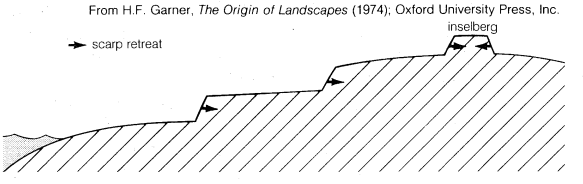



Figure 2: Cross section of an area undergoing erosion by escarpment retreat under Penck's *Treppen* mechanism (see text). Regional base level (± mean sea level) was presumed to provide a limiting downward erosional datum following each episode of uplift marked by stream incision and escarpment development.

The notion of geomorphologists that denudational landforms reflecting tectonic pulses were sufficiently synchronized on a global basis to be correlatable has suffered much from the development of the plate tectonics theory (see PLATE TECTONICS). The separate notion that hillslopes, once developed, retreated laterally to produce a low-inclination surface worthy of a special name (pediment–pediplain) has found more support.

In retrospect, Penck's *Treppen* concept seems to suffer much of the same theoretical damage as Davis' geographical cycle, but it is generally less ingenious. Like Davis, Penck and King made no dynamic use of climatic influences, and in fact the latter went so far as to claim that climate makes no difference. And like Davis, neither King nor Penck acknowledged the isostatic implications for erosion established nearly a century earlier. King suggested that sheetfloods "mold" the surfaces of pediments and depicted sparsely vegetated regions where this might be possible under the label semiarid. More recent work suggests that sheetfloods may be a product, rather than a cause, of the "flat" terrain on which they occur. The so-called molding would appear to be the result of desert stream-flood processes operating to local base levels in the absence of appreciable plant cover, as will be discussed below.

There is an implied landform "chronology" for a geomorphic system tied to intermittent uplift, as suggested by Penck and King, though dating such events is not readily accomplished. Furthermore, King tied his planation method to a regional sea-level erosional datum that the aforementioned constraints throw into question. Perhaps the principal contribution of the Penck–King theoretical ensemble has to do with the concept of lateral escarpment retreat, as opposed to the wearing down of lands favoured by Davis. There are in fact landforms that are widely acknowledged to be pediments. They are planar in form, truncate a wide variety of bedrock types, and can most readily be explained by scarp retreat under non-vegetated conditions. Debate continues about how much or how little moisture best encourages this process. Yet, at least the general nature of the mechanism seems to have been identified (largely by detailed studies in the area of process geomorphology) and the hydraulic constraints established by Gilbert and others seem to be satisfied.

In essence, it has been found that runoff deposits sediment in deserts where its excess transport energy is dissipated by volume loss caused by infiltration and evaporation. Runoff upslope from the depositional base level established by the long-term locus of deposition cannot erode below the resulting deposit (Figure 3). Such overland flow must expend its energy against non-vegetated hillslopes, resulting in their backwearing.

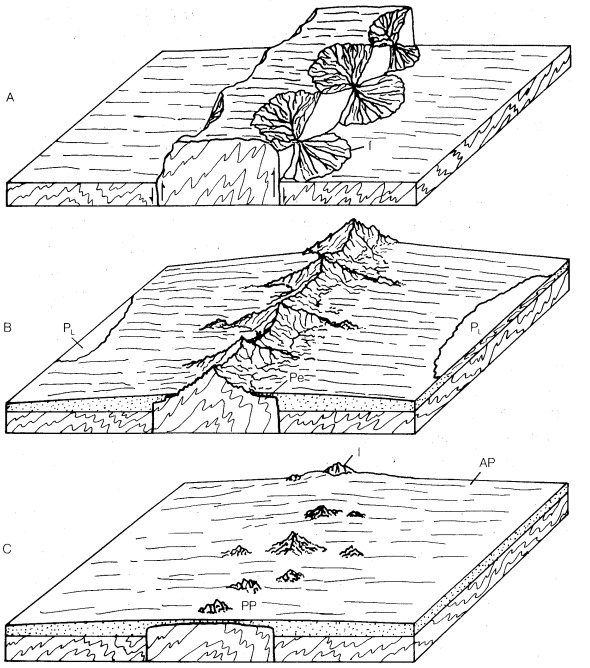



Figure 3: Three-phase block diagram of pedimentation of an upland in a desert. The process of "scarp" retreat and planation is accomplished by sheet wash on non-vegetated surfaces, but it cannot begin until a local base level of erosion–deposition is established. Streams dissecting the upland cannot cut below the level created where deposition of alluvium begins as runoff dissipates. The long-term locus of that deposition established the datum for lateral stream-bank and valley-wall recession at higher elevations. Representative phenomena are designated by letters: alluvial fans (f), marginal pedimentation of horst (Pe), possible playa lake development ($P_L$), pediplain (PP), inselbergs (I), and aggradation plain (AP).

From H.F. Garner, *The Origin of Landscapes* (1974); Oxford University Press, Inc.

Significance of the pedimentation phenomenon

The pedimentation phenomenon must rank as one of the more astute geomorphic insights, regardless of the fact that the hydraulic and sedimentologic details involved were not established until later. Today, this form of land planation in association with alluvial aggradation in deserts, stream incision that establishes regional drainage networks and augments relief under humid conditions as described by Davis, and glacial scour and deposition as elucidated by Charpentier, Agassiz, and others stand as the three most widely established morphogenetic systems on Earth.

## CLIMATIC MORPHOGENESIS

Recognition of climatic dynamism

**Morphogenetic area.** Notions that climate plays a major dynamic role in landform evolution only began to emerge about the mid-1900s. At that time the French, particularly Jean Tricart, André Cailleux, and Louis C. Peltier, began to employ the concept of a morphogenetic area—a region in which a particular set of landforms is being generated under a particular climate. Only slowly, however, and mainly from studies in the tropics did it come to be appreciated how extreme the regional climate shifts between arid and humid have been on the different continents. Davis long ago understood how distinctive the geomorphic mechanisms of humid and arid lands were. It was, however, the new evidence of wide geographic mobility for such environments that forced the recognition of the morphogenetic, or geomorphic, system. Such a system is defined as a group of agencies and processes interacting under a particular environment to produce a landscape. Because morphogenetic areas and their systems can displace each other, it follows that they would leave behind relict landforms, soils, deposits, organisms, and so forth.

The discovery of widespread climatic dynamism and the correlative recognition of plate-tectonic phenomena created a whole new theoretical situation for geomorphologists. Not a single theory of regional landform development existing in 1950 accounted for the constraints imposed by the new climatic and tectonic findings in any significant way.

**Interactions between geomorphic systems.** Climates change and periodically impose one of the foregoing geomorphic systems on the relicts left by one of the others. In addition, areas of each climatic type export matter to adjacent morphogenetic areas and thereby modify the resulting landforms. For example, deserts export dust by eolian means, and the resulting deposits modify soil profiles in downwind regions, as in the eastern United States, or create actual depositional landforms of loess, as in Shansi Province of China on the lee of the Ordos Desert. River systems arising in humid lands develop their drainage networks therein and then may encroach on downslope deserts to create alluvial riverine plains where their flow will not maintain their sediment transport to some distant ocean. Alternatively, rivers form deltas following climate change when their sediment loads and flow are sufficient and the débouché (point of emergence) is protected. Glaciers produce their changes on ice-covered realms and then export their outwash deposits into whatever environment is downslope.

## TECTONIC GEOMORPHOLOGY

Tectonic geomorphic systems

In addition to the usual climatic imprints, orogenic tectonism (including volcanism) adds its obvious dimensions of elevation and slope to any surficial environment it encounters. It is now clear that orogenic realms in their early

phases create gravitational opportunities for Earth sculpture that hardly exist elsewhere. The usual mechanisms for concomitantly gradualistic denudation by ice, wind, and running water are set aside in orogenic belts by relatively rapid uplifts of material ranging from nearly unconsolidated sediment to semicoherent but intensely deformed masses of metamorphic and igneous rocks. Under these conditions, masses of rock measured in thousands of cubic kilometres are torn loose by gravity and fall and/or slide, often moving hundreds of kilometres in a "geologic instant" to a lower resting place (in some cases lubricated by subaqueous avenues). The term catastrophic seems most appropriate for an occurrence of this type (see Figure 4C).

Sculpturing of the Earth is thus seen as more than the mere gradual removal of weathered debris by mechanisms

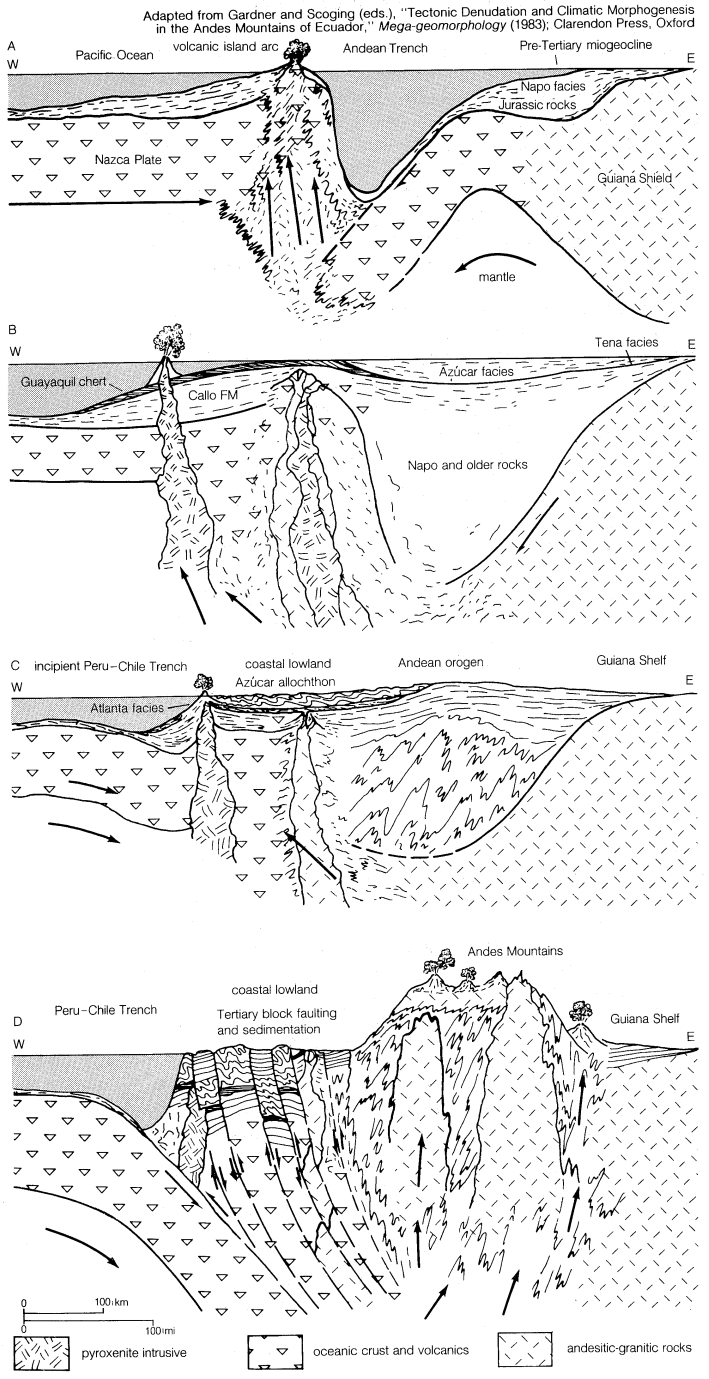



Figure 4: Diagrammatic east–west cross section of the Ecuadorian region in (A) Late Cretaceous, (B) Paleocene, (C) early Eocene, and (D) Oligocene. The Azúcar allochthon reflects tectonic erosion of the rising Andes Mountains. In the early Eocene, about 15,000 to 20,000 cubic kilometres of rock slid at least 250 kilometres in a very brief period of time.

under the control of climatic regimes. The Kamchatka Peninsula in the far eastern part of the Soviet Union is said to have more than 100 active volcanoes. Not surprisingly its terrain is dominated by volcanic landforms. The Afar Triangle at the foot of the Red Sea is shaped by newly formed faults that cut unweathered basaltic lava flows on a newly emergent seafloor in an almost totally tectonic landscape. In the Appalachians, south of the glaciated knobs, an ancient mountain system sheathed by thick saprolitic soils on its upper slopes exhibits ridge-ravine topography and may have been in a humid climatic nucleus for 100,000,000 years. Yet, the same region retains water gaps and entrenched meanders that echo drainage patterns established long ago, probably on alluvial cover masses of Early Mesozoic age (roughly 225,000,000 years old) following an arid-to-humid climate change at the end of the Jurassic Period (about 186,000,000 years ago). In the same area, tropical soils and ridge-top lateritic deposits of Georgia and Alabama reflect weathering conditions established 150,000,000 years ago when southeastern North America was still in the tropics before recent northwesterly continental drift.

### CONSIDERATION OF UNIQUE LANDFORMS AND OTHER DISTINCTIVE TOPOGRAPHIC FEATURES

There are, of course, instances where special types of bedrock combine with particular weathering and erosion regimes to produce unique landforms and landscapes. Best known perhaps are the solutional effects expressed as karst topography. This is most pronounced in limestone terrain, such as that in Kentucky in the southeastern United States and the Karst plateau in Yugoslavia, as well as those in parts of northeast China and on islands like Puerto Rico and Jamaica. In tropical realms where silica is more soluble, similar landforms may develop on other varieties of sedimentary rock or on igneous or metamorphic types, as, for example, quartzite in the isolated plateau remnants of the Venezuelan Guiana Shield. The humid climatic conditions that promote solution production and dripstone formation are readily apparent in such tropical areas.

Karst features

Granitic terrain in several parts of the world also gives rise to a distinctive array of landforms that include domed erosion residuals, often in patterns closely tied to joint spacing in bedrock as noted by the Australian geomorphologist C.R. Twidale. In regions where alternating humid and arid climates or human activity have led to erosional stripping of weathered zones, mammoth boulder piles of exhumed core stones exist. Such features are especially notable on the island of Hong Kong, in southern Brazil, in parts of India and Australia, and in the St. Francois Mountain region of Missouri in the United States.

## Theoretical overview

The complexities of terrestrial surface change demand a theoretical overview that is both flexible and multifaceted. Oversimplified, sweeping landscape generalizations that apply to the whole Earth such as the postulates of Davis and King can hardly be employed when dealing with a planet where virtually every geomorphic element constitutes a potential interruption or complication to every other system. Nevertheless, there do seem to be certain kinds of activity that are repeated sporadically in both tectonic and climatic realms. These repetitions encourage the re-creation of particular suites of landforms and could be taken to imply a certain rationality to events. However, they probably are no more rational than eddies in a river that—in a parody of Murphy's Law—develop only where possible.

Matters of geographic and chronological scale also enter into the question of what is indeed geomorphically possible and repeatable. The interplay between density variations in matter and gravity dictates that the Earth's core (once formed) must remain firmly fixed, and so too must the lighter substances that make up the lithosphere. Concentration of the least dense solids in the continents is involved in a complex process now associated with plate tectonics, and it is at this level that a discussion of landform evolution must begin.

Continents as first-order landforms

Although the designers of the plate tectonics theoretical framework did not single out continents as landforms of a special kind, such is one of the basic consequences of that theoretical construct. Continents are first-order landforms, and there seemingly will be only one cycle of continental denudation in the history of the Earth. It began with the earliest concentration of continental lithosphere at the surface. And it presumably will end, as suggested above, when the last endogenic forces (*i.e.,* those within the Earth) expire and gravity and entropy have their way as the internal systems of the planet run down. The details (in the context of the 8,000,000,000- to 10,000,000,000-year span of this cycle) hardly matter, since the results are inevitable—unless, of course, the Sun becomes a nova and disrupts things.

Second-order features on continents consist primarily of mountains and the relatively low-elevation areas that come into existence as the mountains rise. In the context of continental landforms, mountains and the geomorphic systems that act upon them are unique in that the uplift creates an excess of potential energy, one far above that of the remaining land area. Landform evolution in mountains is necessarily skewed by this special kind of excess energy. Davis seemed to sense this in his theorizing, but he did not understand the limits on slope as a denudational influence and the variety of climatic and tectonic factors at work.

### OROGENIC AND EPEIROGENIC MORPHOGENESIS

**Orogenic geomorphic systems.** Such mountain-building systems evolve in the special contexts of type, setting, and style. The principal orogenic varieties recognized are (1) mountains of continent-continent collision type formed by lithospheric plate interaction along continental margins, (2) mountains of the collision type associated with oceanic trenches (sometimes developed along a single continental margin) with an adjacent plate-tectonic subduction system (see below), and (3) rift-type mountains extending into continental interiors where transcurrent faults shear cratons and deform associated sediment veneers or where spreading zones develop to create fault-block (horst-graben) mountainous terrain. Geologic time is sufficient for several orogenic events of each type to have occurred, and different rules apply to the geomorphic evolution of any given type.

Factors in the evolution of mountainous terrains

Mountains of the continent-continent collision type have special attributes that direct their geomorphic evolution. These distinctive characteristics are the following:

(1) The collision creating the mountains incorporates a finite volume of rock that is not augmented following the collision.

(2) The orogenic rock mass is subject to isostatic uplift during denudation; in general, sedimentary rock types are exposed first, followed by crystalline varieties.

(3) The collision that initiates such orogenesis ultimately adds rock to the adjacent craton, and in thickening the adjacent crust often initiates nearby cratonic tilting and/or uplift.

(4) Because such mountains develop between continents and are thus elevated in the midst of a consequent mega-continent (Pangaea in the case of the Appalachians), they are far from oceanic evaporation sources and therefore often undergo initial denudation under arid geomorphic systems in the manner of the present mountains of central Asia.

(5) As the climatic setting of such mountains is largely established tectonically, it may endure in the same climate for scores of millions of years and, as noted in 1901 by the American geomorphologist Douglas W. Johnson, a desert mountain range tends to bury itself in its own waste.

(6) Re-exposure of such mountains to nearby precipitation sources by plate adjustments may result in dramatic climate changes from arid to humid, so that perennial fluvial erosion is widely initiated on a relict arid, alluvial cover mass with resulting transverse drainage by superimposition—one can compare the Appalachian Mountains of North America and the Zagros Mountains of Iran, as described by the American geomorphologist Theodore M. Oberlander in 1965.

(7) Because of their finite initial rock volume, mountains of the continent-continent collision type can be lowered by erosion, somewhat in the manner visualized by Davis. No such structures more than 500,000,000 years old show mountainous relief.

(8) Volcanic landforms are rarely a part of the topography during orogenesis of this mountain type.

Mountains of the collision type associated with oceanic trenches have their own distinct attributes that control evolution. These are as follows:

(1) The merging of a pair of lithospheric plates along a deep-sea trench initiates orogenesis tied to the subduction process (*i.e.,* the sinking of one plate beneath another at convergent plate boundaries).

(2) Rock mass is added to the orogenic belt via subduction as long as the trench remains "operational."

(3) Denudation accompanies uplift and may reduce rock mass in the orogenic system in the long run, but whether the total mass is growing, shrinking, or static depends on the budget established by additions from subduction versus losses from erosion.

(4) Mountainous elevations tend to increase through much of the life of the orogenic system, since rock lost through erosion is generally removed locally and linearly by rivers and glaciers (the Andes exemplify the type bordering a continent, and they appear to be higher now than at any time since they began to form 150,000,000 years ago).

(5) Because mountains of the trench-associated subduction type develop and endure adjacent to an ocean on at least one side, they are subject to climatic variability tied to such factors as latitudinal position, orientation with respect to prevailing wind patterns, ocean surface temperatures, and progressively increasing elevations.

(6) Examples such as the Andes that border a continent can show alternating segments that are highly volcanic.

(7) Andean types also may display highly contrasting denudational systems under a variety of climatic conditions on opposite sides as well as along the length of the range.

(8) Although an erosion cycle resulting in overall lowering of a trench-associated mountain system does not appear viable as long as the trench endures, a complex steady-state mass situation would seem to be one potential development during this time.

(9) Occasionally orogenesis related to trench-continent interaction may extend far inland; the parts of the Andes exhibiting this trait display mechanical rock deformation but little volcanism, and a similar genetic mechanism has been suggested for the Rocky Mountains of North America.

(10) During their early years, the Rocky Mountains displayed volcanic phases accompanied by upthrusting but now seem tectonically quiescent and are apparently experiencing denudational lowering.

Rift-type mountains are primarily of the block-fault variety. They have the following set of special attributes:

(1) Block-fault mountains appear to originate where a spreading ridge of the plate-tectonic type develops.

(2) On continents, the spreading is expressed in high-angle faulting and may be accompanied by volcanism of tholeiitic basalt type.

(3) Rifting may be limited to linear zones, as in the Rift Valley system of East Africa, or may be more broadly expressed, as in the Basin and Range Province of the western United States.

(4) The extent of rifting may be limited to mere surficial fracturing of the continental crust, or it may extend to actual rupturing of a lithospheric plate and renewal of seafloor spreading, as occurred along the Atlantic seaboard of North America at the end of the Jurassic.

(5) Because block-fault mountains are of endogenic origin, they may occur in and experience a variety of denudational environments. The examples from Africa and North America cited above are in settings ranging from arid to humid. The highest such mountains show glacial effects.

For a detailed discussion of mountains and their evolution, see below *Mountains, mountain ranges, and mountain belts.*

**Climatically dominated epeirogenic realms.** The epeirogenic portions of continents (*i.e.*, those that have escaped orogenesis in the past 500,000,000 years) experience denudation in a situation in which the slope factor, if at all tectonic in origin, is regional in expression and so gentle as to exert little influence beyond giving direction to flowing water or ice. It is these regions that variously exhibit veneers of sedimentary rock largely accumulated in epicontinental seas over the past 500,000,000 years or that expose in shield areas the roots of worn-down mountain systems. In the absence of notable tectonism, it is not surprising to find that morphogenesis on stable cratons is dominated by climate. Vast expanses of cratons situated away from mountain belts either are occupied by temperate and tropical forests and grasslands or are seared by desert heat and wind. Only Antarctica currently sports a continental ice sheet, but both North America and Eurasia show they recently did so as well. It is in these epeirogenic regions that morphogenesis is most significantly punctuated by climate change. With few exceptions, the landforms are polygenetic. Many of the most recent glacial deposits scarcely show the incipient soil development begun under humid conditions only a few thousand years ago. Furthermore, broadly forested, humid regions still exhibit patches of cacti and alluvium left there when they were deserts. Therein, the notable slopes are denudational in origin; the steeper ones were usually developed by stream incision and the more gentle ones commonly were produced by alluviation and/or pedimentation.

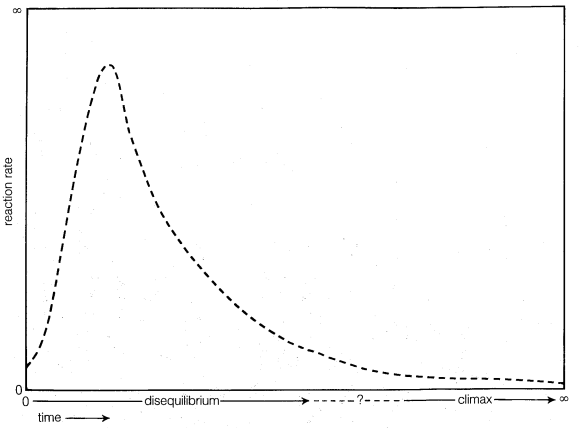



Figure 5: Graph showing the long-term results of a perturbation in a geomorphic system. Where uplift is the cause and creates excess mass subject to gravity and erosion, the disequilibrium phase may endure hundreds of millions of years and the climax, or expression of dynamic equilibrium, would be marked by the elimination of the excess relief and elevation. Where climate change is the cause and leaves relict landforms in a new environment, the disequilibrium continues until the relicts are eliminated or modified into a climax morphology that is in equilibrium with the new system (see text).

From H.F. Garner, "Geomorphic Analogs and Climax Morphogenesis," Arkansas Academy of Science Proceedings (1963).

## A UNIFIED LANDFORM THEORY

Viewed in their entirety, the individual concepts that pertain to landform development so far discussed (catastrophism, uniformitarianism, gradualism, erosion cycle, dynamic equilibrium, disequilibrium, geomorphic system, morphogenetic area, tectonic geomorphology, and orogenic and epeirogenic morphogenesis) have to date been treated by theorists as independent conceptual constructs rather than as geomorphic elements of a unified comprehensive theory. There is a close parallel between this situation and the fable of the several blind men who decided what an elephant is by touching only individual parts of the animal. For each of there geomorphic concepts has a measure of validity, but the earliest ideas were formulated on the basis of very incomplete information. When considered in the context of the entire solar system, in which there is a group of planetary geomorphic entities, the theoretical pieces begin to fall into more distinctly rational positions. Although a degree of variability is imposed by planetary location and by early differentiation of cosmic material, randomness in the solar system is incomplete because of the directional factors imposed by gravity, radiation, and increasing entropy. For any given planet, there are two potential geomorphic factors: (1) exogenic impact phenomena from solar debris possibly modified by tidal disruption caused by nearby planetoids, or radiation phenomena tied mainly to the Sun resulting principally in climatic influences and biologic activity, and (2) endogenic phenomena related to internal heating and expressed as tectonism and volcanism, as on the Earth. Morphogenesis occurs in accordance with interaction between planetary subsystems associated with the above factors.

Consideration of factors in the context of the entire solar system

**Behaviour of geomorphic systems.** Gravity-driven geomorphic systems are potentially cyclical in terms of the elimination of excess relief and elevation. They exhibit activity that graphs in two-phase form—initial disequilibrium when free energy and relief are maximal (and the results are frequently catastrophic), and subsequent dynamic equilibrium where relief and elevation are nearly eliminated and free energy available to do work is so low that change is nearly imperceptible (Figure 5). The latter behaviour is clearly gradualistic. Such systems must be disturbed by outside forces in order for the cycle to be interrupted or reinitiated.

Disequilibrium and dynamic equilibrium

In the solar system the cycle of accretionary, gravity-propelled impact morphogenesis that creates cratered surfaces and high relief is in a distinctly waning phase. Such activity apparently reached a peak within the first 1,000,000,000 years after the planetary system was formed and is not likely to be renewed. Its expression is epitomized by the surface of objects such as the Moon and the planet Mercury, where the near absence of endogenic tectonic forces has left impact effects most intact. On the Earth and a few other planets (or satellites), internal heating propels orogenesis and thereby periodically renews gravity-driven geomorphic cycles. As noted earlier, there will be only one continent-forming cycle in the history of the Earth.

Radiation-driven geomorphic systems are tied to the Sun's nuclear fusion processes and the fluctuations therein. Because of atmosphere and organisms, solar effects are most singularly manifested on the Earth as morphogenetic areas characterized by a particular climate and associated processes. The geomorphic changes in such areas are cyclical largely with respect to the destruction of relict features exposed to the system as the morphogenic areas move and also with respect to the creation of landforms and deposits in morphological equilibrium with the new system. Changes in landforms, deposits, and processes also graph in two phases after the initiation of a system or after a perturbation in one (Figure 5). These landform changes are initially time-indicative, and unless morphogenesis has attained a dynamic equilibrium phase, the partially altered relict features may permit reconstruction of the events of landform evolution.

It will be noted from the above that there is a close relationship between process and form in the dynamic equilibrium phase of radiationally driven geomorphic systems (Figure 6). In morphogenetic areas in states of disequilibrium, form (strongly influenced by relict features) may show little or no conformance to process, which may have just been initiated. Relict features in the process of transformation, such as a desert or a glacial alluvial deposit in a valley being reworked by a perennial stream, thus constitute hybrid features (compare with Davis' mature stream in Figure 1B). The stream valley illustrated has a flat floor unlike that of a late-phase humid valley, as shown in Figure 6, which has a V-shaped cross profile. Furthermore, the "hybrid" stream is not behaving as it would if there were no alluvium, and the alluvium is not the same after the stream has partially reworked it.

Occasionally, the sequence of geomorphic events may conspire to preserve a form that is foreign to the associated geomorphic system and processes. The sinuous paths of entrenched meanders that are cut into bedrock in such regions as the Appalachians express the granular surface and sediment-water volume relations that prevailed when the flow pattern was initiated in the Mesozoic rather than those of the present.

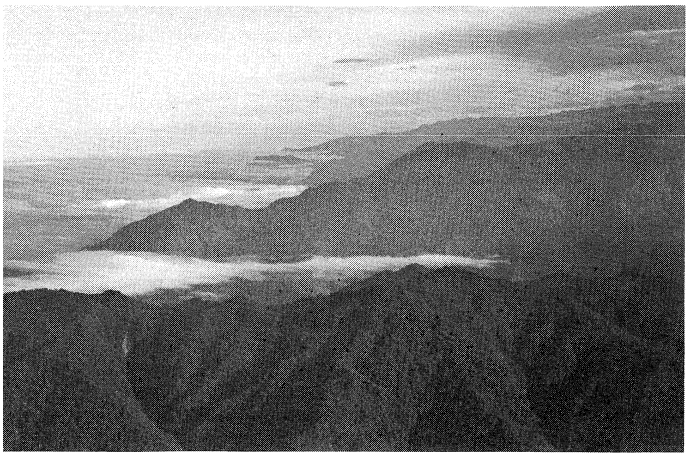


Figure 6: Typical selva (ridge-ravine) topography developed on the forested, humid western spurs of the Andes in central Ecuador. The valleys have V-shaped cross profiles and lack alluvial deposits at their bottoms.
H.F. Garner

**The concept of periodic random dominance.** On the Earth, gravity- and radiation-driven geomorphic systems interact independently, so that their two types of activity can mingle under conditions of periodic random dominance. Thus, peak energy expenditures engendered by each type of system may or may not coincide geographically. Maximum rates of landform change occur where active orogenesis mingles with changing climates. Minimal change occurs where epeirogenic regions are occupied by morphogenic areas that are in states of dynamic equilibrium. In this arrangement of interacting geomorphic systems, there is clearly a place for both catastrophe and gradualism. There also is a place for cycles of erosion of several kinds and for dynamic equilibrium, either as an end phase of enduring climatic morphogenesis and/or as an end phase of relief and elevation reduction by denudation following orogenesis.

The concept of periodic random dominance as an aspect of landform evolution carries with it the implication of polygenetic landforms and landscapes where geomorphic system dominance fails to develop. Indeed, dominance becomes the special case because it is dependent on a particular juxtaposition of tectonic and/or climatic elements over a protracted interval in a given area. One estimate places polygenetic landforms over approximately 80 percent of the Earth's land surface. Perhaps 20 percent is experiencing some type of geomorphic system dominance—less than 10 percent if Antarctica is omitted from the calculations.

The possibility of polygenetic landforms and landscapes

**Process geomorphology and systems equilibria.** Details of landform evolution within a given geomorphic system are matters of process behaviour and terrain response. In the context of geomorphic system dominance versus systemic alternation, two general situations exist: (1) those agencies operating in contact with relicts that they are modifying, often quite rapidly, and (2) those in contact with equilibrium features that they have created and have little or no ability to modify further. The principal surficial geomorphic agencies on Earth—wind, running water, glacial ice, and gravity—in any given geomorphic system induce processes that tend to evolve toward a situation of least work. Polygenetic terrain is usually some combination of hillslopes and "flats," and either topographic type may dominate in the latter part of a geomorphic cycle, depending on whether the system tends to generate relief or reduce it.

Natural geomorphic systems operating along the Earth's surface are classified as open, since they are powered by external energy sources. Because the rates of both endogenetic and exogenetic energy input vary, the coordinate agencies experience changes analogous to power surges in an electrical system. Thus rivers receiving excess runoff periodically flood. The atmosphere locally builds up excess heat, and the transfer of this heat is expressed in storms. Glaciers, normally the epitome of slowness, can acquire a mass-energy excess and consequently surge. In all instances, energy available for erosion, transportation, and deposition of sediment varies greatly over time. In addition, the interaction between solids, fluids, and gases results in turbulence, eddy formation, shearing and vortex activity, and periodic local stagnation.

In response to the foregoing situations, process associations within individual geomorphic systems exhibit typical systems phenomena, including "feedback," "threshhold reactions," and evolution toward dynamic equilibrium (least-work) modes. Where a system is periodically perturbed, processes can pass back and forth between disequilibrium and steady-state conditions rather frequently.

The behaviour and apparent process direction of an individual agency may not reflect the evolution of the overall geomorphic system. For example, a 10,000-year-long episode leading to the formation of an alluvial fan may be seen to include numerous incidents of fan-head trenching that are separately destructive but subordinate to depositional events dominating the trend. Similarly, a river such as the Mississippi that is reworking a relict alluvial deposit in a valley may be seen to be depositing gravel on point bars on the insides of bends. The long-term consequence of the river's activity, however, will be to remove the entire alluvial deposit in its path, including the point bars, unless subject to systemic interruption. (Humankind has of course "short-circuited" the natural evolution of the Mississippi and that of many other rivers with engineering modifications.)

From the foregoing, it seems evident that the direction of landform evolution can only be grasped from the study of geomorphic process if the character and role of relict landforms and deposits are clearly understood. This is an obvious complication in the application of Hutton's doctrine of uniformitarianism.

The concept of periodic geomorphic system dominance provides the rational potential end point of landform evolution under a particular set of conditions. Ideally, it may yield either modified or unmodified tectonic landscapes. These in turn may be either orogenic or epeirogenic. Where modified, they may express marine effects and/or glacial, arid, or humid morphogenesis. Antithetically, where more common polygenetic morphogenesis occurs, some mixture of tectonic, marine, or climatic effects is superimposed on the setting, and a hybrid suite of landforms results. (H.F.G.)

# TECTONIC LANDFORMS

Whereas erosion shapes landforms, their origins lie with tectonic processes that build the major structures of the Earth—mountain belts, mountain ranges, and some individual mountains, as well as plateaus and certain kinds of valleys (rift valleys) and basins. The word tectonic is derived from the Greek word *tekton,* which means "builder." Tectonic processes build landforms mainly by causing the uplift or subsidence of rock material—blocks, layers, or slices of the Earth's crust, molten lavas, and even large masses that include the entire crust and uppermost part of the planet's mantle. In some areas, these processes create and maintain high elevations such as mountains and plateaus. In others, they produce topographic depressions, as exemplified by Death Valley in the western United States, the Dead Sea in the Middle East, or the Turfan Depression in western China. Virtually all areas below sea level have been formed by tectonic processes.

Effects of tectonic processes

Mountain ranges and plateaus result either from the uplift of the Earth's surface or from the emplacement of volcanic rock onto the surface. Many mountain ranges consist of chains of volcanoes that are made up of rocks derived from depths of tens of kilometres below the sur-

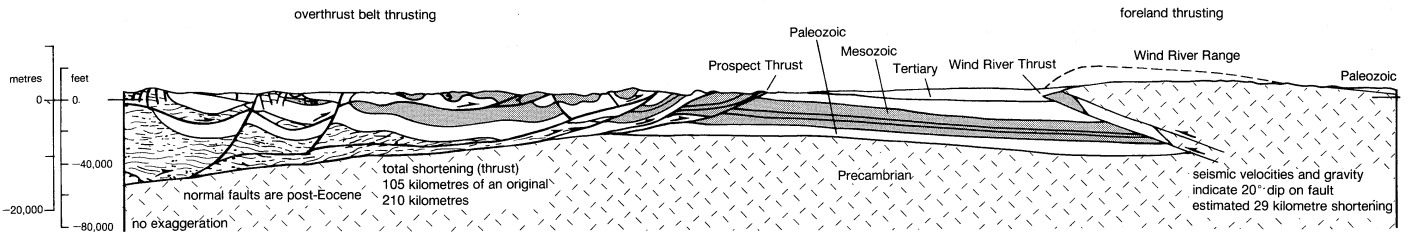



Figure 7: East–west cross section of the fold and thrust belt in eastern Idaho (left) and the block-faulted uplift of the Wind River Range in Wyoming (right).

Adapted from Robbie Gries, "Oil and Gas Prospecting Beneath Precambrian of Foreland Thrust Plates in Rocky Mountains," *The American Association of Petroleum Geologists Bulletin*, vol. 67, no. 1 (January 1983); © copyright 1983, The American Association of Petroleum Geologists, all rights reserved

face. Some plateaus are created by huge outpourings of lavas over vast areas. In addition, the intrusion of molten rock into the crust from below can raise the surface. Many other mountain ranges have been formed by the overthrusting of one terrain or block of crust over an adjacent one, which is another mechanism that uplifts the surface (Figure 7). Similarly, the folding of rocks at the surface creates the ridges and valleys that define some mountain chains (Figure 8). These processes of overthrusting (or underthrusting) and folding result from horizontal forces that cause crustal shortening (in its horizontal dimension) and crustal thickening. Finally, heating and thermal expansion of the outer 100 to 200 kilometres of the Earth can uplift broad areas into either mountain ranges or plateaus.

Similarly, tectonic valleys, basins, and depressions of smaller size can form by the reverse of two of the processes mentioned above. Crustal extension (in its horizontal dimension) and crustal thinning occur where two blocks of crust move apart; a valley or basin forms between such blocks where the intervening segment of crust has been thinned and its top surface subsides (Figure 9). Likewise, subsidence of the Earth's surface can occur by the cooling and the thermal contraction of the outer 100 kilometres of the planet. Plateaus and entire mountain ranges can subside by this mechanism to form large basins in some areas.

Virtually all large-scale landforms are the result of both tectonic processes that built the large differences in elevation and erosional processes that sculpted the relief of such areas into their individual shapes. Thus, it might be said that tectonic processes built the Alps, but erosional processes gave the Matterhorn its unique profile. In all cases, erosion acts to reduce differences in elevation, but when the rate of erosion is not too rapid, landforms created by tectonic processes can persist for hundreds of millions of years after the processes have ceased to operate.

## Mountains, mountain ranges, and mountain belts

A mountain belt is a large tract of land, many tens to hundreds of kilometres wide and hundreds to thousands of kilometres long, that stands above the surrounding surface, which usually lies near sea level. Within mountain belts are individual mountain ranges or chains, which extend tens to hundreds of kilometres in length and consist of individual mountains that are connected by ridges and separated by valleys. Within many such belts are plateaus, which stand high but contain little relief. Thus, for example, the Andes constitute a mountain belt that borders the entire west coast of South America; within it are both individual ranges, such as the Cordillera Blanca in which Peru's highest peak, Huascarán, lies, and the high plateau, the Altiplano, in southern Peru and western Bolivia.

### GEOMORPHIC CHARACTERISTICS

Mountainous terrains have certain unifying characteristics. Such terrains have higher elevations than do surrounding areas. Moreover, high relief exists within mountain belts and ranges. Individual mountains, mountain ranges, and mountain belts that have been created by different tectonic processes, however, are often characterized by different features.

Volcanic chains

Chains of active volcanoes, such as those occurring at island arcs, are commonly marked by individual high mountains separated by large expanses of low and gentle topography. In some chains, namely those associated with "hot spots" (see below), only the volcanoes at one end of the chain are active. Thus those volcanoes stand high, but with increasing distance away from them erosion has reduced the sizes of volcanic structures to an increasing degree.

The folding of layers of sedimentary rocks with thicknesses of hundreds of metres to a few kilometres often leaves long, parallel ridges and valleys termed fold belts, as, for example, the Valley and Ridge province of Pennsylvania in the eastern United States. The more resistant rocks form ridges, and the valleys are underlain by weaker ones. These fold belts commonly include segments where layers of older rocks have been thrust or pushed up and over younger rocks. Such segments are known as fold and thrust belts. Typically their topography is not as regular as where folding is the most important process, but it is usu-

Adapted from Vinton E. Gwinn, "Thin-Skinned Tectonics in the Plateau and Northwestern Valley and Ridge Provinces of the Central Appalachians," Geological Society of America Bulletin, vol. 75, pl. 2, (September 1964)

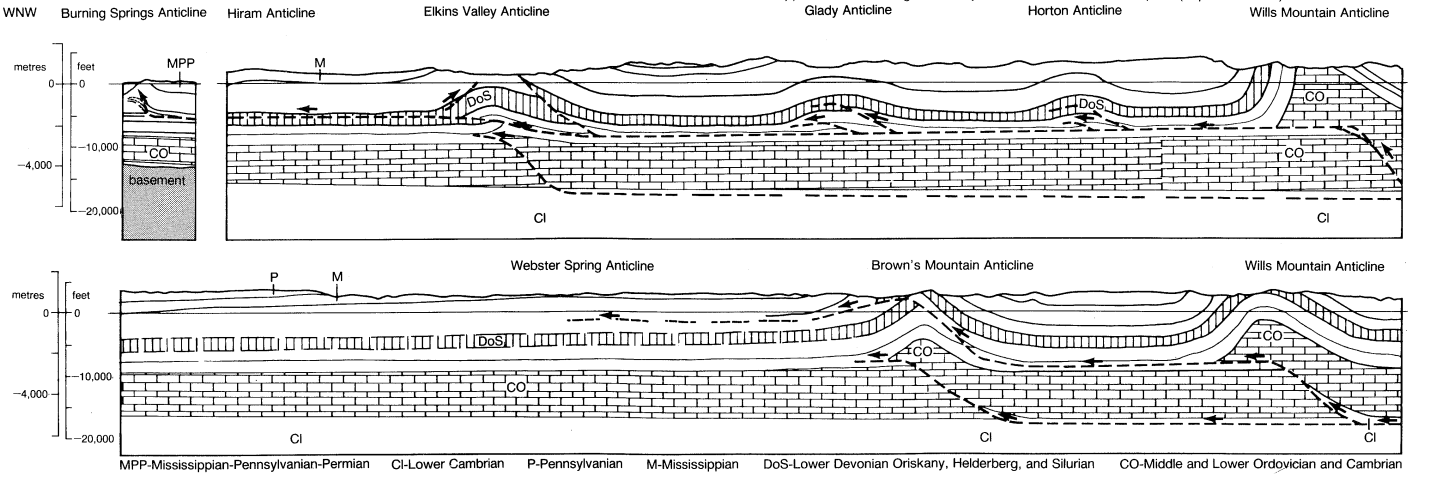



Figure 8: Cross sections of the Appalachian Plateau and northwestern Valley and Ridge provinces in Virginia and West Virginia. Thrust faulting at deeper levels causes folding at shallow depths.

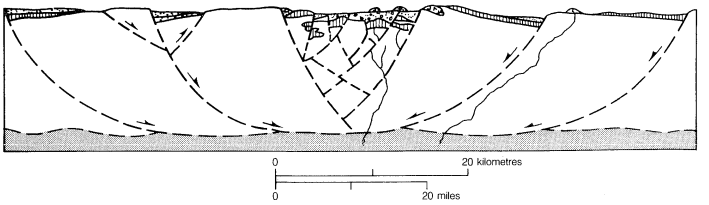



Figure 9: Idealized cross section of a tectonic valley showing the subsidence and rotation of blocks along curved faults.

Adapted from L.A. Wright and B.W. Troxell, "Shallow-Fault Interpretation of Basin and Range Structure, Southwestern Great Basin," *Gravity and Tectonics*, (1973); John Wiley & Sons, New York

ally dominated by parallel ridges of resistant rock divided by valleys of weaker rock, as in the eastern flank of the Canadian Rocky Mountains or in the Jura Mountains of France and Switzerland.

Most fold and thrust belts are bounded on one side, or lie parallel to, a belt or terrain of crystalline rocks—metamorphic and igneous rocks that in most cases solidified at depths of several kilometres or more and that are more resistant to erosion than the sedimentary rocks deposited on top of them. These crystalline terrains typically contain the highest peaks in any mountain belt and include the highest belt in the world, the Himalayas, which was formed by the thrusting of crystalline rocks up onto the surface of the Earth. The great heights exist because of the resistance of the rocks to erosion and because the rates of continuing uplift are the highest in these areas. The topography rarely is as regularly oriented as in fold and thrust belts.

Block-fault mountains

In certain areas, blocks or isolated masses of rock have been elevated relative to adjacent areas to form block-fault mountains or ranges. In some places, block-fault ranges with an overall common orientation coalesce to define a mountain belt or chain, but in others the ranges may be isolated.

Block faulting can occur when blocks are thrust, or pushed, over neighbouring valleys, as has occurred in the Rocky Mountains of Colorado, Wyoming, and Utah in the western United States, or as is now occurring in the Tien Shan, an east–west range in western China and Soviet Central Asia. Within individual ranges, which are usually a few hundred kilometres long and several tens of kilometres wide, crystalline rocks commonly crop out. On a large scale, there is a clear orientation of such ranges, but within them the landforms are controlled more by the whims of erosion than by tectonic processes.

Block faulting also occurs where blocks are pulled apart, causing a subsidence of the intervening valley between diverging blocks. In this case, alternating basins and ranges form. The basins eventually fill with sediment, and the ranges—typically tens of kilometres long and from a few to 20–30 kilometres wide—often tilt, with steep relief on one side and a gentle slope on the other. The uniformity of the gently tilted slope owes its existence to long periods of erosion and deposition before tilting, sometimes with a capping of resistant lava flows on this surface prior to tilting and faulting. Both the Tetons of Wyoming and the Sierra Nevada of California were formed by blocks being tilted up toward the east; major faults allowed the blocks on their east sides to drop steeply down several thousand metres and thereby created steep eastern slopes.

In some areas, a single block or a narrow zone of blocks has subsided between neighbouring blocks or plateaus that moved apart to form a rift valley between them. Mountains with steep inward slopes and gentle outward slopes often form on the margins of rift valleys. Less commonly, large areas that are pulled apart and subside leave between them an elevated block with steep slopes on both sides. An example of this kind of structure, called a horst, is the Ruwenzori in East Africa.

Residual mountains

Finally, in certain areas, including those that once were plateaus or broad uplifted regions, erosion has left what are known as residual mountains. Many such mountains are isolated and not part of any discernible chain, as, for instance, Mount Katahdin in Maine in the northeastern United States. Some entire chains (*e.g.*, the Appalachians in North America or the Urals in the Soviet Union), which were formed hundreds of millions of years ago, remain in spite of a long history of erosion. Most residual chains and individual mountains are characterized by low elevations; however, both gentle and precipitous relief can exist, depending on the degree of recent erosion.

## TECTONIC PROCESSES THAT CREATE AND DESTROY MOUNTAIN BELTS AND THEIR COMPONENTS

Mountains and mountain belts exist because tectonic processes have created and maintained high elevations in the face of erosion, which works to destroy them. The topography of a mountain belt depends not only on the processes that create the elevated terrain but also on the forces that support this terrain and on the types of processes (erosional or tectonic) that destroy it. In fact, it is necessary to understand the forces that support elevated terrains before considering the other factors involved.

**Mechanisms that support elevated terrains.** Two properties of rocks contribute to the support of mountains, mountain belts, and plateaus: strength and density. If rocks had no strength, mountains would simply flow away. At a subtler level, the strength of the material beneath mountains can affect the scale of the topography.

In terms of strength, the lithosphere, the thickness of which varies over the face of the Earth from a few to more than 200 kilometres, is much stronger than the underlying layer, the asthenosphere (see PLATE TECTONICS). The strength of the lithosphere is derived from its temperature; thick lithosphere exists because the outer part of the Earth is relatively cold. Cold, thick, and therefore strong lithosphere can support higher mountain ranges than can thin lithosphere, just as thick ice is better able to support larger people than thin ice.

In terms of chemical composition, and therefore density, the Earth's crust is lighter than the underlying mantle. Beneath the oceans, the typical thickness of the crust is only six to seven kilometres. Beneath the continental regions, the average thickness is about 35 kilometres, but it can reach 60 or 70 kilometres beneath high mountain ranges and plateaus. Thus, most ranges and plateaus are buoyed up by thick crustal roots. To some extent the light crust floats on the heavier mantle, as icebergs float on the oceans.

Differences between the crust and lithosphere

It should be noted that the crust and lithosphere are defined by different properties and do not constitute the same layer. Moreover, variations in their thicknesses have different relationships to the overlying topography. Some mountain ranges and plateaus are buoyed up by a thick crust. The lithosphere beneath such areas, however, can be thin, and its strength does not play a significant role in supporting the range or plateau. Other ranges may overlie thick lithospheric plates, which are flexed down by the weight of the mountains. The crust beneath such ranges is likely to be thicker than normal but not as thick as it would be if the lithosphere were thin. Thus, the strength of the lithosphere supports these mountains and maintains the base of the crust at a higher level than it would have if the strong layer were absent. For instance, the Himalayas have been thrust onto the crust of the Indian shield, which is underlain by particularly cold, thick lithosphere that has been flexed down by the weight of the high range. The thickness of the crust is about 55 kilometres beneath the high peaks, which stand more than 8,000 metres high. The thickest crustal segment of 70 kilometres, however, lies farther north beneath the Plateau of Tibet (or Tibetan Plateau), whose altitude is about 4,500 to 5,000 metres, but whose lithosphere is much thinner than that beneath the Himalayas. The strong Indian lithosphere helps to support the Himalayas, but the buoyancy of the thick Tibetan crust maintains the high elevation of the plateau.

**Tectonic processes that produce high elevations.** As noted above, individual mountains, mountain ranges, mountain belts, and plateaus exist because tectonic processes have elevated terrains faster than erosion could destroy them. High elevations are created by three major processes: volcanism, horizontal crustal shortening as manifested by folding and by faulting, and the heating and thermal expansion of large terrains.

*Volcanism.* Most, but not all, volcanoes consist of material that is thought to have melted in the mantle (at depths of tens of kilometres), which rose through the overlying crust and was erupted onto the surface. To a large extent, the physical characteristics of the erupted material determines the shape and height of a volcano. Material of low density can produce taller mountains than can denser material. Lavas with low viscosity, such as in Hawaii, flow easily and produce gentle slopes, but more viscous lavas mixed with explosively erupted solid blocks of rocks can form steeper volcanic cones, such as Mount Fuji in Japan, Mount Rainier in the northwestern United States, or Mount Kilimanjaro in Africa.

Many volcanoes are built on elevated terrains that owe their existence to the intrusion into the crust of magmas—*i.e.,* molten rock presumably derived from the mantle. The extent to which this process is a major one in mountain belts is controversial. Many belts, such as the Andes, seem to be underlain, at least in part, by solidified magmas, but the volume of the intruded material and its exact source (melting of either the crust or the mantle) remain poorly understood.

*Crustal shortening.* In most mountain belts, terrains have been elevated as a result of crustal shortening by the thrusting of one block or slice of crust over another and/or by the folding of layers of rock. The topography of mountain ranges and mountain belts depends in part on the amount of displacement on such faults, on the angles at which faults dip, on the degree to which crustal shortening occurs by faulting or by folding, and on the types of rocks that are deformed and exposed to erosion. Most of the differences among mountain belts can be ascribed to some combination of these factors.

*Heating and thermal expansion.* Rocks, like most materials, expand when they are heated. Some mountain ranges and plateaus are high simply because the crust and upper mantle beneath them are unusually hot. Most broad variations in the topography of the ocean floor, the mid-ocean ridges and rises, are due to horizontal variations in temperature in the outer 100 kilometres of the Earth. Hot areas stand higher—or at shallower depths in the ocean—than cold areas. Many plateaus, such as the Massif Central in south central France or the Ethiopian Plateau, are elevated significantly because the material beneath them has been heated.

**Tectonic processes that destroy elevated terrains.** Besides erosion, which is the principal agent that destroys mountain belts, two tectonic processes help to reduce high elevations. Horizontal crustal extension and associated crustal thinning can reduce and eliminate crustal roots. When this happens, mountain belts widen and their mean elevation diminishes. Similarly, the cooling and associated thermal contraction of the outer part of the Earth leads to a reduction of the average height of a mountain belt.

Crustal thinning and thermal contraction

## MAJOR TYPES OF MOUNTAIN BELTS

Mountain belts differ from one another in various respects, but they also have a number of similarities that enable Earth scientists to group them into certain distinct categories. Each of these categories is characterized by the principal process that created a representative belt. Moreover, within individual belts different tectonic processes can prevail and can be associated with quite different landforms and topography. Thus, for any category there are exceptions and special cases, as well as subdivisions.

**Mountain belts associated with volcanism.** Volcanoes typically form in any of three tectonic settings. At the axes of the mid-ocean ridge system where lithospheric plates diverge, volcanism is common; yet, high-standing volcanoes (above sea level) rarely develop. At subduction zones where one plate of oceanic lithosphere plunges beneath another plate, long linear or arcuate chains of volcanoes and mountain belts associated with them are the norm. Volcanoes and associated landforms, as well as linear volcanic chains and ridges (*e.g.,* the Hawaiian chain) also can exist far from plate boundaries.

*Mid-ocean ridges and rises.* Where two lithospheric plates diverge, new material is intruded into the gap between the plates and accreted to each of them as they diverge. The vast majority of volcanic rocks ejected onto the surface of the Earth is erupted at the mid-ocean ridges and rises where this process occurs. Thus, such submarine landforms comprise very long, narrow volcanic centres. Although volcanoes do form as isolated seamounts along the axes of mid-ocean ridges, they constitute only a small fraction of the erupted material. Moreover, areas along the ridges and rises where volcanism is particularly abundant are considered unusual; the excess amount of volcanic activity is generally attributed to "hot spots" in the mantle (see below). Finally, most of the relief that defines the mid-ocean ridges and rises is not due to volcanism at all but rather to thermal expansion, as will be explained below.

Submarine landforms

*Volcanic structures along subduction zones.* Linear or arcuate belts of volcanoes are commonly associated with subduction zones. Volcanoes typically lie 150 to 200 kilometres landward of deep-sea trenches, such as those that border much of the Pacific Basin. The volcanoes overlie a zone of intense earthquake activity that begins at a shallow depth near such a trench and that dips beneath the volcanoes. They often form islands and define island arcs—arcuate chains of islands such as the Aleutians or the Lesser Antilles (see OCEANS: *Major geologic and geographic features*). Volcanoes usually are spaced a few to several tens of kilometres apart, and single volcanoes commonly define the width of such belts. Elsewhere, as in Japan, in the Cascade chain of the northwestern United States and southwestern Canada, or along much of the Andes, volcanoes have erupted on the margin of a continent. Nearly all features typical of an island arc, including the narrow belt of volcanoes, deep-sea trench, and intense earthquake activity, can be found at such continental margins.

The landscape of island chains of this kind is characteristically dominated by steep volcanic cones topped by small craters, and the relief between these volcanoes is low. A few such volcanoes have undergone massive eruptions and have expelled a large fraction of their interiors, as did Mount St. Helens in the northwestern United States in 1980. In the most intense eruptions of this sort, the remnants of the volcano collapse into the void at its centre, sometimes leaving a caldera (a very large crater with relatively low rims). Examples of such structures include those formed by Krakatoa in Indonesia in 1883 and by Thera (also called Santorin or Santoríni) in the Aegean Sea a few thousand years ago.

The lavas erupted at these volcanoes are thought to be derived from the mantle in the wedge of asthenosphere above the lithospheric plate plunging into it. Water carried down in the interstices of the subducted rock and by hydrous minerals to which water is loosely bound chemically is expelled into the wedge of asthenosphere above the subduction zone. The introduction of water reduces the melting temperature of the rocks and allows material in the wedge to melt and rise to the surface.

*Landforms associated with hot spot volcanism.* Some volcanic phenomena occur at large distances from plate boundaries (for example, on the Hawaiian Islands or at Yellowstone National Park in the western continental United States). Also, as noted above, volcanism is especially intense at some parts of the mid-ocean ridge system (as in Iceland or the Galápagos Islands in the eastern Pacific). Magmas erupted in these settings originate in the asthenosphere, perhaps at depths of several hundred kilometres or more at what are called hot spots in the mantle. Such sources of melting may be due to chemical differences rather than to heat (see VOLCANISM: *Intraplate volcanism*). Active volcanoes are usually localized in a region with dimensions of 100 to 200 kilometres or less.

A chain of extinct volcanoes or volcanic islands (and seamounts), like the Hawaiian chain, or a volcanic ridge, like Walvis Ridge between the islands of Tristan da Cunha and the east coast of Africa, can form where a lithospheric plate moves over a hot spot. The active volcanoes all lie at one end of the chain or ridge, and the ages of the islands or the ridge increase with their distance from those sites of volcanic activity. Older volcanoes are more eroded than younger ones and are often marked only by coral reefs that grow on the eroded and subsiding volcanic island.

Volcanic chains of this kind are not common in conti-

nental regions, in part because most continental masses move slowly over hot spots. Volcanic activity, however, can be particularly abundant when a plate moves so slowly with respect to a hot spot. Moreover, a long duration of volcanism often results in a warming of the lithosphere. This warming causes a localized thermal expansion and consequently a localized upwarping or doming of the Earth's surface, as in the case of the Yellowstone area or the Massif Central in France. The resulting domes cover areas a few to several hundred kilometres in extent, and the mean elevations are rarely as much as 1,000 metres higher than the surrounding regions. Thus, except for the isolated volcanoes that lie on the upwarps, relief is gentle and due largely to erosion.

Formation of domes

Some hot spots are associated with massive eruptions of lava and ash, primarily of basaltic composition, which cover vast areas as extensive as tens or hundreds of square kilometres. Such flood basalts, or traps, buried the Snake River Plain west of Yellowstone a few million years ago, the Columbia River Valley some 20,000,000 years ago, and central India (the Deccan traps) 60,000,000 to 65,000,000 years ago. Flood basalts create a remarkably flat surface that is later dissected into a network of sharply incised valleys (see below *Plateaus*).

Most volcanoes that cannot be ascribed either to a subduction zone or to seafloor spreading at mid-ocean ridges are attributed to hot spots. There are, however, some volcanoes, volcanic fields, and flood basalts that cannot yet be ascribed to hot spots with any certainty. Nevertheless, the landforms associated with such volcanic phenomena resemble those in other settings for which a simple cause can be offered.

*Mountain belts associated with crustal shortening.* Most mountain belts of the world and nearly all of those in Europe, Asia, and North America have been built by horizontal crustal shortening and associated crustal thickening. The landforms associated with such belts depend on the rates, amounts, and types of crustal deformation that occur and on the types of rocks that are exposed to erosion. To some extent the deformation can be related to different tectonic settings. Large thrusted crystalline terrains and parallel fold and thrust belts are commonly associated with continental collisions in which two separate continents have approached each other and one has been thrust onto the other. Continental collisions are responsible for Alpine-, or Himalayan-, type mountain belts. Fold and thrust belts can also be associated with active continental margins or Andean-type margins, where oceanic lithosphere is subducted into the asthenosphere but where crustal shortening occurs landward of the volcanic arc on the overriding continental plate. Block-faulted ranges commonly form as intracontinental mountain ranges or belts, far from collision zones and subduction zones.

*Alpine- (or Himalayan-)type belts.* These belts are thought to have been created by the movement of one continent beneath another. In general, a thick layer of light, buoyant continental crust cannot be carried deep into the asthenosphere. Instead, the leading edge of the descending continent is scraped off, and the rest of the continent then plunges beneath the off-scraped slice. Eventually the convergence between the two plates carrying the continents comes to a halt, but usually not before several slices of continental material have been removed from the underthrusting continent and stacked on top of it.

The sedimentary rocks deposited on the continental crust and its margin long before the collision often constitute one or part of one of the off-scraped slices. They commonly are deformed into a fold and thrust belt as the basement under them continues to plunge beneath the overriding plate at the subduction zone. Layers of strong sedimentary rock detach from the underlying basement at weak layers that commonly consist of evaporites (salt, gypsum, or anhydrite) or of shale by a process called décollement (from the French word meaning "unglueing"). The stronger layers of sedimentary rock are then folded into linear, regularly spaced folds—alternating anticlines and synclines—and thrust on top of one another. The Valley and Ridge province of Pennsylvania, which was formed during the collision of Africa and North America near the end of Paleozoic time (about 240,000,000 years ago), is a classic example.

Deformation into fold and thrust belts

Convergence between two lithospheric plates can be rapid in such settings—10 to 100 millimetres per year—and the amount of displacement on the major thrust faults also can be large—tens to more than 100 kilometres. Thus, when a slice of crystalline rock from deep in the crust is scraped off the remainder of the continent and is underthrust by it, much of the slice is uplifted and pushed onto the relatively flat, ancient surface of the intact portion of the continent. Erosion generally removes the sedimentary cover of such slices and leaves expanses of crystalline rocks, as can be seen on Himalayan or Alpine peaks.

Faults along which a slice of continental crust is torn from the rest of the continent and thrust onto it are called ramp overthrusts. When the fault first forms, it dips at 10° to 30° (or more). Slip on this fault (*i.e.*, the movement of one face of the fault relative to the other) brings the leading edge of the off-scraped slice of crust to the surface of the Earth, where it then slides along the surface. The intact continent is flexed down by the weight of the material thrust on top of it. As a consequence, its initially flat surface dips at a very gentle angle of only a few degrees. Accordingly, a ramp overthrust consists of two segments. The first segment, the ramp, dips relatively steeply; slip on it causes uplift of the overriding slice and of the crystalline rocks from deep in the crust to create high relief and the high range. The other segment, which was once the top surface of the continent, has been flexed down and dips at a gentle angle. Slip on it allows the overthrust slice to advance over the rest of the continent, where it plows the sedimentary layers in front of it into folds and smaller overthrusts.

Ramp overthrusts

When a major ramp overthrust is active and the intact continent is flexed down in front of the overriding mountain range, a foreland basin is formed by the flexure (see below *Tectonic basins and rift valleys;* also Figure 10). Foreland basins usually exist as subsurface features that have been filled with debris eroded from the advancing overthrust slice of crust. These deposits, called molasse, can in turn be folded and thrust over one another shortly after they are deposited. Fold and thrust belts in such material, as found at the northern edge of the Alps or at the foot of most of the Himalayas, are often narrow, composed of only one or two parallel folds and faults. The topography associated with them generally consists of low, elongated hills of poorly consolidated sedimentary rock that is easily and rapidly eroded.

Collision zones are thus commonly identified by narrow belts of elevated crystalline terrain and parallel fold and thrust belts. The crystalline terrain has been thrust upward and toward the fold and thrust belt. Deformation is generally confined to shallow depths of only a few kilometres at such belts but penetrates deeply into the Earth beneath the crystalline terrains. The rapid uplift of these resistant rocks creates a high range. A crystalline terrain often exhibits large folds in which the rocks appear to have flowed instead of having been bent. Folds of this sort have formed at depths where the rocks were hot and soft before they reached the relatively cold surface of the Earth. The overthrusting of crystalline terrains onto intact continental crust can occur at rates of tens of millimetres per year, which is rapid for rates of slip on faults, and the crystalline rocks can be uplifted 10 to 20 kilometres by slip on ramp overthrusts.

*Andean-type belts.* At some continental margins, oceanic lithosphere is subducted. At some of these sites, the landscape is dominated by volcanoes, such as along the Cascades of western North America or in Japan, but at others, such as along much of the Andes of South America, volcanoes constitute only a small or even negligible part of the relief. At Andean-type margins, the crust is typically thicker than normal, and high mountains can exist even in the absence of volcanoes. Some of the thickened crust is due to the intrusion of magma from the mantle, and some to crustal shortening.

Mountains along active continental margins

Oceanic lithosphere is commonly subducted at active continental margins at rates of tens to more than 100 millimetres per year, but crustal shortening within the

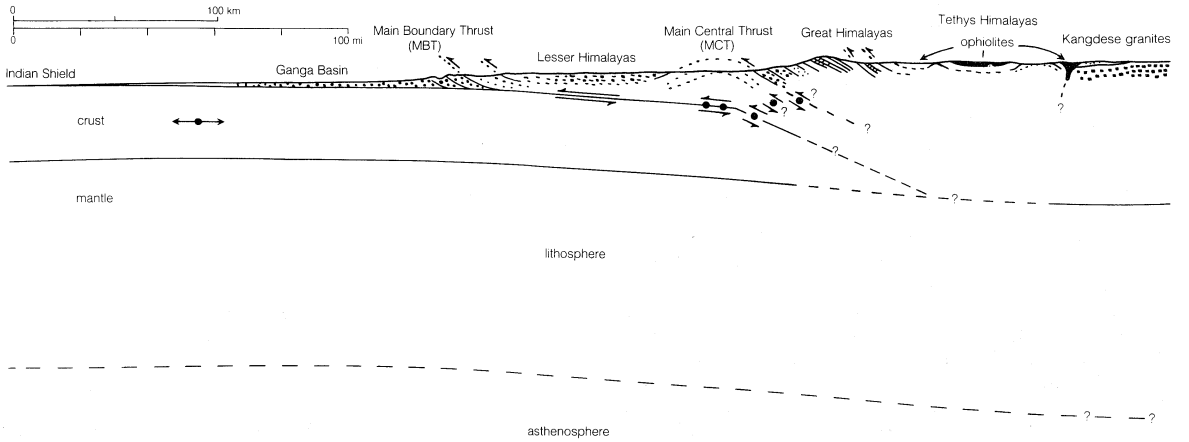



Figure 10: Simplified north–south cross section of the Himalayas, revealing a foreland basin (Ganga Basin), an overthrusting of crystalline terrains onto the Indian Plate, and a steeper thrust fault (a ramp) beneath the Great Himalayas.

Adapted from P. Molnar and W-P. Chen, "Seismicity and Mountain Building," *Mountain Building Processes*, p. 44 (1982); Academic Press, New York

overriding plate typically occurs at rates of only a few millimetres annually. As at continent-continent collision zones, the crustal shortening occurs both by overthrusting of crystalline terrain onto intact continental crust, which in this case lies landward of the volcanic belt, and by the formation of a fold and thrust belt within sedimentary rock lying on the intact continent. The thrusting of crystalline terrain is probably facilitated by a heating and consequent weakening of the rocks near the volcanoes. The presence or absence of a parallel fold and thrust belt depends in part on the presence or absence of thick sedimentary rocks within which detachment of separate layers can take place.

Notwithstanding large variations in topography and in the style of deformation among Andean belts in general, the scales of deformation and uplift are less than those at collision zones. Overthrust crystalline terrains are smaller, and the crystalline rocks themselves have not been thrust up from depths as great as those at collision zones. Much of the Andes, for instance, consists of sedimentary rock that never was buried deeper than a few kilometres and therefore has not been metamorphosed (heated to high temperature or put under high pressure) or at most only has been mildly metamorphosed. Topography in the high parts of the Andes is typically much gentler than in the Himalayas. The most impressive relief is on the eastern flank of the Andes where rivers responding to a wet climate have cut deep canyons.

Fold and thrust belts can be very well developed at Andean margins. The eastern Cordillera of the Bolivian Andes is an extremely wide fold and thrust belt, but only along the eastern third of the cordillera do simple parallel folds control the topography. Farther west, both the greater role of thrust faulting in the evolution of the cordillera and the longer duration of erosion have diminished the role of folding. Except where rivers have cut deep canyons, relief is not exceptionally great. Similarly while oceanic lithosphere was underthrust beneath the west coast of Canada during the Mesozoic Era (248,000,000–65,000,000 years ago), the Canadian shield was underthrust more than 200 kilometres beneath the Canadian Rocky Mountains, with crustal shortening occurring by décollement and by folding and thrust faulting within the sedimentary cover (Figure 11).

Thus Andean-type belts have a narrow belt of volcanoes and often a fold and thrust belt on their landward margin. The volcanoes of some belts are built on a high range that is more of a long, narrow plateau than a mountain range, for relief on it is not necessarily great.

*Intracontinental mountain belts.* In some regions, mountain belts have been formed by crustal shortening within a continental mass, rather than where two continents have collided. Some 40,000,000 to 80,000,000 years ago, the Rocky Mountains of Colorado, Utah, and Wyoming formed in this way, and today both the Tien Shan and the Atlas Mountains of northwestern Africa are actively forming within a continent. In general, intracontinental mountain belts are characterized by block faulting. Blocks, tens of kilometres wide and hundreds of kilometres long, are uplifted along faults that dip beneath them at angles of 25° to 45°. Because of the displacement on steep faults, crystalline rocks commonly crop out in the mountains. The edges of the ranges can be sharply defined. Fold and thrust belts are not common and are usually narrow where present.

At the edges of such ranges, sedimentary rocks are commonly tilted up, and, where resistant, they can form narrow, sharp-crested ridges called hogbacks that are parallel to the front of the ranges. A particularly prominent hogback lies along the east edge of the Front Range in eastern Colorado.

Intracontinental belts generally consist of elongated block-faulted ranges, which in some cases overlap but are not necessarily parallel to one another. Thus, in parts of the Tien Shan, two or three nearly parallel, sharply bounded ranges are separated from one another by parallel basins that are 10 to 30 kilometres wide. The ranges of this great mountain system are being overthrust onto the basins, and one such basin, the Turfan Depression, has dropped below sea level (see below *Tectonic basins and rift valleys*). In contrast with the parallel ranges in the Tien Shan, the northwest-trending Wind River Range in Wyoming,

Adapted from R.A. Price, "The Cordilleran Foreland Thrust and Fold Belt in the Southern Canadian Rocky Mountains," *Thrust and Nappe Tectonics* (1981); The Geological Society of London

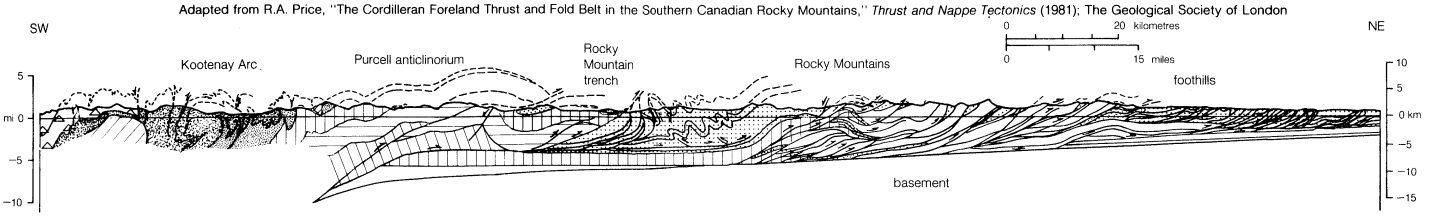



Figure 11: Southwest–northeast cross section of the Canadian Rocky Mountains showing the uplift of crystalline terrains onto the downflexed Canadian shield and the fold and thrust belt on the east side of the range.

the east–west trending Uinta Mountains in Utah, and the north–south trending Front Range in Colorado are all part of the same intracontinental belt, the Rocky Mountains.

## MAJOR MOUNTAIN BELTS OF THE WORLD

Most mountains and mountain ranges are parts of mountain belts that have formed where two lithospheric plates have converged and where, in most cases, they continue to converge. In effect, many mountain belts mark the boundaries of lithospheric plates, and these boundaries in turn intersect other such boundaries. Consequently, there exist very long mountain systems where a series of convergent plate boundaries continue from one to the next. A nearly

Principal mountain systems

continuous chain of volcanoes and mountain ranges surrounds most of the Pacific Basin (the so-called Circum-Pacific System). A second nearly continuous chain of mountains can be traced from Morocco in North Africa through Europe, then across Turkey and Iran through the Himalayas to Southeast Asia; this chain, the Alpine-Himalayan (or Tethyan) System, has formed where the African, Arabian, and Indian plates have collided with the Eurasian Plate. Nearly all mountain ranges on the Earth can be included in one of these two major systems (Figure 12) and most that cannot are residual mountains, which originated from ancient continental collisions that occurred hundreds of millions of years ago.

**The Circum-Pacific System.** A nearly continuous chain of volcanoes surrounds the Pacific Ocean. The chain passes along the west coast of North and South America, from the Aleutian Islands to the south of Japan, and from Indonesia to the Tonga Islands, and to New Zealand. The Pacific Basin is underlain by separate lithospheric plates that diverge from one another and that are being subducted beneath the margins of the basin at different rates. This Circum-Pacific chain of volcanoes (often called the Ring of Fire) and the mountain ranges associated with it owe their formation to the repeated subduction of oceanic lithosphere beneath the continents and the islands that surround the Pacific Ocean. Differences among the various segments of the Circum-Pacific chain arise from differences in the histories of subduction of the different plates.

*The Andes.* The Nazca Plate, which underlies most of the southeastern Pacific, is being subducted beneath most of the west coast of South America at a rapid rate of 80 to 100 millimetres per year. A nearly continuous chain of volcanoes lines the margin of South America, and the world's tallest volcano, Ojos del Salado (6,893 metres), is one of these peaks. The Andean range, however, is more than just a chain of volcanoes, and its highest peak, Mount Aconcagua (6,959 metres), the tallest outside Asia, is not volcanic. Crustal shortening and crustal thickening occur all along the eastern margin of the Andes by the westward underthrusting of the stable areas of Brazil and Argentina beneath the Andes at a rate of a few millimetres per year.

The southern part of the Andes in southern Chile and Argentina consists of a narrow range only 100 to 200 kilometres wide. A chain of volcanoes follows the axis of the range, but crustal thickening due to crustal shortening is a principal cause of the high range, and many of the volcanoes are built on folded and faulted sedimentary rock.

From northern Argentina to northern Peru and Ecuador, the Andes are much wider, with the widest segment across southern Bolivia. There, the mountain belt consists of two parallel ranges, the Cordillera Occidental (or Western Cordillera) and the Cordillera Oriental (or Eastern Cordillera), which surround the high plateau, the Altiplano.

Cordillera Occidental and Cordillera Oriental

The volcanic chain has been constructed on thick crust and forms the Cordillera Occidental. The Brazilian shield has been underthrust beneath the Cordillera Oriental, which comprises the western edge of a wide fold and thrust belt. This fold and thrust belt is marked by north–south trending folds and north–south trending ridges and valleys in northern Argentina and southeastern Bolivia. North of the latitude where the west coast of South America bends, the trend of the Andes, including that of both cordilleras,

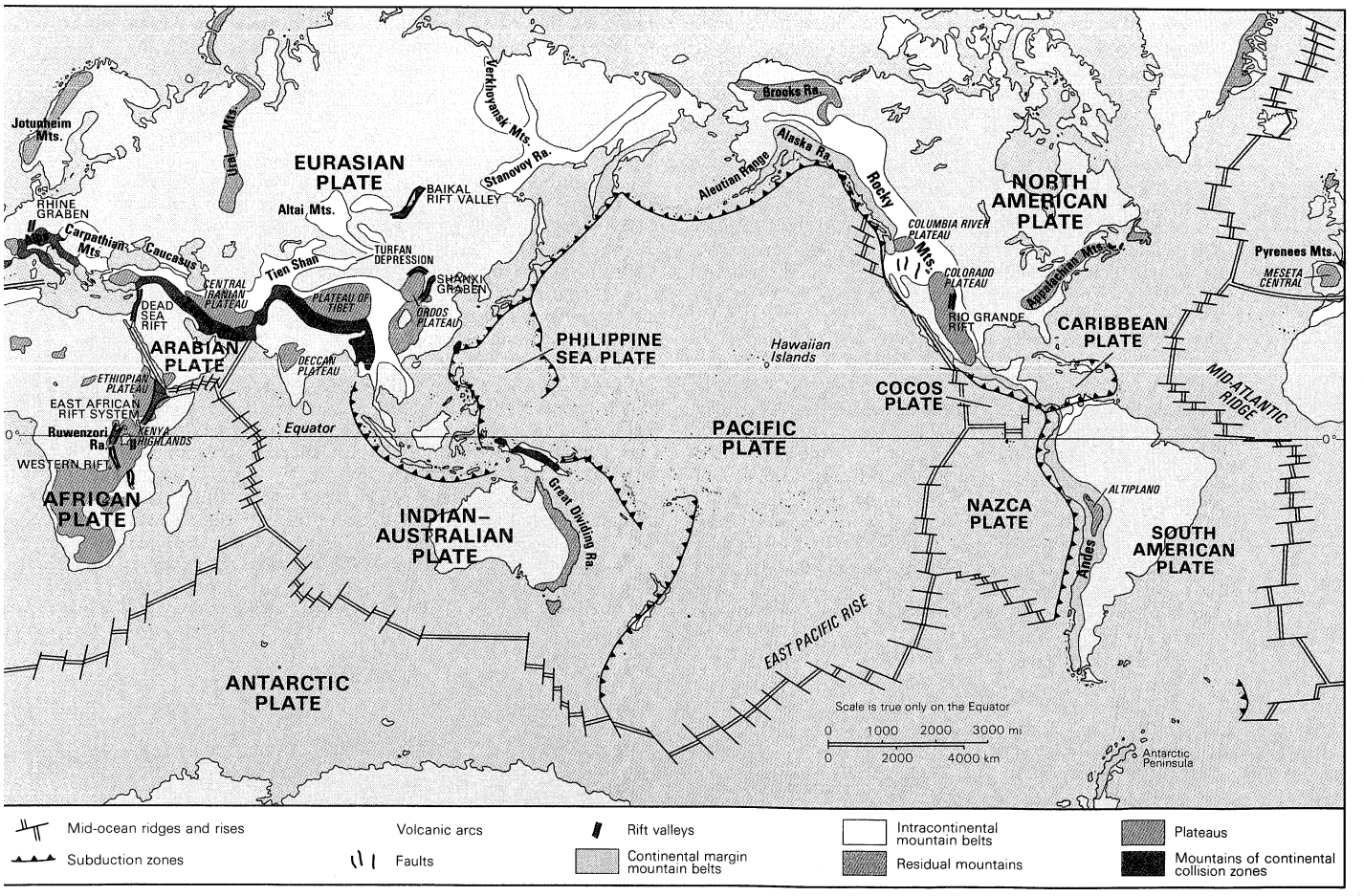



Figure 12: Tectonic map of the world.

is northwesterly parallel to the coast of Peru. The fold and thrust belt east of the Cordillera Oriental is narrower than that farther south but is well defined by a few northwesterly trending ridges and valleys.

Lying between the two cordilleras in northern Argentina, western Bolivia, and southern Peru, the Altiplano stands at an average height of about 3,800 metres. Within it lies Lake Titicaca, the highest navigable lake in the world. The Altiplano is a high arid basin that captures sediment eroded from the eastern and western cordilleras bounding it (see below *Plateaus*). Older rocks that crop out within it have been folded; thus crustal shortening probably has been an important factor in creating the high elevations and the thick crust that underlies this plateau.

In Colombia, the Andean chain diverges into three separate chains, each about 100 kilometres wide. Volcanoes occur in the westernmost chain, but all three have undergone crustal shortening. For example, the easternmost of the three, which continues into Venezuela as the "Venezuelan Andes," is being underthrust from the northwest by the Maracaibo Basin and from the southeast by the Guiana Shield underlying southeastern Venezuela. Thus the Venezuelan Andes are an intracontinental mountain belt.

The divergence of the Andes into three chains in Colombia extends northward. The western chain continues into Panama and through Central America. The central chain continues toward the Caribbean. The Venezuelan Andes intersect an east–west trending chain along the north coast of South America in Venezuela.

*The Caribbean chains.* The mountain range along the coast of Venezuela is a remnant of a phase when the Caribbean Sea was subducted southward beneath Venezuela and where rocks were folded along east–west axes. Right-lateral strike-slip faulting and rather slow mountain building occur there today, as much by slight vertical displacement on predominantly strike-slip faults as by slow obliquely oriented folding and thrust faulting and associated crustal shortening.

At the eastern end of the Caribbean Sea, the Lesser Antilles—volcanic islands that form a typical island arc—mark a zone where a part of the floor of the North Atlantic Ocean underthrusts that of the Caribbean Sea—namely, the Caribbean Plate. This plate has moved east relative to both North and South America at a rate of 10 to 20 millimetres per year for tens of millions of years. This displacement and the consequent overthrusting of the seafloor to the east are responsible for the volcanic arc that constitutes the Lesser Antilles as well as for the strike-slip displacement occurring in Venezuela.

Mountainous islands of the Caribbean

Most of the major islands that define the northern margin of the Caribbean—Puerto Rico, Hispaniola, Cuba, and Jamaica—are mountainous, and these mountainous terrains, like that in northern Venezuela, are remnants of the period of convergence between North and South America and also of complicated deformation along the ancient margins of the Caribbean Plate. At present, crustal shortening occurs at only a very slow rate, if at all, on these islands.

At the western margin of the Caribbean Plate another small plate, the Cocos Plate, is being underthrust beneath Mexico and Central America. A belt of volcanoes extends from northern Panama to western Mexico, and virtually all of the highest mountains in this belt are volcanic. These volcanoes are built on thickened crust, and crustal shortening has occurred within the Central American Cordillera, but the principal tectonic process that has affected the landscape is volcanism.

*The North American Cordillera.* A broad mountain belt extends north from Mexico to Alaska, and it reflects both a more diverse and a more complicated history of convergence between lithospheric plates than that presently occurring at the Andes or Central America.

Subduction of oceanic lithosphere presently occurs only beneath two segments of the coast of western North America. The subduction of a small plate, the Juan de Fuca Plate, beneath northern California, Oregon, Washington, and southern British Columbia is responsible for the Cascade chain of volcanoes, which includes Mount St. Helens. The very large Pacific Plate, which underlies most of the northern and western Pacific Basin, moves north-northwest with respect to North America nearly parallel to the coast, and is subducted beneath southern Alaska and the Aleutians. The volcanic chain that forms the Aleutian Islands and continues into the Alaskan peninsula to the Wrangell Mountains is a consequence of this convergence and subduction.

Cascade chain of volcanoes

Most of the North American Cordillera was built during Mesozoic and Early Cenozoic times (between about 170,000,000 and 40,000,000 years ago) when lithospheric plates converged with North America at rapid rates of many tens to more than 100 millimetres per year. The Juan de Fuca Plate is the last remnant of one of these plates. The others have been subducted beneath western North America and have completely disappeared. Thus, in Mesozoic and Early Cenozoic times, an Andean margin similar to that which presently bounds the west coast of South America bounded western North America.

The Coast Ranges of central and northern California, Oregon, and Washington consist of folded and faulted slices of oceanic crust and its overlying sedimentary rocks. Much of the rock that constitutes these mountains was scraped off the oceanic lithosphere at the trench just west of the continent. The Olympic Mountains in northwestern Washington, for instance, consist largely of off-scraped seamounts. The rock of such coastal mountains was intensely deformed and metamorphosed before being elevated to produce the present range. Specifically, the hard basalt that makes up much of the oceanic crust has been metamorphosed into the easily deformed rock serpentinite, which contains the weak, fibrous mineral serpentine. The gentle relief of the Coast Ranges is due in part to the weakness of serpentinite, a characteristic that gives rise to frequent landslides and rapid erosion.

Granite belt of western North America

A belt of granite lies inland and forms a mountainous zone from the axis of Baja California (in Mexico), through southern California, along the Sierra Nevada in the states of California and Nevada, northwestward into Idaho, and then north-northwestward along the western margin of the Canadian Rocky Mountains to Alaska. This granite belt underlay the volcanoes that marked the subduction zone in Mesozoic and Early Cenozoic times. The intrusion of this granite was most intense between 170,000,000 and 70,000,000 years ago during the Mesozoic Era. The Sierra Nevada of California, which contains Mount Whitney, the highest peak in the contiguous United States, is composed almost entirely of this granite.

While subduction of oceanic lithosphere occurred beneath western North America, a major fold and thrust belt developed east of the granitic belt. During Mesozoic time, the Precambrian basement of Canada and North America was underthrust westward at least 200 kilometres beneath the Andean margin, and the sedimentary rocks covering it were folded and thrust onto one another. Although present in the western United States, this fold and thrust belt is most clearly revealed in the Canadian Rockies along the border between the provinces of Alberta and British Columbia, particularly in Banff and Jasper national parks.

In sum, throughout the latter half of the Mesozoic from about 170,000,000 to 65,000,000 years ago, the topography of western North America probably resembled that of western South America: a trench lay offshore; a belt of volcanoes underlain by granitic intrusions marked the western edge of a high range of mountains; and a fold and thrust belt lay east of the range. The tectonic history of western North America is more complicated, however, because during this period fragments of both continents and sub-oceanic plateaus were carried to the subduction zone and collided with North America. Most of the rock now found in westernmost Canada and Alaska consists of separate terrains of rock that were independently accreted to North America and that were subsequently deformed when the next such terrain collided with it. Moreover, tectonic processes occurring during the Cenozoic (since 65,000,000 years ago) have been different from those that occurred earlier and have severely modified the landscape.

Beginning about 70,000,000 to 80,000,000 years ago, the locus of crustal shortening in the United States shifted

from the fold and thrust belt, whose remnants now lie along the borders of western Utah and eastern Nevada and of western Wyoming and eastern Idaho, to eastern Utah, Colorado, and central Wyoming. Between about 70,000,000 and 40,000,000 years ago, thrust faulting on the margins of the Front Range in Colorado, the Laramie Mountains and the Wind River Range in Wyoming, and the Uinta Mountains in Utah, among others, allowed the uplift of blocks of Precambrian rock that are now exposed in the cores of these ranges. Together, these intracontinental ranges of block-faulted mountains form most of the Rocky Mountains of the United States.

The Rocky Mountains of the United States

During roughly the same period, volcanic rocks were erupted and deposited in parts of the Rockies, such as in southwestern Colorado in what are now the San Juan Mountains. The area that now forms the Colorado Plateau, in southern Utah and northern Arizona, underwent only very mild deformation in the form of small faults and folds and apparently lay at relatively low elevation. Sediment derived from the fold and thrust belt to its west and from the Rockies to its north and east was deposited on this relatively stable area. Thus, some 40,000,000 years ago, a high range of mountains lay along the western margin of North America. This range consisted of a volcanic chain along most of its western edge and an eroded fold and thrust belt on its eastern edge. At the latitude of Wyoming, Colorado, and Utah, another belt of mountains, the present-day Rocky Mountains, lay farther east.

The topography of the western United States has been modified extensively by tectonic processes during the last 20,000,000 years. Much of the mountainous terrain of Utah, Nevada, and California underwent large-scale crustal extension, beginning more than 40,000,000 years ago but accelerating about 15,000,000 years ago. The crustal extension approximately doubled the surface area of the region between central Utah and the Sierra Nevada, presumably with a reduction in the mean elevation of the mountains.

The present topography of the Basin and Range Province of North America is a direct manifestation of this crustal extension (see below *Tectonic basins and rift valleys*). The most prominent basins, such as Death Valley and Owens Valley in California, are small rift valleys that were formed during the last few million years. This phase of the crustal extension continues even today, with such basins becoming deeper and the surrounding ranges increasing in height. This condition is readily discernible in the case of Owens Valley and the Sierra Nevada. The occurrence of a major fault on the east side of the Sierras has allowed the valley to drop with respect to the mountain range, which has been tilted up toward the east.

Concurrent with this extension, the uppermost mantle beneath parts of the western United States has become hotter. The considerable height of the Colorado Plateau, for instance, appears to be the result of the warming of the underlying mantle during roughly the past 10,000,000 years. Such mantle heating also seems to have been responsible, at least in part, for the present elevation of much of the North American Cordillera.

The one area where rapid subduction of oceanic lithosphere (more than 50 millimetres per year) has continued is southern Alaska, where the Pacific Plate is being underthrust beneath the coast. The St. Elias Mountains, the tallest in southeastern Alaska and the Yukon, appear to be the direct consequences of this convergence and rapid underthrusting. Deformation of the southern Alaskan crust extends northward several hundred kilometres to the Alaska Range, where the highest mountain in North America, Mount McKinley, is found.

North–south crustal shortening in southern Alaska occurs both by thrust faulting and by strike-slip faulting on nearly vertical, northwesterly trending planes. Mount McKinley lies adjacent to one such major strike-slip fault, the Denali Fault. The rocks that make up Mount McKinley have been displaced several tens of kilometres northwestward relative to the rocks north of the Denali Fault and a few kilometres upward. This small vertical component, compared with the large horizontal component, has created the high peak.

*Volcanoes and island arcs surrounding the Northwest Pacific Basin.* A chain of volcanoes extends from mainland Alaska down the Alaska Peninsula along the Aleutian Islands and then southwestward down the peninsula of Kamchatka in northeastern Siberia and along the Kuril Islands to Japan. The Pacific Plate is being subducted beneath this long volcanic chain. Most of the relief is the result of volcanism. The Aleutians and Kurils are volcanic islands, and for the most part the volcanoes on the continental areas of the Alaska Peninsula, Kamchatka, and Japan are built up from sea level rather than on high ranges, as is the case with the Andes. For instance, Mount Fuji, a symmetrically shaped volcanic cone, rises from a low elevation to more than 4,000 metres.

Subduction of the Pacific Plate

In the central part of the Japanese island of Honshu, the Circum-Pacific System diverges into two chains. One continues southward along the Izu, Bonin, and Mariana islands. These volcanic islands form island arcs where the Pacific Plate is subducted beneath the floor of the Philippine Sea to the west. Southwest of Honshu, the Ryukyu Islands are another island arc where the Philippine Sea floor is subducted beneath the Yellow Sea.

The Ryukyu island arc ends abruptly at the island of Taiwan, which is not part of the Ryukyu arc. Taiwan is a small mountainous island consisting of folded and thrusted sedimentary rocks on the southeastern margin of the Asian continent. The sedimentary rocks of Taiwan were deposited on that margin under tranquil conditions, much as sedimentary rocks have been deposited on the margins of the Atlantic Ocean. Then, in the last few million years a segment of the Asian continental margin encountered a subduction zone that dipped east-southeast. As that short segment of the margin began to be underthrust, the sedimentary rocks were scraped off its leading edge and thrust back on top of it. Thus, not only the mountains of Taiwan but also virtually the entire island consists of folded and thrust sedimentary rocks that have rapidly piled up on what had been a submerged continental shelf.

A couple of volcanic islands south of Taiwan mark the southward continuation of this subduction zone to Luzon, the large northern island of the Philippines. The mountainous landscape of the Philippine Islands is a consequence both of subduction of the South China Sea floor eastward beneath Luzon and of subduction of the Philippine Sea floor westward beneath the southern Philippine islands. Volcanism and, in Luzon, crustal shortening have built the major mountains.

A major system of island arcs extends across the Indonesian islands of Sumatra and Java and eastward almost to the island of New Guinea and then again eastward along the New Britain, Solomon, and New Hebrides (Vanuatu) chains. Virtually all of the high mountains of the Sunda, or Indonesian, arc are volcanoes, some of which are associated with particularly noteworthy eruptions. In 1883 the massive eruption of the volcano on the island of Krakatoa, in the straits between Java and Sumatra, was followed by a collapse of its caldera, which caused a huge sea wave that was recorded all around the world. The eruption in 1815 of the Tambora Volcano on Sumbawa was perhaps the greatest in recorded history. Debris from this eruption darkened the skies for several months and caused a temporary global cooling that made 1816 "the year without a summer." The Sumbawa volcanic arc is associated with the northward subduction of the Indian Ocean floor beneath Indonesia. Similarly, the volcanic arcs of New Britain, the Solomon and New Hebrides islands, are associated with the northward subduction of the floor of the Solomon Sea and that of the Coral Sea beneath these island arcs.

A high range of mountains forms the backbone of the island of New Guinea between the Sunda and New Britain arcs. Whereas seafloor continues to be subducted beneath these arcs, the northern margin of the Australian continent has encountered the segment of the subduction zone between these arcs. The mountains of New Guinea consist of folded and faulted volcanic and sedimentary rocks. The volcanic rocks include both ancient seafloor and old island arcs that were thrust up and onto the northern

margin of Australia. The sedimentary rock includes a full complement of Paleozoic, Mesozoic, and Cenozoic rock deposited in the tranquil conditions of an ancient continental shelf. Thrust faulting has elevated metamorphic rock to the crest of the high range where glaciers persist even at the Equator, while the sedimentary rock is being deformed in a fold and thrust belt along the southern margin of the range.

East of the New Hebrides Islands, the Circum-Pacific System is defined by the Tonga and Kermadec islands, volcanic islands associated with the westward subduction of the Pacific Plate. The subduction zone continues southward to the North Island of New Zealand, where volcanism is the principal tectonic process that has created mountains and relief. The mountains of the South Island of New Zealand, however, have been produced by different tectonic processes. Whereas the convergence between the Pacific Plate and the seafloor beneath the Tasman Sea manifests itself as subduction of the Pacific Plate at the Tonga-Kermadec-North Island zone, it results in crustal shortening across the South Island. The Southern Alps of New Zealand have resulted from this crustal shortening, which occurs by folding, by thrust faulting, and by vertical components of slip on predominantly strike-slip faults that trend southwest across the northern and western parts of the island. Rapid uplift, possibly as much as 10 millimetres per year, keeps pace with the rapid erosion of the easily eroded schists of the Southern Alps.

The Circum-Pacific System continues southwest of New Zealand along a submarine ridge, the Macquarie Ridge. In short, the Circum-Pacific System consists of a variety of mountain types and ranges where different tectonic processes occurring at different geologic times in the past have shaped the landscape. The grouping of these different belts into this single system is thus only a crude simplification.

**The Alpine-Himalayan, or Tethyan, System.** The interconnected system of mountain ranges and intermontane plateaus that lies between the stable areas of Africa, Arabia, and India on the south and Europe and Asia on the north owes its existence to the collisions of different continental fragments during the past 100,000,000 years. Some 150,000,000 years ago, India and much of what is now Iran and Afghanistan lay many thousands of kilometres south of their present positions. A vast ocean, called the Tethys Ocean, lay south of Europe and Asia and north of Africa, Arabia, and India. Much of the rock that now forms the mountain system, which includes the Alps and the Himalayas was deposited on the margins of the Tethys Ocean.

Origin

As in the case for the Circum-Pacific System, the grouping of these different mountain ranges into a single system is an oversimplification. The various ranges (and plateaus) of the Alpine-Himalayan System formed at different times, at different rates, and between different lithospheric plates, and consist of different types of rocks.

The easternmost segment of the system begins at the western end of the Sunda island arc and continues into the arcuate chain of mountains that constitute the Himalayas, which contain the highest peaks on Earth. This chain was formed as the Indian subcontinent, a passenger on the same plate that currently underthrusts the Sunda arc, collided with the southern margin of Asia and subsequently penetrated some 2,000 kilometres into the rest of Asia. As the leading edge of India, on which Paleozoic and Mesozoic sedimentary rocks had been deposited, plunged beneath southern Tibet, these rocks were scraped off the subcontinent and thrust back onto its more stable parts. With continued penetration of the Indian subcontinent, slices of the metamorphic basement of its leading edge were scraped off the rest of it and thrust onto one another, so that the rocks of the present-day Himalayan chain consist of slices of India's ancient northern continental margin.

The Himalayan chain

Physiographically, this chain can be subdivided into three parallel belts: the Lesser Himalayas, the Great Himalayas, and the Tethys Himalayas. (Some authorities prefer a subdivision into four belts, the additional one designated the Outer, or Sub-Himalayas.) The Great Himalayas are defined by an arcuate chain of the highest peaks. To the south lie the Lesser Himalayas, a belt about 100 kilometres wide with an average elevation of 1,000 to 2,000 metres that is dissected by the rivers emanating from the Great Himalayas and north of it. To the north, the Tethys Himalayas form the southern edge of the Tibetan Plateau.

The rocks of the Lesser Himalayas consist primarily of mildly metamorphosed sedimentary rock largely of Precambrian age. At present, the remainder of the Indian subcontinent underthrusts the Lesser Himalayas on a very gently dipping thrust fault, so that the rocks forming this belt are sliding over the ancient top surface of India. As a result, the uplift of the Lesser Himalayas seems to be relatively slow.

The rate of uplift in the Himalayas seems to be rapid in two parallel zones: (1) at the very front of the range where the ancient metamorphic and sedimentary rocks of the Lesser Himalayas have been thrust up and onto the young sediments, and (2) beneath the Great Himalayas. The thrust fault that carries the Himalayas onto the intact part of India is a ramp overthrust, with the steep part of the ramp dipping north beneath the Great Himalayas. Slip on this steep part allows the rapid uplift of the Great Himalayas, which in turn creates the high peaks and carries rock from deep in the crust to the Earth's surface.

Most of the constituent rocks of the Great Himalayas are metamorphic; they once constituted the middle and lower crust of India's ancient northern margin but were subsequently scraped off and thrust up onto the surface. The very tops of many of the peaks, however, consist of Paleozoic sedimentary rocks, which dip northward. North of the Great Himalayas, in the Tethys Himalayas, these Paleozoic rocks and the Mesozoic sedimentary rocks deposited on them along the southern edge of the Tethys Ocean have been folded and faulted into east–west ridges.

Geologically, the northern margin of the Himalayas follows the Indus River in the west and the Brahmaputra River (also called Tsang-po or Yarlung Zangbo Jiang) in the east. The last remnants of the Tethys Ocean floor can be found in what some refer to as the Indus-Tsang-po Suture Zone, where a jumble of volcanic and sedimentary rocks have been folded and thrust over one another in a narrow zone parallel to these rivers. North of this suture, a belt of granites forms the backbone of the Trans-Himalayan range. These granites were intruded into the crust of the southern margin of Asia between 120,000,000 and 50,000,000 years ago, when the Tethys Ocean floor was being subducted beneath southern Asia and before India collided with it.

Since India collided with Eurasia, it has penetrated 2,000 kilometres or more into the ancient Eurasian continent. The northern edge of India may have been subducted a few hundred kilometres beneath southern Tibet, but most of its penetration has been absorbed by crustal shortening north of the collision zone. The crust of the Tibetan Plateau appears to have been severely shortened; the thickness of its crust has approximately doubled. Although much of Asia underwent extensive deformation during phases of mountain building in Late Precambrian, Paleozoic, or Mesozoic times, the high altitudes of all of the mountain ranges surrounding the Tibetan Plateau—the Pamir, Karakoram, Kunlun, Nan, Ch'i-lien (Qilian), and Lung-men (Longmen) mountains—have formed since India collided with Eurasia.

In some areas, blocks of crust undeformed since Precambrian time, such as beneath the Takla Makan (or Tarim Basin), have remained in that state, but have been displaced in response to India's penetration. The northward displacement of the Tarim Basin has caused intracontinental crustal shortening in the Tien Shan, the aforementioned east–west trending mountain range with peaks exceeding 7,000 metres in height that lies 1,000 to 2,000 kilometres north of India's northern edge. The Tien Shan was the site of Late-Paleozoic mountain building, but by the time India collided with Eurasia erosion had planed down the ancient Tien Shan to a featureless terrain buried in its own sediment. The present elevation therefore seems to be a consequence of India's penetration into Asia notwithstanding its great distance from India itself.

The Tien Shan

The penetration of India into Eurasia not only has

caused crustal thickening in front of itself, but it also is squeezing parts of Asia eastward out of its northward path. One manifestation of this extrusion of material out of its path is the crustal shortening on the eastern margin of the Tibetan Plateau, where crustal thickening is actively occurring. The eastward displacement of crustal blocks along major strike-slip faults also seems to have caused rift systems to open in a northwest–southeast direction. The Baikal Rift Zone in Siberia and the Shansi Graben in northern China seem to have resulted from the east-southeastward extrusion of material out of India's path. Moreover, crustal thickening in the Tibetan Plateau has ceased, and now east–west extension of the plateau contributes to the eastward extrusion. The plateau is laced with northerly trending rift zones that are bounded by northerly trending tilted, block-faulted mountains. Thus, virtually all of the tectonic landforms of Asia seem to be attributable, directly or indirectly, to India's collision with Eurasia and its subsequent penetration into the continent.

The eastward extrusion has been facilitated by the lack of any major obstacle to the eastward displacement of South China. Minor westward displacement of material also has occurred in Afghanistan, but this process has been blocked by the collision of Arabia with southern Iran and Turkey, where to some extent the same processes have occurred as in eastern Asia.

The Zagros and Bitlis mountains

The Arabian peninsula, with its northeastern edge covered by thick sedimentary rocks, has collided with Iran and Turkey at the Zagros and Bitlis sutures to form the Zagros and Bitlis mountains. Thick layers of salt in the sedimentary rock on the Arabian shield have allowed the detachment and folding of the overlying layers, creating a particularly well-developed fold and thrust belt in the Zagros.

While these overlying sedimentary rocks have become detached and folded, the penetration of the Arabian shield into Iran and Turkey has built plateaus in front of it and mountain ranges on the north sides of the plateaus: the Kopet-Dag and Elburz ranges north of the central Iranian plateau and the Caucasus north of the Anatolian plateau. North–south crustal shortening is the principal process by which these ranges were built, but volcanism has contributed in some cases. Many of the high mountains in this area are volcanoes, including Mount Demavand, which towers over the city of Tehrān, Mount Ararat on the border of Turkey and Armenia where Noah reputedly landed, and Mount Elbrus, the highest peak in the Caucasus. The penetration of the Arabian peninsula into eastern Turkey also has induced a westward extrusion of Anatolia, the high central part of Turkey. Thus, the same processes active in eastern Asia have affected the landscape of the western portion but only on a smaller scale.

The evolution of the western segment of the Tethyan System is the most complicated, involving more than just a collision of the African continent with parts of Europe. In Early Jurassic time (about 180,000,000 years ago), Africa, which then lay close to Europe, moved southeastward away from it. In doing so, new ocean floor (Tethys) and new continental margins formed. Much of the rock in the Alps, for instance, was deposited on this newly formed margin of southern Europe. Later, during the Cretaceous (about 100,000,000 years ago), the divergence of Africa and Europe ceased, and convergence between them began. Mountain ranges through northern Greece (the Pindus), Yugoslavia (the Dinaric Alps), Romania, Hungary, and Czechoslovakia (the Carpathians), and Austria, Switzerland, France, and Italy (the Alps) all formed as the Italian peninsula—a promontory on the African continent—moved first north-northeast toward Europe at 20 to 30 millimetres per year and later northwest at a slower rate of about 10 millimetres per year. The change in the direction of motion and the irregular shape of this promontory are two reasons that the tectonic evolutions of the different ranges of Europe are very different from each other. This is unlike the situation of the Himalayas, where the history of the belt is similar throughout the 2,500-kilometre-long range.

The western Alps

The best studied of these ranges is the western Alps in Switzerland and France. The western end of the Tethys Ocean floor was subducted beneath northern Italy until about 35,000,000 to 45,000,000 years ago. At that time, southern Europe and northern Italy collided. As the southern margin of Europe began to be subducted beneath northern Italy, the sedimentary cover deposited on the European margin of the Tethys Ocean was detached and scraped off the margin. Thick layers of relatively strong sedimentary rock (*e.g.*, limestone and sandstone) that had been deposited on weak layers of salt (and in some cases shale) became detached and folded into huge nappes—enormous, flat layers of rock that seem to have been folded and sometimes dragged over one another like sheets of cloth pushed over a table or bed.

As northern Italy continued to override the coast of southern Europe, it not only pushed the sedimentary cover farther onto the European land mass, but it also scraped up bits and pieces of the deeper metamorphic rocks of Europe's basement. Moreover, as the crust thickened, the increase in pressure and temperature metamorphosed the deeply buried rocks. Although there are exceptions, the northern and western parts of the Alps thus are dominated by folded, unmetamorphosed sedimentary rock, and the southern part consists largely of metamorphic rock.

As Europe was flexed down under the weight of the Alps thrust onto it, a foreland basin (see below) formed just north of the Alps: the Molasse Basin of northern Switzerland and southern West Germany. Continental convergence in the last 10,000,000 years has caused folding and thrusting in the Jura Mountains of northwest Switzerland and France, and displacement on ramp overthrusts beneath the front of the Alps has elevated several crystalline massifs: the Belledonne and Mont Blanc massifs in France and the Aare (or Aar) and Gotthard massifs in Switzerland. Moreover, with the elevation of the Alps above the Po plain of northern Italy, a southward overthrusting has carried the southern part of the Alps back onto the basin there as the Italian promontory has continued to penetrate into the rest of Europe.

The Apennines

The Apennines, which form the backbone of the Italian peninsula, were built by the folding and faulting of sedimentary rock deposited on the peninsula (Figure 13). The deformation in a direction nearly perpendicular to that of the Alps was due in part to a phase of the northeastward movement of Italy toward the Adriatic coast of Yugoslavia and also to the rotation of Corsica and Sardinia away from southern France and toward Italy. Thus, while the crust of the Alps was being shortened in its north–south or northwest–southeast dimension, that of the Apennines was being shortened in its northeast–southwest dimension.

While the Alps, the Apennines, and the ranges of eastern Europe were being built, different processes created mountain ranges in parts of western Europe and destroyed others in eastern Europe. For instance, while the last remnant of the Tethys Ocean, the eastern Mediterranean Sea, continues to be subducted beneath Greece and Turkey, north–south crustal extension and associated crustal thinning occurs in the Aegean area and western Turkey. This crustal thinning has already lowered the surface of what may have been a high range or plateau to below the level of the Aegean Sea and is reducing the average elevation of western Turkey.

In contrast, the western Mediterranean Sea—between Italy, Spain, and North Africa—was formed during the last 30,000,000 years and is not a remnant of the Tethys Ocean. Since that time and concurrently with the subduction of the Tethys lithosphere beneath southern Italy, Greece, and Turkey, fragments of crust have separated from southern Europe. As these fragments drifted across the ancient westernmost end of the Tethys Ocean, they opened the new western Mediterranean basin behind them.

The collisions of these fragments with parts of Italy and Africa have contributed to the building of mountain ranges in these areas. Corsica and Sardinia swung out from southern France, and the eastern margin of Corsica, which lies below sea level, collided with Italy. The Calabrian Peninsula of southern Italy once lay against Sardinia, but its southward drift opened the Tyrrhenian Sea. The volcanoes of Italy, including Mount Vesuvius near Naples and Mount Etna on Sicily, were formed as a result of the sub-

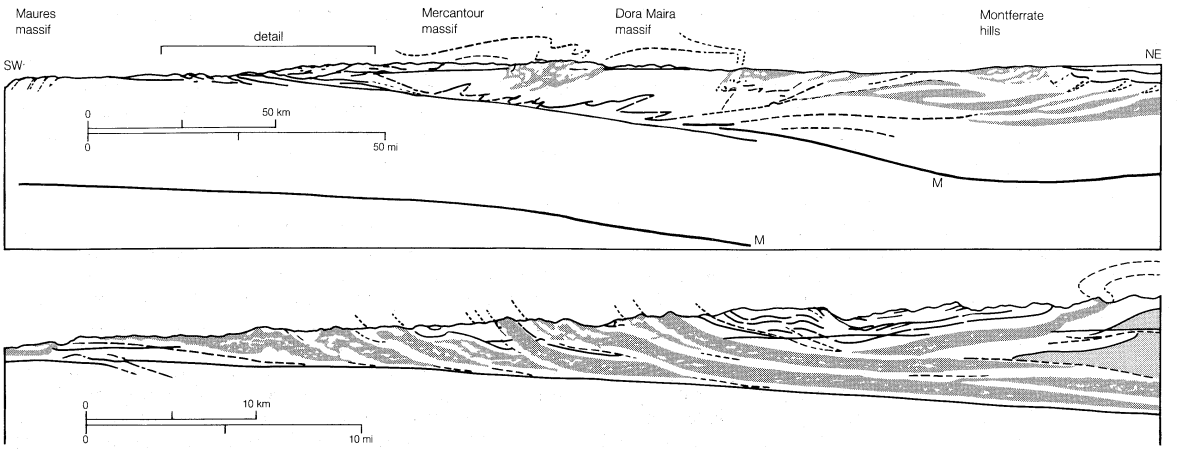



Figure 13: (Top) Small-scale cross section of the Maritime Alps and northern Apennines. (Bottom) Large-scale cross section of a portion of the Alps. Folding and faulting are apparent at both shallow and deep levels.

Adapted from Dietrich Roeder, "Geodynamics of the Alpine-Mediterranean System—A Synthesis," *Eclogae Geologicae Helvetiae*, vol. 73/2, p. 368 (July 1980); published by the Swiss Academy of Natural Sciences

duction of the ancient oceanic lithosphere of the Tethys beneath the Calabrian arc, which only recently collided with the rest of Sicily and the southern part of the Italian peninsula. Small fragments farther west collided with North Africa, causing crustal shortening and mountain building across northern Tunisia, Algeria, and Morocco.

The Pyrenees

The convergence of another small fragment with Europe built the Pyrenees. The Iberian Peninsula lay against the western margin of France until about 90,000,000 or 100,000,000 years ago, when it began to rotate into its present position and opened the Bay of Biscay behind it. As the peninsula moved toward southern France, a combination of crustal shortening and strike-slip deformation along the Pyrenees built the narrow range that separates Spain and France.

**Residual mountain ranges and thermally uplifted belts.** Although isolated mountains and mountainous terrains exist on all continents, most mountain belts not part of either the Circum-Pacific or the Alpine-Himalayan systems either are composed of residual mountains or owe their existences to localized thermally induced uplifts. Most such linear belts are residual ranges. The Appalachians in the eastern United States, for example, emerged as a result of a collision between Africa and North America in Late Paleozoic time before the present Atlantic Ocean formed. The well-developed Valley and Ridge province in the states of Pennsylvania, West Virginia, Virginia, and Kentucky has been eroded, but strong layers remain and define the ridges that once were limbs of folded layers. Similarly, the Ural Mountains were formed by the collision of Europe and Siberia in Late Paleozoic time. Much of the mountainous terrain in northeast Siberia was formed by collisions of continental fragments with the rest of Siberia in Mesozoic time.

Some high areas follow old mountain belts, but the present elevations are the result of recent uplift due to the heating of the lithosphere and to its thermal expansion. Strictly speaking, these belts are not residual mountain ranges. The mountainous topography of Norway and northern Sweden, for instance, follows an Early Paleozoic belt that marks the zone where Europe and North America (including Greenland) collided more than 400,000,000 years ago, long before the present Atlantic Ocean formed. The present-day topography, however, probably exists because this area was heated when Greenland was rifted away from Europe some 55,000,000 years ago when the North Atlantic Ocean began to form. Similarly, the mountains of eastern Australia, including the Snowy Mountains that contain the continent's highest peak, follow a Paleozoic belt; yet, the present topography seems to be the result of the warming of the lithosphere both when New Zealand separated from the east coast of Australia some 80,000,000 to 90,000,000 years ago and again when Australia drifted over a hot spot in the asthenosphere tens of millions of years ago.

The Snowy Mountains of Australia

Except for the chain of mountains across North Africa, virtually all mountains and high terrains on that continent and on Antarctica result from thermal processes. The high margins of the Red Sea and the Gulf of Aden on both Africa and Arabia are due to the heating of the lithosphere that occurred when these narrow bodies of water began to open 20,000,000–40,000,000 years ago and to the existence of a hot spot in the asthenosphere beneath the Ethiopian Plateau. Most of the high plateaus of central and southern Africa, such as the Ahaggar, formed because of hot spots beneath them. The same can be said of the high plateau that surrounds the East African Rift System and of the high volcanoes, such as Mount Kilimanjaro and Mount Kenya, built on that plateau. Similarly, the Transantarctic Mountains probably are high because of recent heating of the lithosphere beneath them. At the end of the range are two volcanoes, Mount Erebus and Mount Terror, which probably owe their existence to a hot spot beneath them.

Most of the highlands of continents that are not characterized by chains of mountains have resulted from a heating of the lithosphere. The majority of them, however, are better described as plateaus than mountain ranges.

## Plateaus

Plateaus are expanses of terrain that stand at higher elevation than surrounding terrain, but they differ from mountain ranges in that they are remarkably flat. Some plateaus, like the Altiplano in southern Peru and western Bolivia, are integral parts of mountain belts. Others, such as the Colorado Plateau (across which the Colorado River has cut the Grand Canyon), were produced by processes very different from those that built neighbouring mountain ranges. Some plateaus, as, for example, the Deccan Plateau of central India, occur far from mountain ranges. The differences among plateaus can be ascribed to the different geologic processes that have created them.

### GEOMORPHIC CHARACTERISTICS

The high flat surface that defines a plateau can continue for hundreds or even thousands of kilometres, as in the case of the Tibetan Plateau. In spite of the paucity of roads, one can drive over most of this plateau, where elevations exceed 4,500 metres, and encounter less relief than in some major cities of the world (*e.g.*, San Francisco or Rio de Janeiro). Although ranges of hills and mountains rise above the rest of the plateau, their topography, too, is rather gentle.

Dissection by rivers and deep canyons

Plateaus dissected by rivers have remarkably uniform maximum elevations, but their surfaces can be interrupted by deep canyons. In the case of some regions described as plateaus, the surface is so dissected that one does not see any flat terrain. Instead, such a plateau is defined by a uniform elevation of the highest ridges and mountains.

The eastern part of the Tibetan Plateau, which constitutes the headwaters of many of the great rivers of Asia (*e.g.*, Huang Ho, Yangtze, Mekong, Salween, and Irrawaddy), is dissected into deep canyons separated by narrow, steep ridges; the high uniform elevation that characterizes plateaus is only barely discernible in this area.

FORMATIVE PROCESSES

The formation of a plateau requires one of the same three types of tectonic processes that create mountain ranges—volcanism, crustal shortening, and thermal expansion (see above). The simplest of these is thermal expansion of the lithosphere (or the replacement of cold mantle lithosphere by hot asthenosphere).

When the lithosphere underlying a broad area is heated rapidly—*e.g.*, by an upwelling of hot material in the underlying asthenosphere—the consequent warming and thermal expansion of the uppermost mantle causes an uplift of the overlying surface. If the uplifted surface had originally been low and without prominent relief, it is likely to remain relatively flat when uplifted to a relatively uniform elevation. The high plateaus of East Africa and Ethiopia were formed this way. As in parts of Africa, plateaus of this sort can be associated with volcanism and with rift valleys, but these features are not universal. Most of the high plateau in East Africa that holds Lake Victoria does not contain volcanic rock and is cut only by small, minor rift valleys.

Where the uplifted surface lay at a low elevation for a very long time and was covered by resistant sedimentary rock, the flatness of the plateau can be particularly marked. The rock underlying the Colorado Plateau has undergone only very mild deformation since Precambrian time, and layers of very resistant limestone and sandstone deposited during the Paleozoic form its top surface in many areas. The warming of the underlying lithosphere in late Cenozoic time caused this area to rise to its present elevation, and those resistant Paleozoic formations define the surfaces that make the remarkably flat horizons at the Grand Canyon (see Figure 17 below).

The Colorado Plateau

The great heights of some plateaus, such as the Tibetan Plateau or the Altiplano, are due to crustal shortening. The geologic structure of plateaus of this kind is entirely different from that of the Colorado Plateau, for instance. Crustal shortening and crustal thickening, as described above, have created high mountains along what are now the margins of such plateaus. In most mountain ranges, streams and rivers transport eroded material from the mountains to the neighbouring plains. When drainage is internal and streams and rivers deposit their debris in the valleys between mountains, however, a plateau can form. The surface of this sort of plateau is defined by very flat, broad valleys surrounded by eroded hills and mountains. The rocks that make up the mountains and the basement of the valleys are often strongly deformed, but the young sediment deposited in the valleys usually lies flat. These plateaus generally survive erosion only in dry climates where erosion is slow. In many cases, the valleys, or basins, are occupied by flat dry lake beds. Thus, plateaus built by crustal shortening are really mountain ranges buried in their own debris.

Lava plateaus

A third type of plateau can form where extensive lava flows (called flood basalts or traps) and volcanic ash bury preexisting terrain, as exemplified by the Columbia Plateau in the northwestern United States. The volcanism involved in such situations is commonly associated with hot spots. The lavas and ash are generally carried long distances from their sources, so that the topography is not dominated by volcanoes or volcanic centres. The thickness of the volcanic rock can be tens to even hundreds of metres, and the top surface of flood basalts is typically very flat but often with sharply incised canyons and valleys.

The separation of plateaus into the above three types is not always easy because two or even all three of the processes involved frequently operate simultaneously. For instance, where the uppermost mantle is particularly hot, volcanism is common. The Ethiopia Plateau, on which Precambrian rocks crop out, stands high because the underlying lithosphere has been heated; however, Cenozoic volcanic rocks cover much of the plateau, especially those areas that are the flattest. Although the scale is different, there are active volcanoes and young lavas covering some broad basins on the northern part of the Tibetan Plateau. All three processes—thermal expansion, crustal shortening, and volcanism—may have contributed to the high, flat elevation of at least part of this plateau.

GEOGRAPHIC DISTRIBUTION

Plateaus of one type or another can be found on most continents. Those caused by thermal expansion of the lithosphere are usually associated with hot spots. The Yellowstone Plateau in the United States, the Massif Central in France, and the Ethiopian Plateau in Africa are prominent examples. Most hot spots are associated with the upwelling of hot material in the asthenosphere, and this hot upwelling not only heats the overlying lithosphere and melts holes through it to produce volcanoes but also uplifts the lithosphere. The relationship of such plateaus to hot spots insures both a wide distribution of plateaus and an absence of belts of plateaus or of interrelated plateaus.

Some plateaus, like the Colorado Plateau, the Ordos Plateau in northern China, or the East African Highlands, do not seem to be related to hot spots or to vigorous upwelling in the asthenosphere, but appear to be underlain by unusually hot material. The reason for localized heating beneath such areas is poorly understood, and thus an explanation for the distribution of plateaus of this type is not known.

Plateaus that were formed by crustal shortening and internal drainage lie within major mountain belts and generally in arid climates. They can be found in North Africa, Turkey, Iran, and Tibet, where the African, Arabian, and Indian continental masses have collided with the Eurasian continent. The Altiplano lies between the Cordillera Occidental composed of volcanoes and the Cordillera Oriental beneath which the Brazilian shield is being thrust. All these areas have undergone crustal shortening during Cenozoic time, and in each case the surface of the plateau includes both strongly deformed pre-Cenozoic rocks and very young, flat-lying sediment.

The Altiplano

There are some plateaus whose origin is not known. Those of the Iberian Peninsula and north-central Mexico exhibit a topography that is largely high and relatively flat. Crustal shortening clearly occurred in Mexico during the Late Cretaceous and Early Cenozoic (between 100,000,000 and 50,000,000 years ago) and in some parts of Spain during the Cenozoic, but the high elevations in either case do not seem to be supported by thick crust. These areas are probably underlain by a hot uppermost mantle, but proof of this is still lacking.

Volcanic plateaus are commonly associated with eruptions that occurred during the Cenozoic or Mesozoic. Eruptions on the scale needed to produce volcanic plateaus are rare, and none seems to have taken place in recent time. The volcanic eruptions that produce lava plateaus tend to be associated with hot spots. For example, the basalts of the Deccan traps, which cover the Deccan Plateau in India, were erupted 60,000,000 to 65,000,000 years ago when India lay in the Southern Hemisphere, probably over the same hot spot that presently underlies the volcanic island of Réunion. The Serra Geral basalts that cap a plateau of the same name on the Atlantic coast of Brazil were erupted some 135,000,000 years ago before Africa and South America separated from each other and when the future continental margins overlay the hot spot now beneath the volcanic island of Tristan da Cunha in the South Atlantic Ocean. In North America, the Columbia River basalts may have been ejected over the same hot spot that underlies the Yellowstone area today. Lava plateaus of the scale of these three are not common features on the Earth.

## Tectonic basins and rift valleys

GEOMORPHIC CHARACTERISTICS

Most tectonic basins and valleys, including rift valleys, are characterized by relatively steep, mountainous sides and flat floors. The steep sides are created by displacement on

faults such that the valley floor moves down relative to the surrounding margins, or, conversely, the margins move up relative to the floor. Differences in the elevations of valley floors and surrounding mountains or plateaus range from only several hundred metres to more than 2,000 metres in major rift valleys. The widths of tectonic valleys and basins vary from as little as 10 kilometres to more than 100 kilometres. Their lengths typically are hundreds of kilometres, but range from a few tens to thousands of kilometres.

The vast majority of tectonic basins and valleys is produced by an extension of the Earth's crust and the subsequent dropping of a block of crust into the space created by the divergence of large crustal blocks or lithospheric plates. The extension of the brittle crust causes it to fracture, and as the adjoining crustal blocks or plates move apart, a smaller block slides down into the resulting gap. The down-dropping of this block between the surrounding fault blocks, which commonly rise during an episode of crustal extension, creates a rift valley or tectonic basin. The geologic term for this type of tectonic depression is "graben," the German word for "ditch" or "trough."

Graben

Tectonic depressions also can be produced by horizontal compression of the crust—*i.e.,* by crustal shortening. Two types of compressional tectonic valleys and basins can be recognized: ramp valleys and foreland basins. A ramp valley is analogous to a rift valley but is formed by the margins of the valley being pushed over its floor. A foreland basin, on the other hand, results from a gentle downward bending or flexing of the entire lithosphere.

## PRINCIPAL TYPES

**Rift valleys.** In the simplest case, a rift valley forms when a block of crust, tens of kilometres wide and hundreds of kilometres long, drops down between two diverging lithospheric plates, much as the keystone in an arch will fall if the walls of the arch move apart. This process is responsible for the relatively symmetrical cross sections of most parts of the East African Rift System, where the valley floor lies 1,000 metres or more below the higher plateaus of Ethiopia and Kenya. In some places, the sides of the rift valley make single, steep walls as high as 1,000 metres. In others, the edges of the valleys consist of steps or tiers with each small inner block dropping with respect to its neighbouring outer block (Figure 14). Thus the deepest part of the rift valley is not always at its centre.

Volcanoes mark the axes of some, but by no means all, rift valleys. Where the lithospheric plates separate and the crust is thinned, the underlying parts of the lithosphere in the mantle also must diverge, allowing hot material from the asthenosphere to rise to shallow depths. Some such material from the asthenosphere has erupted at volcanoes within the eastern rift of the East African Rift System in Ethiopia and Kenya (Figure 14) and within a small section of the western rift in Zaire. Most of the western rift, which extends from Uganda through Lake Tanganyika and Lake Nyasa (Malawi), however, has no volcanoes.

Many rift valleys are asymmetrical with one steep wall and one gentle side. The steep wall is formed by slip on one or two major faults; however, unlike the simple grabens described above, no major fault bounds the other side of the rift valley. Instead, the other side is formed by a flexing of the lithosphere and by a tilting of the surface. Small faults are common, but over all there is a relatively gentle slope into the rift valley. Death Valley, in California, has a very steep eastern margin and a gentler western edge. The floor of Death Valley is moving down along a fault along its eastern margin and is rotating about an axis west of the valley. Thus, the most rapid sinking is along the valley's eastern edge, where the lowest point in the Western Hemisphere, Badwater, lies 86 metres below sea level. Similarly, the Baikal Rift, which contains the deepest lake in the world, Lake Baikal, has a very steep northwestern edge and a gentler southeastern margin.

Within some rift valleys are narrow ridges (10 to 20 kilometres wide) that are bounded by steep sides, separating the ridges from neighbouring parts of the valleys. A ridge of this kind is called a horst, a block of crust bounded by faults such that the flanks of the range have dropped

Horst

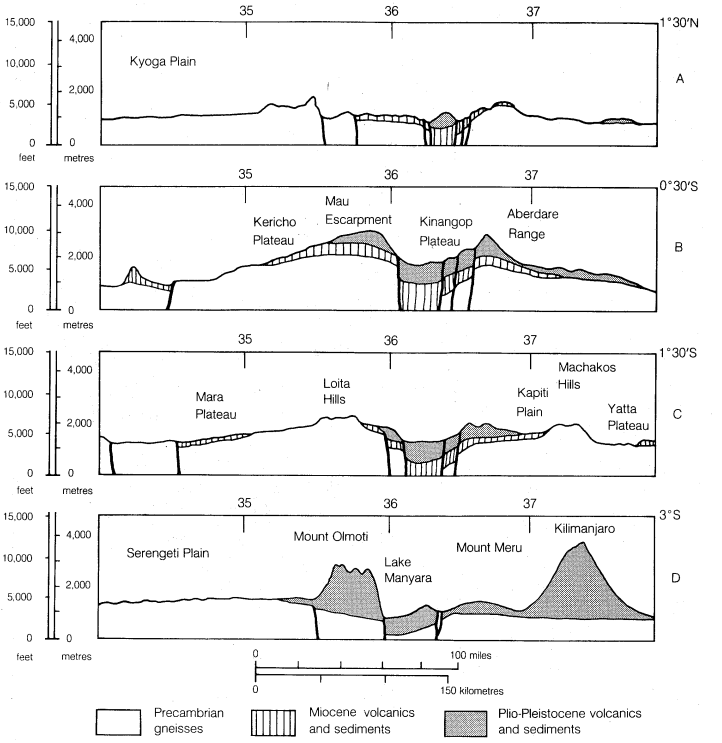



Figure 14: Cross sections of the East African Rift showing down-dropped blocks and thick volcanic rocks.

Adapted from B.H. Baker, P.A. Mohr, and L.A.J. Williams, "Geology of the Eastern Rift System of Africa"; The Geological Society of America, Inc. Special paper 136

with respect to it. A horst is the opposite of a graben. The third highest mountain in Africa, Margherita Peak of the Ruwenzori Range (located along the border of Uganda and Zaire) marks the highest point on a horst within the western rift of the East African Rift System.

Horsts can be found in most rift valleys, but unlike the Ruwenzori, they rarely dominate the landscape. The floors of most rift valleys have dropped relative to the surrounding landscape, but the tops of horsts rarely stand higher than the surface outside the valleys. Thus most horsts are merely blocks that have remained at nearly the same height as the unbroken crust outside of the rift valleys. Most horsts exist because rift valleys formed adjacent to them, not because they were elevated.

Some rift valleys, such as the East African Rift Valley in Ethiopia and Kenya, have formed over large domes (Figure 15). Upwelling of hot material within the underlying asthenosphere not only pushes the overlying lithosphere up but heats it as well, causing it to expand. To some extent the upward bulging of the lithosphere causes it to stretch, and this stretching manifests itself as a rift valley. Rift valleys that have formed in this way are commonly associated with extensive volcanism.

Certain rift valleys seem to be created by distant forces acting upon the lithosphere. These valleys cannot be associated with large domes, and in general volcanism is rare or absent. The Baikal Rift, for example, seems to be associated with the same forces that are pushing India into the rest of Eurasia. Moreover, though the elevations of the flanks are high (more than 3,000 metres in some places), the overall elevation decreases rapidly to only a couple of hundred metres at distances of just 50 to 100 kilometres northwest of Lake Baikal. Thus, a broad dome is not present.

**Basins and ranges.** Some areas, such as the Basin and Range Province of the western United States (Utah, Nevada, and California), contain an extensive network of relatively small tectonic depressions closely akin to rift valleys. The topography consists of basins 10 to 30 kilometres wide and 50 to 200 kilometres long, separated by ranges of similar dimensions. The basins contain young sediment derived from the neighbouring ranges and are quite flat. The sides of the basins can be steep or gentle. Where a major fault separates a basin from a range, the

Basin and Range Province of the western United States

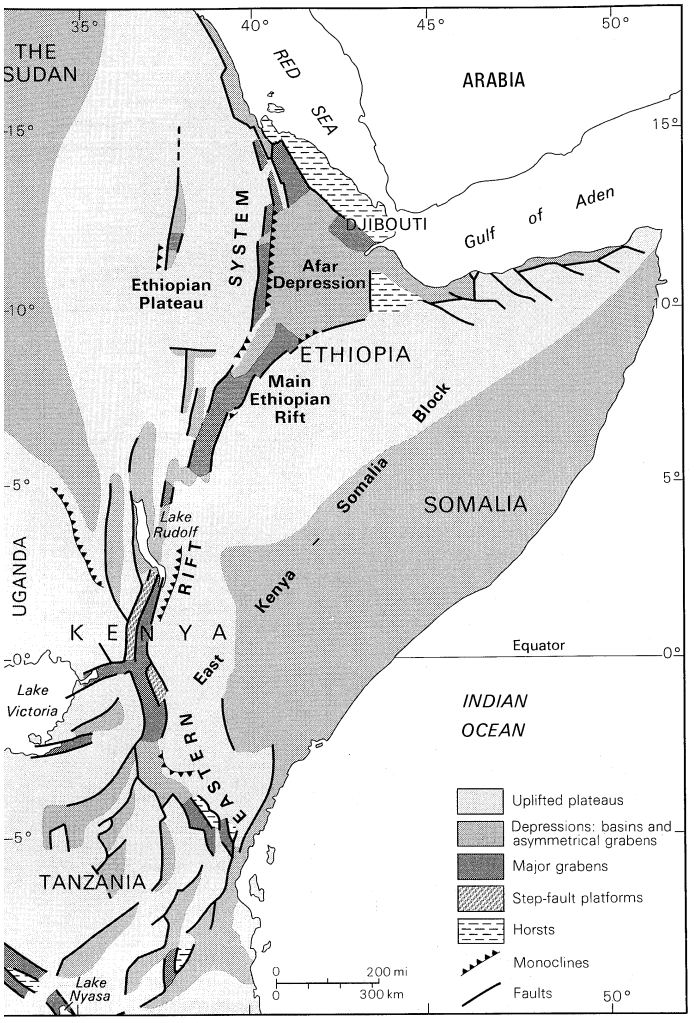



Figure 15: Topography of the eastern rift of the East African Rift System (see text).

Adapted from B.H. Baker, P.A. Mohr, and L.A.J. Williams, "Geology of the Eastern Rift System of Africa"; The Geological Society of America, Inc. Special paper 136

edge of the basin is often steep. Where the edge of the basin is produced by the tilting of the basin down and of the range up, the flank is gentle, with average slopes of from a few to 15°. These tilted, gently dipping slopes are particularly apparent wherever lavas, resistant to erosion in dry climates, had flowed onto the surfaces before they were tilted. Such tilted lava-capped surfaces are known as louderbacks. In sum, the tectonic basins of the Basin and Range Province are similar to rift valleys, but their dimensions are smaller, and the ranges are tilted blocks or horsts.

Networks of basins and ranges exist in several other high plateaus. Northerly trending basins lace the Tibetan Plateau; however, unlike those of the western United States, they are more widely spaced, occurring hundreds of kilometres apart. Moreover, a single northerly trending range in Tibet does not in general separate neighbouring basins from one another. The development of a basin and range morphology in Tibet is at a much earlier geologic stage than that of the western United States. The landscape of western Turkey likewise is cut by easterly trending basins and neighbouring ranges that were formed by crustal extension in its north–south dimension. This morphology of basins and ranges extends westward beneath the Aegean Sea. Many of the islands in the Aegean are ranges between basins that stand high enough to poke above sea level. Thus, whereas the dominant feature in a rift valley is the deep wide valley itself, the ranges and valleys are of comparable importance in basin and range topography.

**Pull-apart basins.** Some tectonic valleys are rectangular or rhomb-shaped basins, bounded by as many as four steep sides. The Dead Sea, the lowest place on Earth, lies 396 metres below sea level at the bottom of just such a basin. Another is the Imperial Valley of southern California, most of which also lies below sea level. These tectonic valleys are closely related to major strike-slip faults—nearly vertical faults along which material on one side moves horizontally with respect to that on the other.

In regions such as the Dead Sea or southern California, nearly parallel strike-slip faults bound two sides of the tectonic valley and end at the valley. Slip on the overlapping segments of the strike-slip faults results in crustal extension in the region between the two faults. Thus two sides of the tectonic valley are bounded by faults with primarily horizontal displacement, and the other two sides are bounded by faults with vertical components of slip. These basins are called pull-apart basins because the crust is literally pulled apart in the section between the two strike-slip faults.

**Ramp valleys.** As previously noted, these depressions are similar to rift valleys, but they have been formed by the opposite process—crustal shortening. A ramp valley develops when blocks of crust are thrust toward one another and up onto an intervening crustal block. The latter is forced down by the weight of this material, resulting in the formation of the valley. The thrusting of the material onto the intervening crustal block creates high mountains adjacent to the valley.

Formation by crustal shortening

Ramp valleys are characterized by steep sides tens of kilometres apart, and flat floors, which contain debris eroded from the neighbouring mountains. Escarpments on the edges of ramp valleys are not as sharply defined as for simple rift valleys, but the surrounding mountains can be higher than those that bound the latter. To a casual observer, the landscapes of ramp and rift valleys are very similar. In fact, early theories for rift valleys incorrectly attributed their origin to that of ramp valleys.

The most spectacular example of a ramp valley is the Turfan Depression, the second lowest place on Earth (154 metres below sea level), which lies within the Tien Shan of western China and along the northern margin of the Gobi. In general, the rapid filling of ramp valleys in all but the most arid climates makes them ephemeral features; however, small, young ramp valleys can be found in the South Island of New Zealand east of the Southern Alps, and remnants of ramp valleys lie within the Rocky Mountains of the western United States.

**Foreland basins.** These lie in front of major mountain ranges—*e.g.*, south of the Himalayas, north of the Alps, and east of the Canadian Rocky Mountains. Most basins of this kind are subsurface features, filled with sediment eroded from the adjacent mountain ranges; thus, they are not easily recognized in the flat landscape that is visible. Foreland basins are formed because the overthrusting of the mountains onto a neighbouring lithospheric plate places a heavy load on the plate and flexes it down, much as a diving board is flexed down by the weight of the diver. Foreland basins are deepest and young sediments are thickest next to the mountain range, and the thickness of material decreases gradually and smoothly away from the mountains. The rapid deposition of sediment from the mountains makes a nearly flat surface, such as the Indo-Gangetic Plain of northern Pakistan and India where the Indus and Ganges rivers flow south of the Himalayas. Foreland basins can be important sites of oil and gas reserves. (P.H.M.)

## Volcanic and tectonic caves

### VOLCANIC CAVES

Caves of various types and sizes occur where volcanic rocks are exposed. These are caves formed by flowing lava and by the effects of volcanic gases rather than by dissolution of the bedrock. (For a discussion of the latter type, solution caves, see below *Caves and karst landscape.*) Because volcanic caves form very close to the land surface, they are easily destroyed by erosional processes. As a result, such caves are usually found only in recent lava flows, those that are less than 20,000,000 years old.

**Lava tubes.** These are the longest and most complicated of volcanic caves. They are the channels of rivers

Formation in pahoehoe flows

of lava that at some earlier time flowed downslope from a volcanic vent or fissure. Lava tubes develop best in highly fluid lava, notably a basaltic type known as pahoehoe. They rarely form in rough, clinkery aa flows or in the more massive block lavas. In pahoehoe flows volatile components remain in solution in the molten rock where they decrease both the rate at which the lava solidifies and its viscosity. Because of this, pahoehoe lava flows like a sticky liquid, sometimes rushing down steep slopes and forming lava falls.

*Process of formation.* Near the vent of a volcano, the overflowing lava is directed toward whatever natural channels or gullies are available. As the flow advances downslope, the sides begin to congeal, so that more and more of the flowing lava is confined to a progressively narrowing channel. At this stage, the lava flow behaves like a river moving at relatively high velocity in a narrow canyon. Gradually the surface of the flow becomes crusted over and may also be covered with solid blocks of lava that have been rafted along the flow. As more and more of the surface crusts over, the supply of fluid lava feeding the advancing front of the flow is confined to a roughly cylindrical tube beneath the surface. It is possible in the later stages of crusting to observe the lava river through the few remaining "windows" in the crust.

The development of a channel that feeds the advancing front of the lava flow represents the initial stage in the formation of a lava tube. The second stage is the draining of the original conduit. If the source of lava is cut off at the vent, the fluid lava in the tube continues to flow and the tube drains. The combustion of gases released from the lava maintains a high temperature, and the walls of the conduit may be fused to a black glaze. The draining of the tube may take place in stages, so that benches or ledges are formed along the walls. Lava dripping from the ceiling congeals to form lava stalactites, while lava dripping onto the floor gives rise to lava stalagmites. The floor of a lava tube often has a ropey pattern parallel to the flow direction, showing how the last dregs of the draining lava were frozen into place. Other features of the moving fluid such as trenchlike channels in the floor, lava falls over ledges, ponded lava, and embedded blocks may also be found frozen in place.

*General characteristics.* In their simplest form, lava tube caves are long tunnels of uniform diameter oriented down the slope of the volcano from which they had their origin (Figure 16). Their roofs and walls consist of solidified lava. In some cases, the floor is covered with sand or other unconsolidated material that has been washed into the cave by water. The roof of a lava tube commonly breaks down, and some caves of this type are littered with blocks of fallen ceiling material. Complete collapse of segments of the roof forms "skylights." When such openings occur at the upper end of a tube, the tube acts as a cold air trap.

D&J McClurg

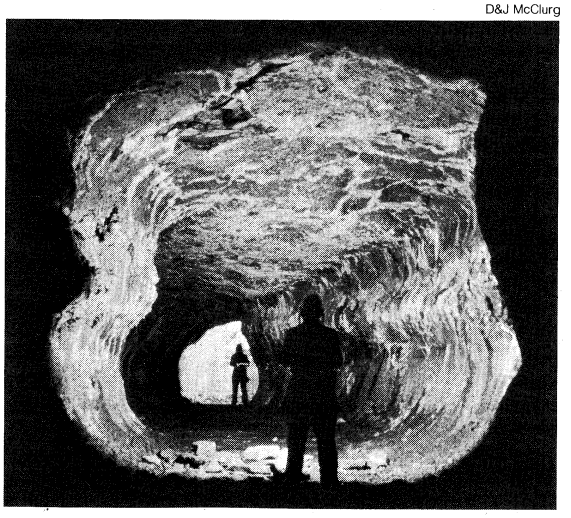


Figure 16: Rectangular-shaped passage in Bat Cave, Medicine Lake Highlands, near Lava Beds National Monument, California.

Many lava tubes contain ice formations—ponded ice as well as icicles and ice stalagmites where seepage water has frozen in the cold air trapped within the tubes. Some of these ice deposits persist far into the summer.

Complex forms

Lava tubes that have more complicated shapes also occur. Where slopes are gentle, the original lava river may branch into a distributary pattern near the toe. If these are all drained, the remaining tube branches in the downstream direction. New lava flows may override older flows and result in the formation of additional lava tubes on top of existing ones. Sometimes they are connected by younger flows falling through the roof of the older one, thus rejuvenating the older tube. Because most lava flows are thin, lava tubes form near the land surface. Portions of the roof frequently collapse, and the resulting sinkholes provide entrances to the lava tubes. The collapse process also segments the tubes, so that most lava caves have lengths of only a few hundred to a few thousand metres. Often one can line up the individual caves on maps to identify the course of the original tube. Some lava tube caves are found tens of kilometres from the vent where the flow originated.

**Other types of lava caves.** Small caves are produced in regions of active volcanism by at least three other processes. These are (1) pressure-ridge caves, (2) spatter cone chambers, and (3) blister caves.

The solidified crust of pahoehoe flows often buckles from the movement of lava underneath. The buckled crust appears as ridges several metres to a few tens of metres high, elongated perpendicular to the flow. So-called pressure-ridge caves can be formed beneath the ridges by the mechanical lifting of the roof rock. Such cavities typically measure one to two metres in height, have a roughly triangular cross section, and extend several hundred metres in length. Unlike lava tube caves that are oriented along the flow, pressure-ridge caves are oriented perpendicular to the flow.

Spatter cones

Liquid lava can be forced upward through cracks in the congealed surface layers of the flow. When the ejected blobs of liquid freeze and weld together, they form spatter cones. If the lava subsequently drains from the feeder channel, a dome-shaped chamber is formed beneath such a cone. The depths of these spatter cone pits range from several metres to a few tens of metres.

Trapped steam or other gases can lift layers of lava while it is still in a plastic state to form small blister caves. These cavities consist of dome-shaped chambers somewhat resembling those of spatter cones. They are generally small, ranging from one to a few metres in diameter, but they often occur in great numbers in many lava flows rich in volatile components.

In the United States lava caves are found chiefly in the Pacific Northwest—northern California, Washington, Oregon, and Idaho—and in Hawaii. One of the longest (measuring 3.4 kilometres) is Ape Cave on the flank of Mount St. Helens in Washington. The cave is located on the side of the volcano opposite that involved in the catastrophic eruption of 1980 and so survived the outburst. Ape Cave is only one fragment of a series of interrelated lava tubes that mark a continuous flow path down the volcano. A large number of lava tubes also occur beneath a nearly flat plain in the Bend region of central Oregon. Many of these are related to fissure eruptions rather than to a single volcanic cone. Lava tubes are commonly found in other young volcanic regions of the world, notably in the Canary Islands, on Iceland, along the East African Rift Valley, and in parts of Australia.

## TECTONIC CAVES

Tectonic caves are formed by a mass movement of the bedrock. The rocks separate along joints or fractures, and are pulled apart mechanically. The resulting cave is usually a high, narrow fissure that has nearly planar walls with matching patterns on opposite sides of the passage. The ceiling is often a flat bed of rock that did not move or that moved along some different fracture. The floor of a tectonic cave may consist of massive bedrock or of a rubble of fallen blocks, or it may be covered with soil and other material washed in from the surface.

Because tectonic caves are formed by mechanical processes, the most important characteristic of the bedrock is that it be mechanically strong. Massive, brittle rocks such as sandstones and granites are the best host rocks for tectonic caves.

Importance of gravity sliding in the formation of tectonic caves

Although tectonic caves can be formed by any geologic force that causes rocks to move apart, the key mechanism is gravity sliding. The optimum setting for the development of tectonic caves occurs where massive rocks dip gently to the sides of ridges or mountains. The presence of shale layers between beds of massive sandstone can act as a lubricating layer and facilitate mechanical slippage. Gravity causes the massive rocks to slip and separate along vertical fractures, which then become tectonic caves. The amount of slippage must be small for the cave to maintain its roof. Too much slippage and consequent roof collapse will form an open canyon. Still more slippage can result in a landslide.

Tectonic caves occur in many geologic settings and in great numbers, since they are produced by minor slippages in outcrops of massive sandstones, granites, basalts, and even limestone. Tectonic caves are among the most common caves, but they are rarely noticed or catalogued. They contain few, if any, features that attract attention and usually are quite small. Most such caves measure from several metres to a few hundred metres in length. Many of them consist of a single passage that extends into hillsides along major fractures. Some of the larger tectonic caves have a grid or network pattern that matches the pattern of the fractures or joints. (W.B.Wh.)

# STRUCTURAL LANDFORMS

Treated in this section are the topographic features formed by the differential wearing away of rocks and the deposition of the resulting rock debris under the influence of exogenetic geomorphic forces. Such forces operate at the interface of the planetary atmosphere, lithosphere, cryosphere, and hydrosphere. The processes generating these forces are the major agents of erosion, transport, and deposition of debris. These include fluvial, eolian, glacial, groundwater, and coastal-marine processes, as well as those associated with mass movement. Structural landforms result from forces generated by these processes interacting with resistances, imposed by rocks and sediments. For change to occur, the forces must exceed the thresholds of resistance imposed by the earth materials on which they act. The landform itself, however, may alter the forces by developing specific shapes. Sand dunes, beaches, river valleys, and glacial drumlins are all examples of landforms that modify the forces imposed upon them. Such self-regulation of landform development is a quality of landscapes that achieve equilibrium.

Nature of structural landforms

Although structure and lithology establish the resistance factors for structural landforms, climate defines the nature of the exogenetic geomorphological processes. In cold regions ice-related processes dominate in the development of landscapes, while in warm-wet regions fluvial processes exert primary control. Thus, a climatically controlled style of landscape development is imposed on the structurally defined surface. Moreover, process and structure interact through geologic time on an evolving landscape. As pointed out by the eminent William Morris Davis, landscape is a function of the trilogy of structure, process, and time (see above).

## Stream valleys and canyons

Wherever sufficient rainfall occurs, opportunity exists for the land surface to evolve to the familiar patterns of hills and valleys. Of course, there are hyperarid environments where fluvial activity is minimal. There also are geomorphological settings where the permeability of rocks or sediments induce so much infiltration that water is unable to concentrate on the land surface. Moreover, some landscapes may be so young that insufficient time has elapsed for modification by fluvial action. The role of fluvial action on landscape, including long-term evolutionary processes, is considered here in detail. For additional information on fluvial and hillslope processes relating to valley formation, see GEOMORPHIC PROCESSES and RIVERS.

Probably the world's deepest subaerial valley is that of the Kāli Gandak River in Nepal. Lying between two 8,000-metre Himalayan peaks, Dhaulāgiri and Annapūrna, the valley has a total relief of six kilometres. Because the Himalayas are one of the Earth's most active areas of tectonic uplift, this valley well illustrates the principle that the most rapid downcutting occurs in areas of the most rapid uplift. The reason for this seeming paradox lies in the energetics of the processes of degradation that characterize valley formation. As will be discussed below, the steeper the gradient or slope of a stream, the greater its expenditure of power on the streambed. Thus, as uplift creates higher relief and steeper slopes, rivers achieve greater power for erosion. As a consequence, the most rapid processes of relief reduction can occur in areas of most rapid relief production.

A canyon is a deep and very narrow type of river valley. Perhaps the most famous example is the Grand Canyon of the Colorado River in northern Arizona (Figure 17). The Grand Canyon is about 1.6 kilometres deep and 180 metres to 30 kilometres wide and occurs along a 443-kilometre long reach where the Colorado River incised into a broad upwarp of sedimentary rocks.

### GEOMORPHIC CHARACTERISTICS

The relief of valleys and canyons is produced by the incising action of rivers. Hillslope processes are indeed critical in the development of valley sides (see below), but it is rivers that lower the level of erosion through degradation. Rivers ultimately adjust to a base level, defined as the lowest point at which potential energy can be transformed to the kinetic energy of river flow. In most cases, the ultimate base level for rivers is sea level. Some rivers drain to enclosed basins below sea level, as, for example, the Jordan River, which flows to the Dead Sea in Israel and Jordan. Moreover, rivers may adjust to local base levels, including zones of resistance to incision, lakes, and dams (both natural and artificial).

Primacy of fluvial action

**Valley longitudinal profiles.** The longitudinal profile of a valley is the gradient throughout its length. Valleys formed by river action typically have a concave upward profile, steep in the headwaters and gentle in the lower reaches. The lower end of such a profile is adjusted to an effective lower limit of erosion defined by the base level (Figure 18).

In an ideal case of river adjustment to uniformly resistant materials, the longitudinal profile of a stream assumes a characteristic form that minimizes variations in trans-

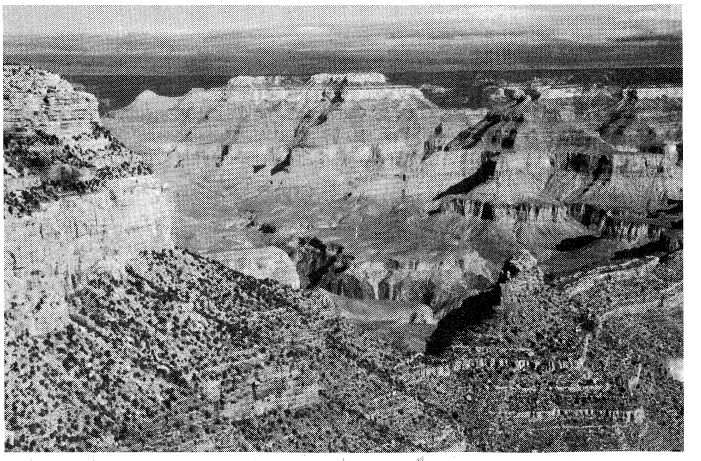


Figure 17: Grand Canyon of the Colorado River in northern Arizona.

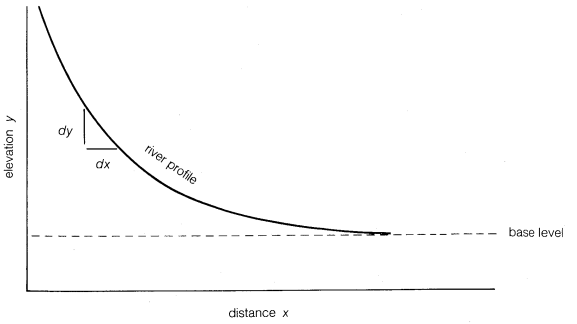



Figure 18: Simplified concave upward longitudinal profile of a river that illustrates the concept of base level and the defining terms used in calculating stream power.

porting power. Power in a river derives from the rate of transfer of potential energy, $dE/dt$, which depends on the rate of fall in elevation of water, $dy/dt$, according to

$$\frac{dE}{dt} = m\,g\,\frac{dy}{dt}, \tag{1}$$

Where $E$ is energy, $t$ is time, $m$ is mass, $g$ is the acceleration of gravity, and $y$ is elevation. The rate of fall in elevation, in turn, can be expressed as follows:

$$\frac{dy}{dt} = \frac{dy}{dx}\,\frac{dx}{dt} = S\,V, \tag{2}$$

where $S$ is the slope (fall in elevation, $dy$, with downstream horizontal distance, $dx$), and $V$ is the flow velocity (change in horizontal distance, $dx$, with time, $dt$).

Combining equations (1) and (2) and using the fluid density $\rho$ (mass per unit volume of water), one obtains

$$\frac{dE}{dt} = \rho\,g\,(W \cdot D \cdot L)\,S\,V, \tag{3}$$

where $W$ is channel width, $D$ is channel depth, $L$ is a unit length of stream, and the other parameters are as defined above. Because flow discharge $Q$ is defined as

$$Q = W \cdot D \cdot V, \tag{4}$$

the power per unit length of flow, $\Omega$, can be expressed as

$$\Omega = dE/dt/L = \rho\,g\,Q\,S. \tag{5}$$

It should be noted that in order to minimize variation in power, a river increasing its discharge in a downstream direction must decrease its slope. Thus, slope must be constantly decreasing downstream, explaining the concave upward character of the longitudinal profile.

The idealized concave upward longitudinal profile defined purely by energy considerations, noted above, only occurs where channel bed resistances and adequate adjustment time permit. Resistant zones of bedrock require greater power for a stream to incise at a given discharge $Q$ than do less resistant zones. Therefore, by equation (5) the stream gradient $S$ must be locally steeper at resistant zones. Similarly, a rapid base-level change, such as a fall of sea level, may not allow adequate time for the entire longitudinal profile to adjust. One indication of such effects on a longitudinal profile is a nick point, or abrupt change in slope of the profile.

**Valley cross profiles.** The cross profiles of valleys involve a combination of fluvial and hillslope processes. Although slopes and rivers are often studied separately by process geomorphologists, hills and valleys are the features that dominate landscapes. In upland areas cross profiles of valleys are often narrow and deep. Canyon morphologies are most common. Further downstream, valley floors are wider and often dominated by floodplains and terraces.

Classification scheme

*Types of valleys.* One of the few classifications of valleys is that used by the German climatic geomorphologists Herbert Louis and Julius Büdel. In areas of rapid uplift and intense fluvial action such as tropical mountains, *Kerbtal* (German for "notched valley") forms occur. These are characterized by steep, knife-edge ridges and valley slopes meeting in a V-shape. Where slopes are steep but a broad valley floor occurs, *Sohlenkerbtal* (meaning precisely a valley with such characteristics) is the prevailing form. Valleys of this kind develop under the influence of groundwater flow in Hawaii (see below *Processes*). Gutter-shaped valleys with convex sides and broad floors are called *Kehltal;* and broad, flat valleys of planation surfaces are termed *Fachmuldental.*

It is important to remember that the form of valleys reflects not only modern processes but also ancient ones. The entire valley or some landforms within it may be relict, with features inherited from past geologic periods during which occurred tectonic and climatic processes of intensities quite different from those prevailing today.

*Hillslopes.* Hillslopes constitute the flanks of valleys and the margins of eroding uplands. They are the major zones where rock and soil are loosened by weathering processes and then transported down gradient, often to a river channel.

Principal hillslope types

Two major varieties of hillslopes occur in nature (Figure 19). On weathering-limited slopes, transport processes are so efficient that debris is removed more quickly than it can be generated by further weathering. Such hillslopes develop a faceted or angular morphology in which an upper free face, or cliff, contributes debris to a lower slope of accumulation. Slopes of this sort are especially common on bare rock where the profile of the slope is determined by the resistance of the rock, not by the erosional processes acting on it. One consequence of this is that many rock slopes retreat parallel to themselves in order to preserve the characteristic slope angle for a rock type of given strength. If the features of the rock change with depth into the slope, however, the characteristic angle of the slope will change. Rock slopes develop where weathering and soil erosion are slow (as in arid regions) and where rock resistance is high.

The second major variety of slope is transport limited. Transport-limited slopes occur where weathering processes are efficient at producing debris but where transport processes are inefficient at removing it from the slope. Such slopes lack free faces and faceted appearances, and they are generally covered with a soil mantle. The profile of this type of slope generally has a sigmoid appearance, with convex, straight, and concave segments. The shape of the slope is an expression of the process acting upon it.

Convex slope segments commonly occur in the upper parts of soil-mantled slopes, as near the drainage divide. The noted American geomorphologist G.K. Gilbert elucidated the principles applying to convex slopes in his study of piles of mining-waste debris in California. The processes of soil creep and raindrop splash erode soil on the upper parts of slopes. Since soil eroded from the upper slope must pass each point below it, the volume of soil moved increases with distance from the divide. Since the transport rate for creep and rain splash is proportional to the slope angle, the slope angle must also increase from the divide, resulting in the slope convexity.

Straight slope segments are dominated by mass movement processes. Talus slopes are a type in which debris piles up to a characteristic angle of repose. When new debris is added to the slope, thereby locally increasing the angle, the slope adjusts by movement of the debris to reestablish the angle. Again, the result is a dynamic equilibrium in which the landform adjusts to processes acting upon it.

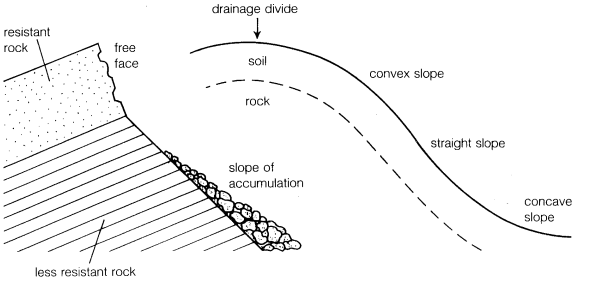



Figure 19: Comparison of idealized profiles for weathering-limited, faceted hillslopes (left) and transport-limited, sigmoid hillslopes (right).

Concave slopes are especially common where overland-flow runoff transports sediment derived from upper slopes. Because the collection area for wash increases downslope and discharge $Q$ is proportional to collection area, stream power—equation (5)—can be maintained at lower slope angles. In addition, the size of particles being transported decreases downslope because of weathering and abrasion. Because the finer particles are easier to transport, slope angles can be reduced in the downslope direction. The result is a concave shape to the slope profile.

### ORIGIN AND EVOLUTION

River valleys figure prominently in the evolutionary sequence of landscape development conceived by W.M. Davis (see above *Davis' erosion cycle theory and related concepts*). Unfortunately Davis' marvelous deductive scheme of progressive landscape change with time was somewhat abused by those who employed it merely for description and classification. By the mid-20th century, the focus of geomorphological research shifted from evolutionary sequences to studies of processes. Today, new procedures for radiometric dating have rekindled interest in long-term landscape evolution.

Valley development with time can be conceived of as a functional relationship, as follows:

$$v = f(c, r, l, p, t), \tag{6}$$

where $v$ is the valley morphology, $c$ is the climate, $r$ is relief factors including slope, $l$ is lithology and rock structure, $p$ is the type of process operating (surface runoff or spring sapping), and $t$ is time.

Significance of drainage density

Valley morphology can be described in numerous ways. A useful measure is drainage density $Dd$, which relates the length of valleys (or streams) $L$ to the area $A$ in which they occur:

$$Dd = \frac{\Sigma L}{A}. \tag{7}$$

In many applications, $A$ is defined as the drainage area in which a network of valleys is developed. There is a close relationship of drainage density to hillslope angles and local relief. For a given relief, higher drainage density results in short, steep valley-side slopes. For the same relief, a lower drainage density results in long, gentle slopes.

**Processes.** *Runoff processes.* When rain falls on a land surface, part of it may infiltrate, depending on the rate of rainfall and the permeability of the substrate. The amount of rainfall that exceeds the infiltration capacity collects in pools and eventually flows over the land surface. This process of overland flow is quite inefficient because a large surface area greatly resists water movement. Depending on the substrate resistance and power of the flow, the tendency is to incise to form a channel. This transition from overland flow to channel flow is the first step toward a response to rainfall input. Eventually the dissection by channels leads to the differentiation of hills from valleys.

Not all the rainfall is transformed to overland flow and infiltration to groundwater. A portion is lost to evaporation and to transpiration by plants. What eventually flows off the landscape from surface and subsurface sources is the runoff $R$ given by

$$R = P - ET \pm S, \tag{8}$$

where $P$ is the precipitation and $ET$ is the combination of evaporation and transpiration; $S$ is a storage term for water held in plants, soils, and subsurface rocks. The overland flow component of runoff appears very quickly after storms, while the subsurface flow components appear much more slowly. In channels, all forms of runoff generate increased stream power because of increased discharge. This allows streams to incise, thereby deepening valleys, which may widen through hillslope processes.

*Sapping.* Sapping is a process of hillslope or scarp recession by the undermining of an overlying resistant material in the form of weathering or water flow occurring in an underlying less-resistant material. A variation of this process, spring sapping, occurs where groundwater outflow undermines slopes and, where appropriately concentrated, contributes to the development of valleys. The action of groundwater in sapping may be concentrated at valley heads, leading to headward growth. Both enhanced weathering and direct erosion by the concentrated fluid flow lead to slope undermining and collapse at sites of groundwater outflow.

Spring sapping

A conceptual model of valley development by sapping can be envisioned with the initial condition of a water table having a regional slope toward a hydraulic sink provided by a depressed region. Water emerging along a spring line would then foster chemical weathering and thereby increase the porosity of the seepage zone, reducing the local rock tensile strength and rendering the weathering zone more susceptible to erosional undercutting of adjacent slopes. Local zones of heterogeneity in the rock will result in some zones achieving the critical conditions necessary for such undermining before other zones achieve them. Joints, faults, and folds serve this function. These critical zones then experience enhanced undermining. Once initiated, this process becomes self-enhancing because the lines of groundwater flow converge on the spring head. The increased flow accelerates chemical weathering, which leads to further piping at the same site.

The farther a spring head retreats, the greater the flow convergence that it generates, thereby increasing the rate of headward erosion. Headward sapping proceeds faster than valley widening because the valley head is the site of greatest flow convergence. Headward growth, however, may intersect other zones that are highly susceptible to sapping. A particularly favourable zone will result in a tributary that also experiences headward growth and that may generate tributaries of its own. Thus, sapping that occurs in a zone of jointing or faulting will develop a pattern aligned with those structures. It will, however, be organized by the hydraulic controls on the groundwater flow.

This process of sapping, headward retreat, and branching eventually forms a network of valleys. The developing network works to counteract the self-enhancing effect of flow concentration mentioned above. As spring heads migrate to the neighbourhood of one another, their demands for the available groundwater compete with each other. Eventually an equilibrium is achieved at some optimum drainage density.

Excellent examples of valleys formed by sapping are found in the massive sandstone terrains of the Colorado Plateau. Groundwater seepage from the sandstone contributes to local disintegration of the bedrock at the bases of cliffs, thereby undermining slopes and leading to backwearing. Because of structural concentration of water flow along joints and faults, valleys grow headward along zones of structural weakness. Canyons formed by sapping have prominent structural control vertical to overhanging walls, flat floors, elongate shape, low drainage density (leaving undissected uplands), relatively short tributaries to main trunk valleys, irregular variation in valley width as a function of valley length, and theatre-like valley heads. Many of the sapping valleys of the Colorado Plateau are probably relict features, since lowered water tables and/or dessicating climatic conditions have in all likelihood resulted in reduced groundwater flow to the valley floors today. During wetter climatic episodes of the Quaternary (from about 2,000,000 years ago to the present), which probably coincided with periods of mountain glaciation, spring sapping activity would have been more pronounced. Under modern climatic conditions, the results of past spring-sapping processes are obscured by the modifying action of non-sapping morphogenetic processes.

Canyons formed by sapping processes

*Valley evolution in Hawaii.* The Hawaiian Islands comprise a chain of volcanic islands, with ages increasing progressively to the northwest from the island of Hawaii with its active volcanoes, Kilauea and Mauna Loa. In general, the dissection of the Hawaiian volcanoes also increases with age to the northwest, but the details of dissection are considerably influenced by climate, factors related to parent material, and changes in process. Nevertheless, a remarkable opportunity to study valley development with time is afforded by the phenomenon of the northwesterly movement of the Pacific Plate carrying a succession of volcanoes away from a stationary mantle plume (rising jet

of partially molten rock material) located at the southern tip of Hawaii.

Rainfall is heaviest on the northeastern slopes of the volcanoes because of the prevailing trade winds. Although this results in generally higher drainage densities on the windward rather than leeward slopes of islands such as Hawaii, there are important exceptions. Mauna Loa, for example, lacks dissection on its northeastern flanks in spite of having the same amount of rainfall as highly dissected parts of Mauna Kea. Such is the case because the basaltic lava flows of the volcanoes are so permeable that drainage will not develop until a less permeable ash mantle is emplaced or until weathering reduces infiltration. Examples of both phenomena occur in Hawaii. Kilauea Volcano, the youngest of the Hawaiian shields, displays essentially no dissection except where ash from the 1790 Keanakakoi eruption was emplaced. The older Mauna Loa and Mauna Kea shields display V-shaped ravines only where their flanks were mantled by Pahala ash. Dissection is more pronounced on Mauna Kea, which is older than Mauna Loa. Kohala Volcano is the oldest shield on the island of Hawaii, having formed about 700,000 years ago. Deep weathering of its basalt has reduced infiltration sufficiently to promote high-density drainage on its northeastern slopes.

Valley initiation on the Hawaiian volcanoes thus depends on rainfall and infiltration capacity. When runoff valleys are initiated, their streams incise to form V-shaped ravines. The ravine systems eventually become deep enough that they expose deeper layers where groundwater activity and spring sapping become more important (Figure 20). The deepest incision produces U-shaped, theatre-headed valleys. Because the layered basalt flows are most permeable parallel to dip, there is efficient groundwater movement toward the sea. The regional water table on the islands is near sea level, with a slight bulge in the central parts of the islands that have a gentle seaward slope in all directions.

Formation of deep U-shaped valleys

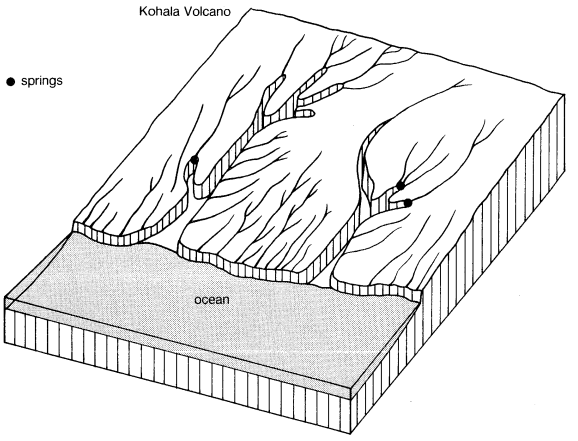



Figure 20: Association of V-shaped ravines and deep valleys enhanced by spring sapping as typical of valley development on Hawaiian volcanoes.

The U-shaped sapping valleys of the older Hawaiian volcanoes display enhanced weathering at the water table. This undermines the side slopes of the valleys, so that their steep-sided walls meet their floors at a sharp angle (Figure 21). The valleys widened laterally as they were developed by headward growth of springs at the valley heads. Perennial flow was maintained by large springs.

*Channels and valleys on Mars.* At least one other planetary body in the solar system besides the Earth is dissected by valleys of fluvial origin—namely, Mars. The heavily cratered terrains of Mars are extensively dissected by interconnected, digitate networks of valleys. Many of the valleys are steep-walled and have theatre-like headward terminations, especially near the equatorial regions of the planet. Additional properties include common structural control of the networks, low drainage densities, and low junction angles with tributaries. This combination of features seems best explained by a sapping mechanism for much of the valley formation.

Network of ancient stream valleys

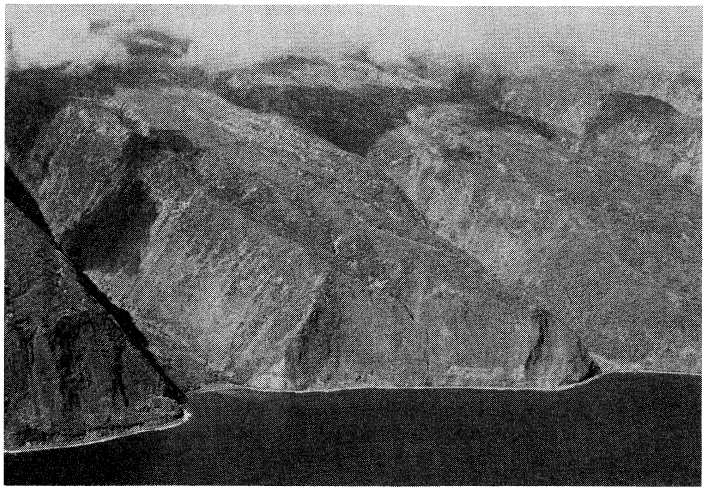


Figure 21: Large U-shaped valleys developed on Kohala Volcano, Hawaii.

The valleys of Mars are for the most part extremely ancient. Very large numbers of craters are superimposed on the valleys, indicating that they formed about the time of the phase of heavy bombardment early in the history of the solar system.

Another variety of valley on Mars occurs at fairly high latitudes where temperatures are colder. These valleys have rounded, subdued wall topography, and their floors are covered with debris that appears to have been produced from the walls and flowed across the floors. Masses of similar debris surround isolated massifs. It is probable that subsurface ice facilitated the production and flowage of the debris in a manner similar to what is observed in the Earth's periglacial regions.

**Influence of structure.** The role of structure in drainage development may be passive, in which case the composition of rocks and various rock discontinuities (joints, faults, and bedding) dictate the details of erosion. In this way, structure provides the boundary conditions for landscape degradation. During tectonism, such as faulting and folding, structural controls change with time, and the erosional system must adjust to changing resistances. Different structural controls also are encountered as incision of streams exposes lower units in the Earth's crust.

In the Soviet Union and eastern Europe, structural control is emphasized in geomorphic analysis. The Soviet academician I.P. Gerasimov defined structural units of the landscape called morphostructures as terrain types generated by a combination of tectonic activity and climate. Various morphostructures are produced by alternating periods of uplift (with resulting dissection) and stabilization (yielding planation surfaces). The history of a morphostructure and regional tectonism can be studied by analyses of river terraces, planation surfaces, and correlative sedimentary deposits.

Morphostructures

*Drainage patterns.* The pattern of fluvial dissection of a landscape is of considerable importance in understanding the structural influence on drainage evolution. Dendritic patterns (Figure 22), so called because of their similarity to branching organic forms, are most common where rocks or sediments are flat-lying and preferential zones of structural weakness are minimal. The conveyance properties of a dendritic network are analogous to blood circulation systems and tree branching. Rectangular and angular patterns occur where faults, joints, and other linear structures introduce a grain to drainage. Where a broad tilt or regional slope occurs on a surface of otherwise uniform resistance, a parallel pattern occurs. Special drainage patterns characterize belts of parallel folds (trellis pattern), domes or volcanoes (radial pattern), and other landscape types.

In a series of tilted sediments the differential erosion of softer units, such as clay and shale, results in valleys developed perpendicular to the dip or tilt of the units. These strike valleys are paralleled by ridges of the tilted sediments called cuestas. Another term for a strike stream, which

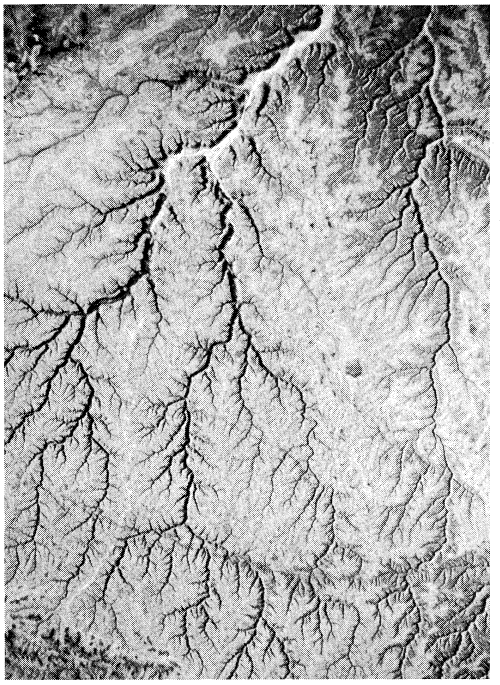


Figure 22: Dendritic drainage pattern developed on flat-lying limestone in central Yemen.
By courtesy of National Aeronautics and Space Administration

parallels the structural grain, is a longitudinal stream. In contrast, transverse streams cut across structural trends. Streams flowing down the tilted sediments of the cuesta are called dip streams because they parallel the structural dip of the strata. Streams draining the cuesta scarp into longitudinal valleys flowing opposite to the structural dip are called antidip streams.

*Cross-axial drainage.* One of the most interesting anomalies that occurs in drainage evolution is the development of stream courses across the axes of structural zones (*e.g.,* upwarps and fold belts). Some examples of cross-axial, or discordant, drainage include rivers that appear to take the most difficult routes possible through folded regions such as the Appalachian Mountains of the United States and the Zagros Mountains of Iran. The classical studies of cross-axial drainage were made during the exploration of the Colorado River system in the 19th century by the American geologist John Wesley Powell. The Colorado River and its tributaries cross great structural upwarps. Rather than flowing around domes or plunging folds, the rivers carved canyons into what appears to be paths of greatest resistance. One theory posed by Powell for such relationships is that of antecedence. According to this view, the rivers were already in their present positions when the various anticlinal folds and upwarps began to grow. A relevant analogy is a saw into which a log is being pushed. The saw represents the river and its continuing degradation, and the log represents the growing upwarp.

Possible explanations for cross-axial drainage

Another possible origin of cross-axial drainage is superimposition. According to this theory, a cover of sedimentary material must bury older structures. The river develops on this overlying sedimentary cover and subsequently imposes its pattern across the underlying structures as they are exposed by continuing degradation.

A third explanation for cross-axial drainage is that of inheritance. In this hypothesis, an erosion surface is developed across the structure zone by long-continued planation. When the streams incise, abandoning the former planation surface, they become imposed across the structures. Alternatively, the stream may actually exploit zones of weakness or minimal resistance as it downcuts from former levels. Stream capture (also called stream piracy) occurs as more aggressively eroding portions of the drainage cut through divides. In many cases, a complex combination of the above processes probably occurs to yield the final result.

Stream capture

Stream capture is especially common where longitudinal streams flowing on the weaker rocks of a fold belt erode into the valleys of transverse streams that must cross the resistant strata. Sections of valley abandoned after such captures are known as wind gaps. These contrast with the water gaps that still contain transverse streams. The famous water gaps of the Appalachians are excellent examples of such patterns.

**Influence of climate.** The importance of climate in landscape evolution, particularly valley development, has been emphasized by many European geomorphologists. Jean Tricart and André Cailleux of France and Julius Büdel of Germany developed climatic geomorphology as a synthesis of relief-forming processes. Climatic geomorphologists define systematic morphoclimatic zones on the globe in which relief-forming mechanisms differ as a function of climate. Some of the important morphoclimatic zones are briefly outlined in the following sections.

*Periglacial zone.* The term periglacial relates to cold-climate processes and landforms (see below *Periglacial landforms*). The most important periglacial influence on valleys is frost action, which produces abundant debris by freeze-thaw action on rock and soil. During the coldest periods of the Quaternary (the last 1,600,000 years), the periglacial zone was enlarged to approximately twice its present extent. Hillslopes became mantled with frost-shattered rubble that moved downslope during cycles of freezing and thawing. The relicts of this periglacial activity characterize much of the modern humid-temperate zone—*e.g.,* in portions of Pennsylvania and Wisconsin in the United States and England and Poland.

*Arid zone.* In arid regions moisture conditions are inadequate to support abundant vegetative cover of the land surface. As a result, the land is subjected to intense fluvial, eolian, and mass-wasting processes. The importance of fluvial action may seem ironic for an arid region. Although most arid regions receive little rainfall, the amount that falls is especially effective. Rare but intense arid-region rainstorms act upon a landscape that is unprotected by vegetation. Of course, some hyperarid regions receive such infrequent rainfall that fluvial processes are indeed ineffective. Nevertheless, even the most arid places on Earth show evidence of fluvial activity, either because of wetter conditions in the past or because of very rare rainstorms.

*Tropical zone.* Tropical regions are dominated by dense vegetative cover and deep weathering profiles. In continuously humid tropical zones, fluvial activity is facilitated by intense rainfall but inhibited by the protective effect of rain forests. The lateritic soils of these regions, however, do not promote deep root penetration, and the vegetative cover may be undermined by fluvial erosion or mass movement. Fluvial activity may be quite intense in the tropics, especially in tectonically active areas. In more stable cratons, however, the landscape is dominated by low-relief planation surfaces. Rivers flowing on the deeply weathered regolith of these surfaces have low stream power and transport mainly fine-grained weathering products. Thus, immense contrasts in fluvial activity exist in the tropics.

At higher tropical latitudes, the continuously humid zone of the Equator changes to a zone of seasonal rainfall. Such regions have savanna vegetation because of prolonged dry seasons. Erosion rates may be extremely high in savanna environments.

**Role of climatic change.** Because the Earth's climate has changed profoundly during the Quaternary and Tertiary (roughly the past 65,000,000 years), many landscapes are palimpsests—*i.e.,* they are composed of relict elements produced under the influence of past climates and modern elements produced in the present climatic regime. The study of such landscape changes is sometimes called climato-genetic geomorphology. Some researchers in the field, notably Büdel, have maintained that little of the extant relief in humid temperate regions of the Earth results from modern relief-forming processes. Rather, they believe, much of the familiar humid temperate landscape is inherited from past climatic conditions, including periglacial, arid, and tropical.

The views of the climato-genetic geomorphologists

Büdel focused attention on differences in the nature of valley formation as a function of climate through the

history of a landscape. He argued that very rapid valley formation accompanied periods of periglacial activity in central Europe. Modern rivers in the region seem less effective at valley incision. Most of them flow on fills within great bedrock valleys, indicating that aggradation rather than downcutting is occurring today. Prior to the Quaternary phase of valley cutting, central Europe in the Tertiary seems to have experienced a prolonged period of planation, resulting in low-relief plains. Büdel proposed that the remnants of these Tertiary plains, now preserved as broad planar uplands, are inherited from a time of tropical planation. Remnants of a former tropical regolith on the uplands provide some evidence for this hypothesis.

*Paleovalleys.* The southwestern desert of Egypt is one of the most arid places on Earth. The region lacks surficial traces of active fluvial processes and is dominated by eolian activity. In this region, a research team headed by John F. McCauley of the U.S. Geological Survey discovered in 1982 that the local drift sand had buried an array of valleys and other relict fluvial features. The discovery was made possible by the imaging radar system of the U.S. Space Shuttle, which penetrated several metres of the extremely dry sand to reveal the previously unknown valleys. The relict valleys were probably part of Late-Tertiary river systems that drained the eastern Sahara during relatively wet climatic conditions prior to the onset of hyperaridity in the Quaternary.

Relict river valleys of the eastern Sahara

A very important American paleovalley involves the complex history of the Ohio River. Prior to the glacial phases of the Quaternary, the preglacial predecessor of the Ohio drained northwestward from the Appalachians across the Midwest, but far north of its present course. Numerous water wells in Ohio, Indiana, and Illinois are located along this paleovalley, which is called the Teays River System. The advances of Quaternary ice over the course of the Teays River eventually caused the drainage to shift from the Teays route to one roughly paralleling the glacial boundary. The modern Ohio River is the product of this heritage.

*Misfit streams.* Another manifestation of the impact of climatic change is the misfit stream. Such streams are those for which some practical measure of size, most often the meander wavelength, indicates that the modern river is either too large or too small for the valley in which it flows. The former condition, known as an overfit stream, is relatively rare. An example, described below, occurs where cataclysmic glacial floods invaded valley systems formed by overland flow processes in a non-glacial climatic regime. The more common case is the underfit stream, in which valley morphology indicates a larger ancient stream (Figure 23).

Overfit and underfit streams

The English-born geomorphologist George H. Dury developed a theory for the widespread phenomenon of stream underfitness. He believed that, when the larger valley forms developed, climatic change was required to reduce the channel-forming discharges from past highs to the modern shrunken channel dimensions. Dury argued that the last phase of stream shrinkage occurred at the end of the last glaciation when the global climate changed from cool and moist to warm and dry. He quantified his theory, utilizing the relationship between the wavelength of modern meandering rivers ($\lambda$) and their bank-full discharge ($q_b$),

$$\lambda = 54.3\ q_b^{0.5}, \qquad (9)$$

where the units of $\lambda$ and $q_b$ are metres and cubic metres per second, respectively.

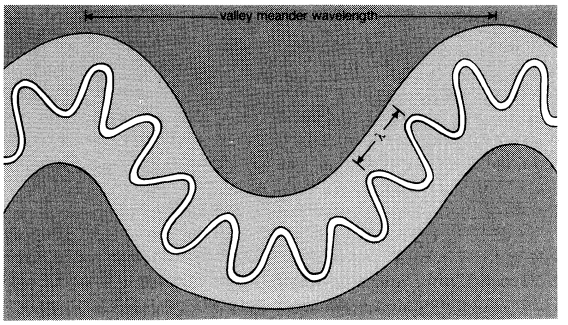



Figure 23: Idealized map view of a stream that is underfit in relation to its valley. The contrast between the meander wavelength $\lambda$ of the modern river and that of the valley is apparent.

Using equation (9), Dury found that since valley meanders were five to 10 times larger than modern river meanders, the ancient bank-full discharges must have been 25 to 100 times larger than the modern values. Such large modifications implied a phenomenal climatic change that was not accepted by the general scientific community. Numerous other factors besides climatic change play a role in the development of underfitness. These include changes in the type and amount of sediment transported by streams, the role of different rock types in shaping valley dimensions, and the role of large, rare floods (as opposed to bank-full discharge) in defining channel dimensions. The problem of underfitness remains a challenge awaiting complete geomorphological explanation.

Probably the most remarkable example of a misfitness is the channeling of the basaltic plain of eastern Washington in the northwestern United States by cataclysmic glacial floods. The great floods emanated from glacial Lake Missoula, which was impounded between about 17,000 and 12,000 years ago by a lobe of the Cordilleran ice sheet that extended into northern Idaho. Failure of this ice dam released a lake volume of about 2,500 cubic kilometres at discharges of up to $2 \times 10^7$ cubic metres per second. These immense flows completely overwhelmed the preglacial stream valleys of the Columbia Plain in eastern Washington. As the floods eroded loess and bedrock from former valley divides, a great plexus of scoured channel ways known collectively as the Channeled Scabland was formed. Because preglacial valleys were filled to overspilling, this process is really an example of stream overfitness. Numerous diagnostic landforms, including great cataracts, characterize the Channeled Scabland.

The Channeled Scabland

The above relationships were first described in the 1920s by the American geologist J Harlen Bretz, who contended that the Channeled Scabland could only be explained by the action of cataclysmic flooding. He encountered vehement opposition to this hypothesis but was eventually able to convince most of his critics of its validity by carefully documenting the overwhelming evidence for flood-produced landforms. Of considerable importance was the discovery of giant current ripples composed predominantly of gravel. More than five metres high and spaced 100 metres apart, these current ripples occurred on large bars of gravel and boulders.

*The channels of Mars.* The landforms produced by large-scale fluid flow in the Channeled Scabland are remarkably similar to those in the channeled terrains of Mars. In contrast to the Martian valley networks (see above), the channels of the planet display evidence of large-scale fluid flows on their floors. Most Martian channels show that the erosive fluid emanated from zones of complex terrain. Apparently the fluid was derived from subsurface reservoirs, and the overlying materials collapsed as fluid was released. The channels, unlike the valley networks, probably formed over a considerable span of Martian history. The fluid for carving the channels was most likely water, perhaps with substantial amounts of entrained ice and sediment. Ground ice in Martian permafrost may have provided a source for the immense ancient floods.

## Other landforms associated with stream erosion

### PLANATION SURFACES

Among the most common landscapes on Earth is that of a low-relief plain cutting across varied rocks and structures. There has been much scientific controversy over the origins of these surfaces. Because genetic implications are so often associated with various names, it seems best to refer to these features as simply planation surfaces.

Figure 24 shows a spectacular planation surface that bevels sandstone cuestas in the James Range in central Australia. Clearly an erosive process cut across rocks of

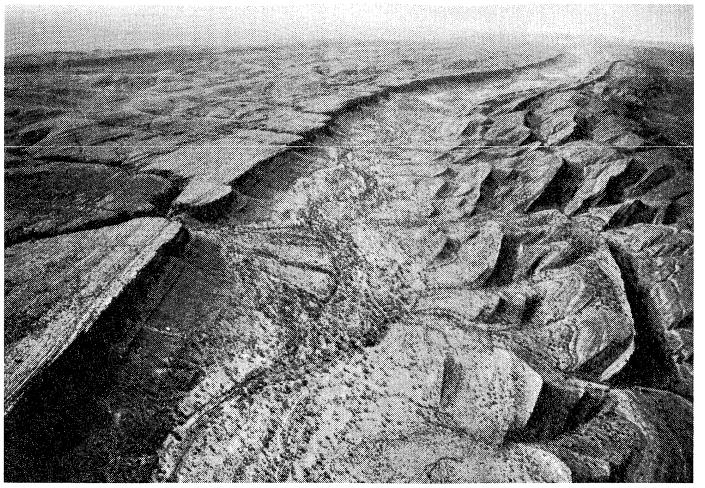


Figure 24: Planation surface cut across dipping Paleozoic sandstone in the James Range, central Australia.

varying resistance. The rock structure would never have developed such a flat surface unless a lateral erosive process had been at work in the past at a particular base level. Where many beveled cuestas of this sort line up at the same approximate elevation, they define the regional planation surface.

Planation surfaces, such as that shown in Figure 24, are especially common in the ancient, tectonically stable land masses of the Southern Hemisphere. The South African geomorphologist Lester C. King identified several phases of cyclic planation, which he correlated on a global basis. The oldest surfaces he recognized, termed Gondwana, were Mesozoic in age and related to the ancient landmass of Pangaea and its subsequent breakup during the Mesozoic. A younger surface, called the African or Moorland, developed during the Late Cretaceous and Early Cenozoic by the stripping of weathered materials from the ancient Gondwana surfaces. Younger surfaces developed during the Late Tertiary and Pleistocene as incomplete planation at levels below the remnants of the ancient plains.

Piedmonts

**Pediments.** The portion of a plain adjacent to mountain slopes is known as a piedmont. In desert regions the characteristic faceted slopes of the mountain front result in a pronounced juncture of mountain and piedmont, the piedmont angle. Where piedmonts have experienced extensive erosion, often to a degree that bedrock is exposed, they constitute pediments. There may be a veneer of alluvium over the erosional surface, particularly where soft rocks (*e.g.,* shales) occur on the piedmont. Massive rocks, such as granite, may develop spectacular bare-rock pediments that sharply mark a mountain front.

Peter Kresan

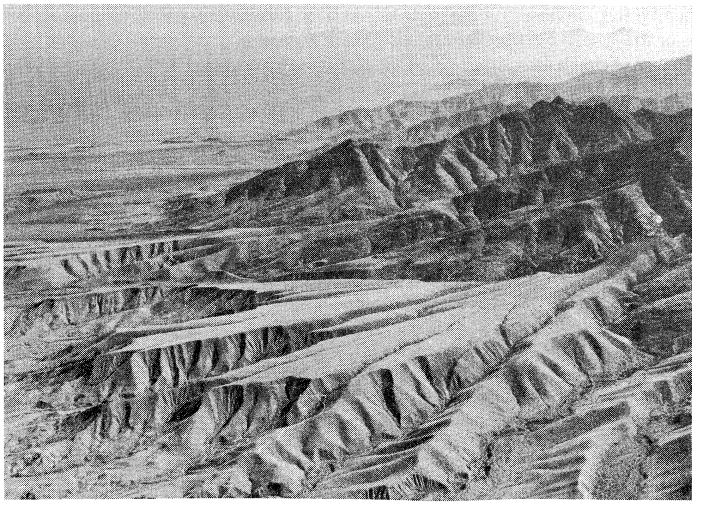


Figure 25: Dissected pediment surfaces on the northeastern flank of Mount Graham, southeastern Arizona. The pediments are mantled with early Pleistocene–late Pliocene alluvium.

Pediments probably develop through a complex of processes, including the backwearing of the front slopes of a mountain, mantling and weathering of the pediment surface, and removal of weathered mantles by fluvial and slope processes. In many cases, the erosional history of the pediment does not allow for the preservation of diagnostic evidence as to its origin. Pediments are most common in areas where tectonism is relatively slow, since rapid uplift increases the capability of streams to deliver sediment to piedmont areas, leading to a dominance of deposition over erosion. In nature, the distinction between depositional piedmont landforms (alluvial fans) and erosional ones (pediments) is often unclear. The history of a piedmont influenced by climatic and tectonic changes is almost certainly to be marked by changes in the relative erosional or depositional character. A common situation is to find ancient pediments that were once mantled with fluvial deposits (they may have constituted fans) and that were subsequently left as relict forms when mountain front streams incised the surface (Figure 25).

The growth of pediments at the expense of the mountain mass results in retreat of the mountain front. For a small mountain range in an area of tectonic stability, the entire range may be eroded. This leaves a dome-like surface composed of the coalesced pediments. Cima Dome in the eastern Mojave Desert of California is an excellent example of this advanced stage of planation.

Dome-like surfaces consisting of coalesced pediments

**Pediplain.** Where pedimentation occurs over broad regions, the coalesced surface is termed a pediplain. King believed that this process was responsible for many of the ancient planation surfaces of the world. Most geomorphologists, however, consider pedimentation to be a local process at mountain fronts, perhaps capable of generating planation surfaces for an individual mountain range but not uniquely the cause of globally correlated surfaces.

**Etchplain.** Where deep weathering occurs on a landscape, a dichotomy is set up between the thick regolith of weak, weathered rock and the underlying zone of intact rock. If subsequent erosion removes the weathered regolith, then a new planation surface develops through exposure of the old weathering front. This process often results in the exposure of structurally defined compartments of resistant rock. A subsurface landscape is essentially etched from the rock by deep weathering and subsequent removal of weathered products.

Etchplanation appears to have been especially characteristic of the ancient, stable cratonic areas of Gondwanaland, the supercontinent that many researchers believe once existed in the Southern Hemisphere. Deep weathering, generally tropical laterization, occurred through the Mesozoic and Early Tertiary. Great planation surfaces developed as tectonic factors influenced the periodic removal of weathering products.

**Peneplain.** The concept of a peneplain (the word meaning "almost a plain") emerged from W.M. Davis' cyclic view of landscape evolution. As rivers and hillslopes reduced relief through the phases of youth, maturity, and old age, explained Davis, the eventual result was a plain of extremely low relief. This plain could only change very slowly, since potential energy for fluvial action was greatly reduced. Such a peneplain, as with any planation surface, could become relict when renewed uplift induced stream incision below its former position on the plain.

Because it is tied genetically to the Davisian theory of landscape development, the concept of peneplains is rarely used in modern geomorphology. There is, however, frequent reference to peneplanation in older literature. For modern applications, it is best to use a purely descriptive term such as planation surface or erosion surface for features that were formerly classified as peneplains.

## INSELBERGS

The early German explorers of southern Africa were impressed by immense low-relief plains (planation surfaces) from which isolated highlands rose abruptly. The domed or castlelike highlands were dubbed inselbergs (literally "island mountains" in German). Spectacular examples include Ayers Rock and the Olga Rocks (Figure 26) in central Australia.

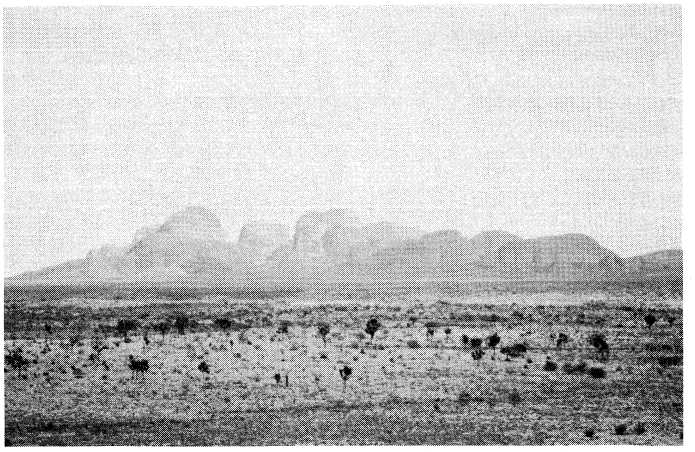


Figure 26: Granitic inselbergs known as the Olga Rocks in central Australia.
Malcolm S. Kirk—Peter Arnold, Inc.

Inselbergs are relict features. They have maintained their relief as the adjacent surrounding landscape was lowered. C.R. Twidale of Australia demonstrated the role of subsurface weathering in shaping the flanking hillslopes and pediments of granitic inselbergs.

Inselbergs as a type of paleoform

The occurrence of inselbergs implies immense variations in the rates of degradational activity on the land surface. These structures are one of several varieties of landform called paleoforms that can survive with little modification for tens of millions of years. In inselberg landscapes, the active erosional processes are confined to valley sides and valley floors.

## Landforms associated with stream deposition

### ALLUVIAL FANS AND BAJADAS

Alluvial fans are fan-shaped zones of fluvial deposition that occur where streams emerge from highlands or mountain fronts to adjacent plains. Like pediments, alluvial fans characterize piedmonts, and the two landform types may be considered, respectively, erosional and depositional end members in a continuum of piedmont landscapes. A bajada is a piedmont that consists of multiple coalescing alluvial fans.

Conditions favourable to alluvial fan formation

Tectonically active mountain fronts in desert regions are excellent areas for the development of alluvial fans. Abundant sediment production is achieved from hillslopes that lack vegetative cover, and this sediment is deposited downstream of the point where the stream emerges from the mountain front. Characteristic examples occur in Death Valley in southeastern California (Figure 27). On the western side of the valley, very long fans extend eastward from the Panamint Range. Because of the eastward tilting of the valley, the toes of fans on the eastern side are being buried by the playa sediments of the valley centre. Consequently, the eastside fans are much smaller.

L.K. Lustig—E.B. Inc.

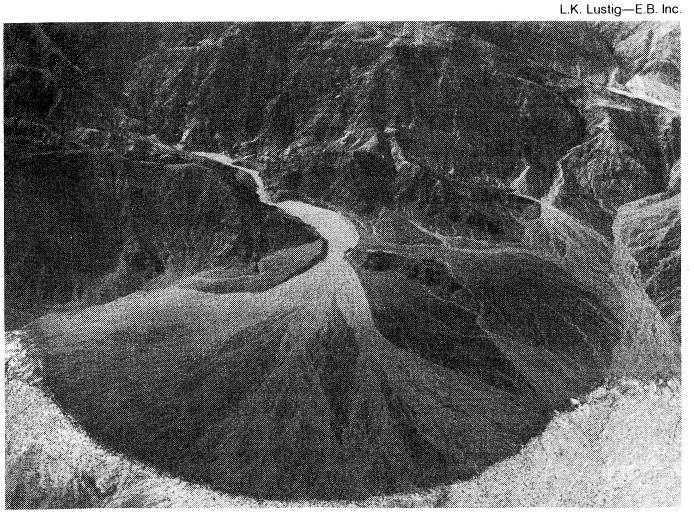


Figure 27: Alluvial fan at the mouth of Copper Canyon, Death Valley, California, an area of internal drainage.

Some of the biggest alluvial fans occur in humid regions. For example, the fan of the Kosi River, which drains the Himalayas covers an area of 15,000 square kilometres. The Kosi fan has an extremely low gradient. The slope at its apex is one metre per kilometre, and this decreases distally to less than 0.2 metre per kilometre near the Ganges River. The paleo-channels at the fan apex have diffuse patterns that indicate probable braiding when they flowed. Near the fan toe, the abandoned paleo-channels are meandering.

From 1736 to 1964, the Kosi River shifted 110 kilometres from east to west, reworking all the land in between. Unlike alluvial fans of arid regions that show somewhat random shifts of erosion and deposition, the Kosi fan was developed by a progressively shifting channel. The shifting occurred as deposition raised the active portions of the fan, and the river moved to lower terrain on the inactive portions.

Alluvial fans and bajadas may accumulate sediments in tectonic basins over millions of years. The processes that operate on them and the locations of deposition or erosion may vary through time, reflecting changes in climate and tectonism. Processes on alluvial fans include debris flows and mudflows, which are most common at the proximal apex of arid-region fans with mountainous sources of appropriate rock types. Channelized and sheetflood water flows may persist to the distal portions of fans, depending on the climate and infiltration capacity of the fan surface. Fans in humid regions are dominated by water flows, which tend to produce lower gradient fan surfaces than do debris flows. (For a detailed discussion about the characteristics and formation of alluvial fans, see RIVERS.)

### PLAYAS, PANS, AND SALINE FLATS

Playas, pans, and saline flats are among the flattest known landforms. Their slopes are generally less than 0.2 metre per kilometre. When filled with only a few centimetres of water, many kilometres of surface may be inundated. It is the process of inundation that develops and maintains the near-perfect flatness so characteristic of these arid-region landforms.

Relation to flooding

Playas occupy the flat central basins of desert plains. They require interior drainage to a zone where evaporation greatly exceeds inflow. When flooded, a playa lake forms where fine-grained sediment sand salts concentrate. Terminology is quite confused for playas because of many local names. A saline playa may be called a salt flat, salt marsh, salada, salar, salt pan, alkali flat, or salina. A salt-free playa may be termed a clay pan, hardpan, dry lake bed, or alkali flat. In Australia and South Africa small playas are generally referred to as pans. The low-relief plains of these lands contrast with the mountainous deserts of North America, resulting in numerous small pans instead of immense playas. The terms *takyr, sabkha,* and *kavir* are applied in the Soviet Union, Saudi Arabia, and Iran, respectively.

Saline flats are specialized forms located adjacent to large bodies of water, as, for example, along coasts, lakeshores, and deltas. They flood during storms, either with surface runoff or with surges from the nearby body of water. The saline crusts of saline flats are quite similar to those that develop in playas.

**Physical characteristics.** Enclosed basins of salt and clay accumulation may originate from numerous causes. Tectonic causes include faulting, as in the East African Rift Valley and Death Valley, and warping, as in Lake Eyre in Australia, Lake Chad in central Africa, and Shaṭṭ al-Jarīd (Chott Djerid) in Tunisia. Wind deflation can produce shallow basins with downwind dunes, as in southeastern Australia. Even very large basins, such as the Qattara Depression of Egypt, have been ascribed to deflation. Local cataclysmic disruptions of drainage (*e.g.,* volcanism, landslides, and meteorite impacts) may produce playas in desert regions.

Modern playa surfaces are not passive receptors of sed-

iment as they were once believed to be. They serve as important sources of dust and salts, which are blown to the surrounding uplands. Complex assemblages of minerals and sediments occur on the playa surfaces. These directly reflect their environment of deposition and may be used to interpret ancient environmental conditions.

Types of playa

Two broad classes of playas may be defined on the basis of past histories. One type develops from the desiccation of a former lake. Sediments in such a playa are primarily lacustrine, rather than derived from modern depositional processes. The second type of playa has no paleolacustrine heritage. Small salt pans in South Africa, called *vokils*, are of this type.

The supply of material, basin depth, and duration of accumulation all contribute to variations in the thickness of playa deposits. Very thick playa sequences may have alternating layers of lacustrine clays and salt beds. The former generally reflect periods of high floodwater runoff into the closed basins, perhaps induced by higher rainfall (a so-called pluvial period). Saline sediments or pure evaporite beds reflect arid climatic phases. The precise climatic interpretation of paleolacustrine playa sequences, however, can be problematic.

*Role of flooding and groundwater.* Playas affected by occasional surface floods are usually dry. Their surfaces consist of silt and clay deposited by the floodwaters that enter closed basins during the occasional flow events. Salts develop as ponded floodwater in the centre of such a basin gradually evaporates. Water also can be supplied to closed basins by groundwater flow. In basins dominated by groundwater inputs, sediment influxes are minimized, and saline crusts dominate. Moist areas may persist as groundwater flows to the lowest portion of playas. Very large playas may exhibit dry, sediment-dominated sections and moist, salt-dominated sections.

*Saline minerals.* The salt deposits of a salt pan are zoned like bathtub rings, with less-soluble sulfates and carbonates at the outer margin and highly soluble sodium chloride (table salt) at the centre. The crystallization of these salts can be compared with the evaporation of brine in a dish. The first precipitates from the evaporating brine are calcium carbonate ($CaCO_3$) and magnesium carbonate ($MgCO_3$). These form the outer "bathtub ring." The next ring consists of sulfates of calcium and sodium ($CaSO_4$ and $Na_2SO_4$, respectively). If sufficient calcium is present, gypsum ($CaSO_4 \cdot 2H_2O$) will form. If less calcium is present, thenardite ($Na_2SO_4$) and sodium carbonate ($Na_2CO_3$) may be deposited. The last remaining brines of exceptionally high salinity precipitate highly soluble chlorides of sodium, calcium, magnesium, and potassium.

Zoning of salt deposits in a salt pan

Another kind of zoning occurs in saline playas with respect to the hydration of different minerals. Dehydrated minerals, such as anhydrite ($CaSO_4$), occur on surface areas protected against flooding and in wet saline areas.

Some playas also contain exotic minerals. The Death Valley playa is famous for borate minerals, including borax ($Na_2B_4O_7 \cdot 10H_2O$) and Meyerhofferite ($Ca_2B_6O_{11} \cdot 7H_2O$).

*Surface relief and structures.* Surface properties of playas depend on sediments (sand, silt, and clay) and salts. Near-surface groundwater may give rise to evaporite crusts formed by rigorous evaporative concentration. Thick salts may form rugged crusts, as at Devil's Golf Course in Death Valley (Figure 28). Regular flooding of evaporative layers may form a very smooth surface, as at Bonneville Salt Flats in Utah. For thick, soluble crusts, dissolution may occur during fluctuations of a high water table. Solution cavities in the crust can produce a salt karst topography.

The muds deposited on playas are subject to drying and shrinking. The amount of volume change varies with the clay minerals present. Smectite clays experience the greatest shrinkage on drying. The presence of salts enhances the effect, since deposition and crystallization of salts in the cracks creates a polygonal network of salt wedges.

Some clay-rich playas have experienced unusually deep drying and sediment contraction during prolonged droughts. Giant desiccation polygons formed under these conditions are as large as 90 metres across. Individual cracks more than one metre wide and 15 metres deep have been observed.



Figure 28: The Devil's Golf Course near Badwater, Death Valley, California. Thick salts on a playa floor may form rugged crusts of this kind.
Stephen J. Krasemann—Peter Arnold, Inc.

**Geomorphic evolution.** *Impact of climatic change.* Playas, pans, and saline flats are exceptionally sensitive to environmental change. They have been most profoundly influenced by changes in hydrologic regimen induced by the climatic variations of the Quaternary. All have experienced episodes of expanded lake levels in the past. Such predecessors are often called pluvial lakes, thereby implying periods of increased rainfall. It is also possible, however, that lakes could have expanded because of other factors, including increased groundwater inflow and/or decreased evaporation/transpiration.

*Paleolake chronologies.* Modern geochronologic techniques, such as radiocarbon dating, permit the comparison of fluctuations in the paleolakes that were predecessors to many modern playas. In northern Africa lakes were at a moderately high level from 30,000 to 22,000 years ago. During the maximum cold, dry phase of the last glacial period, from approximately 20,000 to 12,000 years ago, most African lakes were at low levels, and many were dry. From 10,000 to 8,000 years ago, lakes rose to maximum high levels. Lake Chad expanded to the size of the modern Caspian Sea. Small volume lakes, however, are more sensitive to climatic change, recording higher frequency oscillations in the hydrologic balance. Since about 4,000 years ago, the north African lakes have fallen to the range of their modern lows.

Pluvial lakes in the southwestern United States, including Lake Lahontan in western Nevada and the lakes of eastern California draining to Death Valley, seem to have achieved their most recent high levels between 14,000 and 11,000 years ago. The period from 30,000 to 24,000 years ago was marked by low lake levels. Another low was reached in the middle Holocene, about 7,000 years ago. Many of the lakes of the southwestern United States, however, seem to have been not quite in phase with one another.

*Effects of wind action.* Playas and saline flats are particularly susceptible to wind action. Clays and salts form crusts that curl and flake upon drying. The flakes and curls are readily deflated, and these wind-eroded sediments are then deposited leeward of the playas and saline flats from which they were removed.

In Australia many playas have large transverse crescentic

Clay dunes, or lunettes

foredunes on their leeward side. Because of their silt and clay composition, these features are sometimes called clay dunes. In Australia they are known as lunettes. James M. Bowler, an Australian Quaternary stratigrapher, produced a precise chronology of playa development and associated eolian activity in the desert of western New South Wales, Australia. There, numerous small lakes reached their maximum extent 32,000 years ago, approximately coincident with the age of the first human remains in Australia. From about 26,000 years ago, the lakes fell to low levels. Playas formed roughly 16,000 years ago at a time when eolian activity peaked. High lakes again occurred about 9,000 to 5,000 years ago, but playas were reestablished after that.

The present association of playas, lunettes, and linear dunes in the Australian deserts may imply a causative association. C.R. Twidale proposed that the linear dunes developed as lee-side accumulations of sand trapped by the growth of lunettes. Climatic change is critical to the association. (V.R.B.)

## Sand dunes

Sand dunes are accumulations of sand grains shaped into mounds or ridges by the wind under the influence of gravity. They are comparable to other forms that appear when a fluid moves over a loose bed, such as subaqueous "dunes" on the beds of rivers and tidal estuaries and sand waves on the continental shelves beneath shallow seas. Dunes are found wherever loose sand is windblown: in deserts, on beaches, and even on some eroded and abandoned farm fields in semiarid regions, such as northwest India and parts of the southwestern United States. Images of Mars returned by the U.S. Mariner 9 and Viking spacecrafts have shown that dunes are widely distributed on that planet both in craters and in a sand sea surrounding the north polar ice cap.

True dunes must be distinguished from dunes formed in conjunction with vegetation. The latter cover relatively small areas on quiet humid coastlands (see below *Beaches and coastal dunes*) and also occur on the semiarid margins of deserts. True dunes cover much more extensive areas—up to several hundred square kilometres—primarily in great sand seas (ergs), some of which are as big as France or Texas. However, they also occur as small isolated dunes on hard desert surfaces, covering an area of as little as 10 square metres (107 square feet). Areas of gently undulating sandy surfaces with low relief are classified as sand sheets.

Sand sheets

They commonly have a nearly flat or rippled surface of coarse sand grains and are only a few centimetres to metres thick. Minor sand sheets cover only a few square kilometres around the margins of dune fields. A few, such as the Selima Sand Sheet in southwestern Egypt and the northwestern Sudan, are probably almost as extensive as some of the great sand seas.

During the last 2,000,000 years or so the conditions of very low rainfall under which true dunes form expanded beyond the margins of the Sahara and other present-day arid regions into areas that are now more humid. The best evidence for these changes is the presence of sand seas that are immobilized by vegetation. Dunes formed under similar climates in the geologic past and at certain times occupied deserts as extensive as modern ones. Rocks formed by the solidification of ancient sand seas occur, for example, in the walls of the Grand Canyon in the southwestern United States; in the west Midlands of England; and in southern Brazil.

### GEOMORPHIC CHARACTERISTICS

An understanding of sand dunes requires a basic knowledge of their sands, the winds, and the interactions of these main elements. These factors will be treated in turn in the following sections.

**Sands.** Dunes are almost invariably built of particles of sand size. Clay particles are not usually picked up by the wind because of their mutual coherence, and if they are picked up they tend to be lifted high into the air. Only where clays are aggregated into particles of sand size, as on the Gulf Coast of Texas, will they be formed into dunes. Silt is more easily picked up by the wind but is carried away faster than sand, and there are few signs of dunelike bed forms where silt is deposited, for instance as sheets of loess. Particles coarser than sands, such as small pebbles, only form dunelike features when there are strong and persistent winds, as in coastal Peru, and these coarse-grained features are generally known as granule ripples rather than dunes. Larger particles, such as small boulders, can be moved by the wind only on slippery surfaces (*e.g.*, ice or wet saline mud) and never form into dunes.

Common dune sands have median grain diameters between 0.02 and 0.04 centimetre (0.008 and 0.016 inch). The maximum common range is between 0.01 and 0.07 centimetre. Most dune sands are well sorted, and a sample of sand from a dune will usually have particles all of very similar size. The sand on sand sheets, however, is poorly sorted and often bimodal—*i.e.*, it is a mixture of coarse sands, often about 0.06 centimetre in diameter, and much finer sands, as well as particles of intermediate size. Windblown sands, especially the coarser particles, are often rounded and minutely pitted, the latter giving the grains a frosted appearance when seen under a microscope.

Minerals in windblown sand

Most windblown sand on the Earth is composed of quartz. Quartz exists in large quantities in many igneous and metamorphic rocks in crystals of sand size. It tends to accumulate when these rocks are weathered away because it resists chemical breakdown better than most minerals, which are taken away in solution. Most of the great sand seas occur in continental interiors that have been losing soluble material for millions of years; as a consequence, quartzose sandstones are common. These sandstones are eroded by rainwash and stream runoff, processes that are spasmodic but violent in deserts. The eroded products are transported to great interior basins where they are deposited. Such alluvial deposits are the sources of most windblown sand. Quartz also predominates in most coastal dune sands, but there usually are considerable mixtures of other minerals in dunes of this kind.

Dune sands not composed of quartz are rarer but not unknown. Near volcanic eruptions in Hawaii, some western states of the continental United States, and Tanzania, for example, dunes are built of volcanic ash particles. In many arid areas, gypsum crystals of sand size are deposited on the floors of ephemeral lakes as the water dries out; they are then blown like sand to form gypsum dunes. Gypsum dunes occur in the White Sands National Monument in New Mexico, as well as in northern Algeria and southwestern Australia.

**Winds.** Winds have three sources of variation that are important—namely, direction, velocity, and turbulence. Most of the great deserts are found in the subtropical areas of high atmospheric pressure, where the winds circulate in a clockwise direction in the Northern Hemisphere and a counterclockwise direction in the Southern Hemisphere. The high-pressure systems tend to dip down to the east so that winds are stronger there, a pattern mirrored by the dunes. Poleward of these circulation systems are the zones of eastward moving depressions in which there are generally westerly winds that mold the dunes of the North American and Central Asian deserts and of the northern Sahara. The boundaries between these two circulation systems migrate back and forth seasonally, so that complicated dune patterns are found in the zones of overlap. Only a few deserts, notably the Thar Desert of India and the Sonoran Desert of the American Southwest, are affected by monsoonal wind systems. Some dunes are built by sea breezes and local winds, as in coastal Peru.

The direction of the wind at any one place in the desert is affected by a number of local factors. Winds are particularly channeled around topographical features, such as the Tibesti Massif in the Sahara, so that dunes are affected by different winds on different sides of the obstruction. Winds also can be channeled around the dunes themselves, thereby developing patterns of secondary flow that modify the shapes of the dunes.

Wind velocity patterns

The pattern of wind velocity also is important. Like many natural phenomena, wind velocities have a log-normal distribution: there are a large number of moderate breezes and a diminishing number of increasingly more violent winds. The greatest volumes of sand are probably moved

by unusually strong winds, because the amount of sand moved by wind is a power function (exponential factor) of the wind speed. For example, a 10-kilometre-per-hour wind carries 13 grams per hour (0.39 ounce per hour), a 20-kilometre-per-hour wind carries 274 grams per hour, and a 30-kilometre-per-hour wind carries 1,179 grams per hour. A wind of a particular velocity will move fewer larger than smaller grains. Strong winds often blow from a particular direction, as in the southern Sahara, where the intense winds of sandstorms come predominantly from one direction. Such winds are responsible for the undulations of the sand sheets, because they alone can move coarse sands. Lighter winds blow from several different directions, and the dunes, being of finer sand, are therefore affected by several winds.

The wind is retarded near the surface by friction. Above the ground the wind velocity increases rapidly. The near-surface velocity must rise above a certain threshold value before sand will be picked up, the value depending on the size of the sand grains; for example, a wind of 12 kilometres per hour measured at a height of 10 metres is required to move sands 0.02 centimetre in diameter, and a 21-kilometre-per-hour wind is required to move 0.06-centimetre sands. Once sand movement has been initiated by wind of such velocity, it can be maintained by winds blowing at lower speeds. Because instantaneous wind speeds in eddies can rise well above the average velocity, turbulence also is important, but it is difficult to measure.

### FORMATION AND GROWTH OF DUNES

The dune-forming process is complex, particularly where many thousands of dunes have grown side-by-side in sand seas. Yet, an introductory account can be given based on the example of a single dune on a hard desert surface.

Most of the sand carried by the wind moves as a mass of jumping (saltating) grains; coarser particles move slowly along the surface as creep and are kept in motion partly by the bombardment of the saltating grains. Saltating sand bounces more easily off hard surfaces than off soft ones, with the result that more sand can be moved over a pebbly desert surface than over a smooth or soft one. Slight hollows or smoother patches reduce the amount of sand that the wind can carry, and a small sand patch will be initiated. If it is large enough, this patch will attract more sand.

The wind adjusts its velocity gradient on reaching the sand patch; winds above a certain speed decrease their near-surface velocity and deposit sand on the patch. This adjustment takes place over several metres, the sand being deposited over this distance, and a dune is built up. The growth of this dune cannot continue indefinitely. The windward slope is eventually adjusted, so that there is an increase in the near-surface velocity up its face to compensate for the drag imposed by the sandy surface. When this happens, the dune stops growing and there is no net gain or loss of sand.

Angle of repose and the slip face

As the dune grows, the smooth leeward slope steepens until the wind cannot be deflected down sharply enough to follow it. The wind then separates from the surface leaving a "dead zone" in the lee into which falls the sand brought up the windward slope. When this depositional slope is steepened to the angle of repose of dry sand (about 32°), this angle is maintained and the added sand slips down the slope or slip face. When this happens, the dune form is in equilibrium, and the dune moves forward as a whole, sand being eroded from the windward side and deposited on the lee.

If the regional rate of sand flow can be calculated from measurements of wind speed and direction, and if it is assumed that the dune has a simple cross section that migrates forward without change of form, a formula for the rate of movement of a dune that agrees with actual measurements can be derived. In Peru dunes have been observed to move at 30 metres per year; in California rates of 25 metres per year have been measured; and in the al-Khārijah Oases (or Kharga Depression) in southern Egypt dunes have been reported to move 20 to 100 metres per year, depending on dune size (in general, small dunes move faster than large dunes because their smaller cross-sectional area requires less sand to be transported to reconstitute their form one dune-length downwind).

### DUNE AND SHEET PATTERNS

If the wind were a homogeneous stream of air blowing from one constant direction, long straight dune ridges oriented at right angles to the wind would result. Most dunes, however, are neither straight nor at right angles to the wind, and this indicates that the winds are not a uniform stream or that they blow from different directions. The fairly uniform geometric shapes of several basic types of dunes can be recognized from desert to desert on Earth, and some of the same types have been identified on Mars as well.

Barchan dunes are common to both the Earth and Mars. These small crescent-shaped sand bodies occur in areas where the regional wind blows consistently from one direction (Figure 29). Their crescentic shape must be due to spatial variations in wind velocity, and the regular repetition of dune shapes and spacings when they are close together indicate that the variations in the wind are also regular. This is a property common to all bed forms. It is thought that the flow of a fluid arranges itself in long spiral vortices parallel to the direction of flow, which, with zones of faster and slower velocities arranged transverse to the flow, gives a regular sinuous pattern on the bed.

Where there is a continuous sand cover, a varied dune pattern results from the pattern of flow. The main forms are transverse ridges composed of alternating crescentic elements, like barchans, facing downwind, and other crescentic elements facing upwind. These enclose between them a regular pattern of small hollows. Superimposed on this are small straight ridges parallel with the flow. These elements form a network pattern that is extremely common in the great sand seas. The dunes commonly reach a height of nearly 200 metres and are spaced hundreds of metres to more than two kilometres apart.

One of the important features of sandy terrains is that their forms occur in a number of distinct sizes. Large features are covered with smaller ones, and the smaller ones are covered with ripples (Figure 30). In most of the larger sand seas there is usually a network pattern of very large dunes known as compound dunes, mega-dunes, or *draa*. These are sometimes arranged parallel to the apparent flow, in long ridges, and occasionally transverse to it in great sand waves. The compound dunes are usually covered with a smaller, secondary dune pattern, and the smaller dunes with ordinary sand ripples in most cases. Within each of the size groups of the hierarchy (ripples, dunes, or compound dunes) there are variations in size depending on the grain size of the sand and wind velocity; for example, whereas most ripples are spaced only a few centimetres apart, "mega-ripples," built in very coarse sand, are spaced almost as far apart as small dunes; and whereas most dunes are about 100 metres apart, the low

Compound dunes

By courtesy of the Servicio Aerofotografico Nacional, Lima, Peru

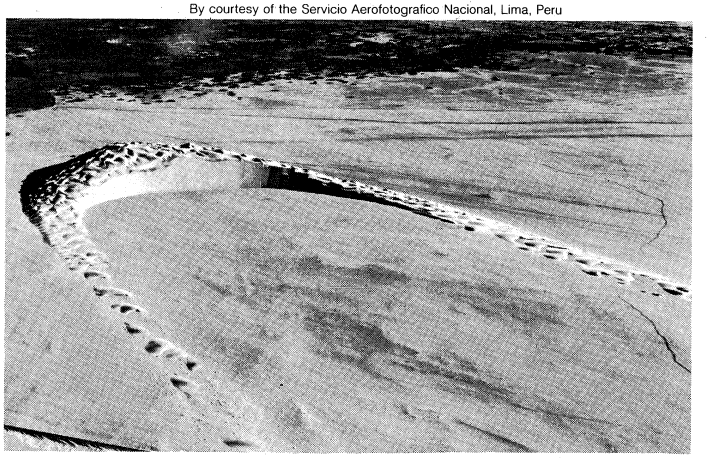


Figure 29: Mega-barchan dune (in Virú Valley, Peru), with small barchan dune migrating from the horns that extend downwind. The wind direction is from the upper left corner of the photograph to the lower right.

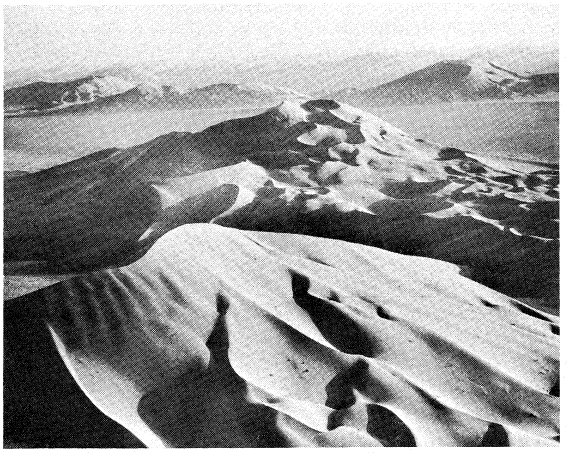


Figure 30: Giant sand mountains in the Rub' al-Khali, southern Saudi Arabia. These dunes attain heights of more than 200 metres; smaller dunes of several types may form on their surfaces.
By courtesy of Arabian American Oil Company

undulations of coarse sand on sand sheets are up to 500 metres apart. The relation between sand grain size and the shape of a dune is not, however, one of simple cause and effect, for the relation is not constant in all dunes of a given shape or in all localities.

Some dune forms can be related to variations in the overall wind direction, usually on a seasonal cycle. In some areas, winds from opposed directions blow during different seasons, so that "reversing dunes" are formed, in which the slip faces face first in one direction and then in the other. Distinct dunes are formed around topographic obstructions and in sheltered zones on the lee of small hills into which the sand migrates. If the wind meets a high scarp or large hill massif, a so-called echo dune is deposited on the upwind side separated from the scarp by a rolling eddy of air that keeps a corridor free of sand. Many oases and routeways are found in this kind of corridor. Echo dunes are among the largest dunes in the desert, sometimes reaching a height of more than 400 metres.

FIXED DUNES IN SEMIARID REGIONS

Dunes also form around plants in the desert where groundwater is available for vegetation. The usual dune forms that occur in such instances are isolated mounds around individual plants. These forms are known as coppice dunes, or *nebkha*. Further, in many regions that are now subhumid or humid, one finds areas of older dunes fixed by vegetation, providing undeniable evidence that these regions were once more arid than they are today. On the North American high plains, in Hungary, and in Mongolia, the fixed sands have a cover of rich grassland. In Poland they are covered with coniferous forests. The dune patterns on these fixed sands bear a close resemblance to those in active sand seas, except that their forms are rounded and subdued. (A.Wa./W.J.Br./C.S.Br.)

## Glacial landforms

Glacial landforms are the product of flowing ice and meltwater. These landforms are forming today in glaciated areas, such as Greenland, Antarctica, and many of the world's higher mountain ranges. In addition, large expansions of present-day glaciers have recurred during the course of Earth history. At the maximum of the last ice age, which ended about 20,000 to 15,000 years ago, more than 30 percent of the Earth's land surface was covered by ice. Consequently, if they have not been obliterated by other landscape-modifying processes since that time, glacial landforms may still exist in regions that were once glaciated but are now devoid of glaciers.

Periglacial features, which form independently of glaciers, are nonetheless a product of the same cold climate that favours the development of glaciers, and so are treated in this section as well.

GENERAL CONSIDERATIONS

Before describing the different landforms produced by glaciers and their meltwater, it seems useful to briefly discuss the glacial environment and the processes responsible for the formation of such landforms.

**Types of glaciers.** There are numerous types of glaciers, but it is sufficient here to focus on two broad classes: mountain, or valley, glaciers and continental glaciers, or ice sheets, (including ice caps). For information about other types, see the article ICE AND ICE FORMATIONS.

Valley glaciers and ice sheets

Generally, ice sheets are larger than valley glaciers. The main difference between the two classes, however, is their relationship to the underlying topography. Valley glaciers are rivers of ice usually found in mountainous regions, whose flow patterns are controlled by the high relief in those areas. In map view, many large valley glacier systems, which have numerous tributary glaciers that join to form a large "trunk glacier," resemble the roots of a plant. Pancakelike ice sheets, on the other hand, are continuous over extensive areas and completely bury the underlying landscape beneath hundreds or thousands of metres of ice. Within continental ice sheets, the flow is directed more or less from the centre outward. At the periphery, however, where ice sheets are much thinner, they may be controlled by any substantial relief existing in the area. In this case, their borders may be lobate on a scale of a few kilometres, with tonguelike protrusions called outlet glaciers. Viewed by themselves, these are nearly indistinguishable from the lower reaches of a large valley glacier system. Consequently, many of the landforms produced by valley glaciers and continental ice sheets are similar or virtually identical, though they often differ in magnitude. Nonetheless, each type of glacier produces characteristic features and thus warrants separate discussion.

**Glacial erosion.** Two processes, internal deformation and basal sliding, are responsible for the movement of glaciers under the influence of gravity (see ICE AND ICE FORMATIONS: *Glaciers*). Glaciers that are cold-based, (*i.e.*, frozen to the ground at their base) move solely by internal deformation of the ice mass. Internal deformation, or strain, results from shear stresses in the ice due to a slope on the surface of the glacier and the weight of the ice. The velocity at the bottom of the glacier is zero and increases upward until it reaches a maximum at the surface. Since there is no relative movement between the ice at the bottom of the glacier and the ground to which it is frozen, it is impossible for any erosion to occur at this interface. Investigators reportedly have found dead lichens emerging at the front of a receding cold-based glacier still attached to the rocks on which they originally grew before being buried for hundreds of years by the glacier during its advance. Although this is an extreme case, it illustrates how little cold-based glaciers may alter a landscape. This situation is rare, however, because few glaciers are entirely or partly cold-based. Cold-based glaciers produce no basal meltwater and therefore no glaciofluvial landforms.

The ice in a warm-based, or temperate, glacier is at the pressure melting point throughout. Since it is not frozen to its substrate, a warm-based glacier generally moves not only by internal strain of the ice mass itself but also by sliding along its base. It is this sliding that enables temperate glaciers to erode their beds and carve landforms. Ice is, however, much softer and has a much lower shear strength than most rocks, and pure ice alone is not capable of substantially eroding anything but unconsolidated sediments. Most temperate glaciers have a basal debris zone from several centimetres to a few metres thick that contains varying amounts of rock debris in transit. In this respect, glaciers are comparable to sandpaper; while the paper itself is too soft to sand wood, the adherent hard grains make it a powerful abrasive system. The analogy ends here, however, for the rock debris found in glaciers is of widely varying sizes—from the finest rock particles to large boulders—and also generally of varied types as it includes the different rocks that a glacier is overriding. For this reason, a glacially abraded surface usually bears many different "tool-marks," from microscopic scratches to gouges centimetres deep and tens of metres long. Over thousands of years glaciers may erode their substrate to

a depth of several tens of metres by this mechanism, producing a variety of streamlined landforms typical of glaciated landscapes.

Glacial plucking

Several other processes of glacial erosion are generally included under the term glacial plucking. This process involves the removal of larger pieces of rock from the glacier bed. Various explanations for this phenomenon have been proposed. Some of the mechanisms suggested are based on differential stresses in the rock caused by ice being forced to flow around bedrock obstacles. These pressures have been shown to be sufficient to fracture solid rock, thus making it available for removal by the ice flowing above it. Other possibilities include the forcing apart of rock by the pressure of crystallization produced beneath the glacier as water derived from the ice refreezes (regelation) or because of temperature fluctuations in cavities under the glacier. Still another possible mechanism involves hydraulic pressures of flowing water known to be present, at least temporarily, under nearly all warm-based glaciers. It is hard to determine which process is dominant because access to the base of active glaciers is rarely possible. Nonetheless, investigators know that larger pieces of rock are plucked from the glacier bed and contribute to the number of abrasive "tools" available to the glacier at its base. Other sources for the rock debris in glacier ice may include rockfalls from steep slopes bordering a glacier or unconsolidated sediments overridden as a glacier advances.

**Glacial deposition.** Debris in the glacial environment may be deposited directly by the ice (till) or, after reworking, by meltwater streams (outwash). The resulting deposits are termed glacial drift.

As the ice in a valley glacier moves from the area of accumulation to that of ablation, it acts like a conveyor belt, transporting debris located beneath, within, and above the glacier toward its terminus or, in the case of an ice sheet, toward the outer margin. Near the glacier margin where the ice velocity decreases greatly is the zone of deposition. As the ice melts away, the debris that was originally frozen into the ice commonly forms a rocky and/or muddy blanket over the glacier margin. This layer often slides off the ice in the form of mudflows. The resulting deposit is called a flow-till. On the other hand, the debris may be laid down more or less in place as the ice melts away around and beneath it. Such deposits are referred to as ablation till. In many cases, the material located between a moving glacier and its bedrock bed is severely sheared, compressed, and "over-compacted." This type of deposit is called lodgment till. By definition, till is any material laid down directly or reworked by a glacier. Typically, it is a chaotic mixture of rock fragments and boulders in a fine-grained sandy or muddy matrix (non-stratified drift). The exact composition of any particular till, however, depends on the materials available to the glacier at the time of deposition. Thus, there are some tills consisting entirely of lake clays deformed by an overriding glacier. Other tills are composed of river gravels and sands that have been "bulldozed" and striated during a glacial advance. Tills often contain some of the tools that glaciers use to abrade their bed. These rocks and boulders bear striations, grooves, and facets, and characteristic till-stones are commonly shaped like bullets or flat-irons. Till-boulders of a rock type different from the bedrock on which they are deposited are dubbed "erratics." In some cases, erratics with distinctive lithologies can be traced back to their source, enabling investigators to ascertain the direction of ice movement of ice sheets in areas where striations either are absent or are covered by till or vegetation.

Tills

Meltwater deposits, also called glacial outwash, are formed in channels directly beneath the glacier or in lakes and streams in front of its margin. In contrast to till, outwash is generally bedded or laminated (stratified drift), and the individual layers are relatively well sorted according to grain size. In most cases, gravels and boulders in outwash are rounded and do not bear striations or grooves on their surfaces, since these tend to wear off rapidly during stream transport. The grain size of individual deposits depends not only on the availability of different sizes of debris but also on the velocity of the depositing current and the distance from the head of the stream. Larger boulders are deposited by rapidly flowing creeks and rivers close to the glacier margin. Grain size of deposited material decreases with increasing distance from the glacier. The finest fractions, such as clay and silt, may be deposited in glacial lakes or ponds or transported all the way to the ocean.

Finally, it must be stressed that most glacier margins are constantly changing chaotic masses of ice, water, mud, and rocks. Ice-marginal deposits thus are of a highly variable nature over short distances, as is much the case with till and outwash as well.

## EROSIONAL LANDFORMS

**Small-scale features of glacial erosion.** Glacial erosion is caused by two different processes: abrasion and plucking (see above). Nearly all glacially scoured erosional landforms bear the tool-marks of glacial abrasion provided that they have not been removed by subsequent weathering (Figure 31). Even though these marks are not large enough to be called landforms, they constitute an integral part of any glacial landscape and thus warrant description here. The type of mark produced on a surface during glacial erosion depends on the size and shape of the tool, the pressure being applied to it, and the relative hardnesses of the tool and the substrate.

*Rock polish.* The finest abrasive available to a glacier is the so-called rock flour produced by the constant grinding at the base of the ice. Rock flour acts like jewelers' rouge and produces microscopic scratches, which with time smooth and polish rock surfaces, often to a high lustre.

*Striations.* These are scratches visible to the naked eye, ranging in size from fractions of a millimetre to a few millimetres deep and a few millimetres to centimetres long. Large striations produced by a single tool may be several centimetres deep and wide and tens of metres long.

Because the striation-cutting tool was dragged across the rock surface by the ice, the long axis of a striation indicates the direction of ice movement in the immediate vicinity of that striation. Determination of the regional direction of movement of former ice sheets, however, requires measuring hundreds of striation directions over an extended area because ice moving close to the base of a glacier is often locally deflected by bedrock obstacles. Even when such a regional study is conducted, additional information is frequently needed in low-relief areas to determine which end of the striations points down-ice toward the former outer margin of the glacier. On an outcrop scale,

Drawing by Gunnar Schlieder

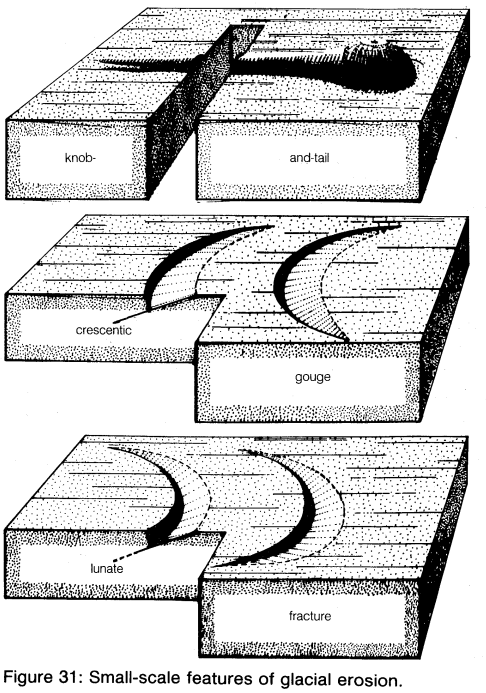



Figure 31: Small-scale features of glacial erosion.

"Chatter marks"

such information can be gathered by studying "chatter marks"—namely, crescentic gouges and lunate fractures. These marks are caused by the glacier dragging a rock or boulder over a hard and brittle rock surface and forming a series of sickle-shaped gouges. Such depressions in the bedrock are steep-sided on their "up-glacier" face and have a lower slope on their down-ice side. Depending on whether the horns of the sickles point up the glacier or down it, the chatter marks are designated crescentic gouges or lunate fractures. Another small-scale feature that allows absolute determination of the direction in which the ice moved is what is termed knob-and-tail. A knob-and-tail is formed during glacial abrasion of rocks that locally contain spots more resistant than the surrounding rock, as, for example, silicified fossils in limestone. After abrasion has been active for some time, the harder parts of the rock form protruding knobs as the softer rock is preferentially eroded away around them. During further erosion, these protrusions protect the softer rock on their lee side and a tail forms there, pointing from the knob to the margin of the glacier. The scale of these features depends primarily on the size of the inhomogeneities in the rock and range from fractions of millimetres to metres.

*P-forms and glacial grooves.* These features, which extend several to tens of metres in length, are of uncertain origin. P-forms (P for plastic) are smooth-walled, linear depressions which may be straight, curved, or sometimes hairpin-shaped and measure tens of centimetres to metres in width and depth. Their cross sections are often semicircular to parabolic, and their walls are commonly striated parallel to their long axis, indicating that ice once flowed in them. Straight P-forms are frequently called glacial grooves, even though the term is also applied to large striations, which, unlike the P-forms, were cut by a single tool. Some researchers believe that P-forms were not carved directly by the ice but rather were eroded by pressurized mud slurries flowing beneath the glacier.

**Erosional landforms of valley glaciers.** Many of the world's higher mountain ranges—*e.g.,* the Alps, the North and South American Cordilleras, the Himalayas, and the Southern Alps in New Zealand, as well as the mountains of Norway, including those of Spitsbergen—are partly glaciated today. During segments of the Pleistocene, such glaciers were greatly enlarged and filled most of the valleys with ice, even reaching far beyond the mountain front in certain places. Most scenic alpine landscapes featuring sharp mountain peaks, steep-sided valleys, and innumerable lakes and waterfalls are a product of several periods of glaciation.

Alpine landscapes

Erosion is generally greater than deposition in the upper reaches of a valley glacier, whereas deposition exceeds erosion closer to the terminus. Accordingly, erosional landforms dominate the landscape in the high areas of glaciated mountain ranges.

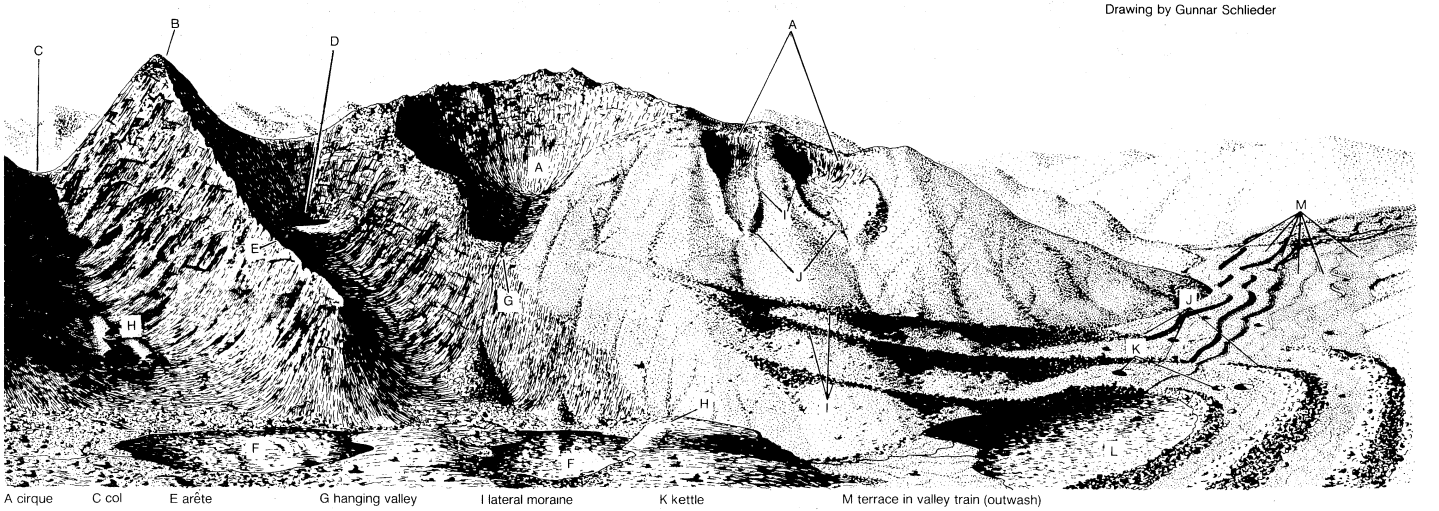
*Cirques, tarns, U-shaped valleys, arêtes, and horns.* The heads of most glacial valleys are occupied by one or several cirques (or corries). A cirque is an amphitheatre-shaped hollow with the open end facing down-valley (Figure 32). The back is formed by an arcuate cliff called the headwall. In an ideal cirque, the headwall is semicircular in plan view. This situation, however, is generally found only in cirques cut into flat plateaus. More common are headwalls angular in map view due to irregularities in height along their perimeter. The bottom of many cirques is a shallow basin, which may contain a lake. This basin and the base of the adjoining headwall usually show signs of extensive glacial abrasion and plucking. Even though the exact process of cirque formation is not entirely understood, it seems that the part of the headwall above the glacier retreats by frost shattering and ice wedging (see below *Periglacial landforms*). The rock debris then falls either onto the surface of the glacier or into the bergschrund—a crevasse between the ice at the head of the glacier and the cirque headwall. The rocks on the surface of the glacier are successively buried by snow and incorporated into the ice of the glacier. Because of a downward velocity component in the ice in the accumulation zone, the rocks are eventually moved to the base of the glacier. At that point, these rocks, in addition to the rock debris from the bergschrund, become the tools with which the glacier erodes, striates, and polishes the base of the headwall and the bottom of the cirque.

Bergschrund

During the initial growth and final retreat of a valley glacier, the ice often does not extend beyond the cirque. Such a cirque glacier is probably the main cause for the formation of the basin scoured into the bedrock bottom of many cirques. Sometimes these basins are "over-deepened" several tens of metres and contain lakes called tarns.

In contrast to the situation in a stream valley, all debris falling or sliding off the sides and the headwalls of a glaciated valley is immediately removed by the flowing ice. Moreover, glaciers are generally in contact with a much larger percentage of a valley's cross section than equivalent rivers or creeks. Thus glaciers tend to erode the bases of the valley walls to a much greater extent than do streams. A river only erodes an extremely narrow line along the lowest part of a valley. The slope of the adjacent valley walls depends on the stability of the bedrock and the angle of repose of the weathered rock debris accumulating at the base of and on the valley walls. For this reason, rivers tend to form V-shaped valleys. Glaciers, which inherit V-shaped stream valleys, reshape them drastically by first removing all loose debris along the base of the valley walls and then preferentially eroding the bedrock along the base and lower sidewalls of the valley. In this way, glaciated valleys assume a characteristic parabolic or U-shaped cross profile, with relatively wide and flat bottoms and steep, even vertical sidewalls. By the same process, glaciers tend to narrow the bedrock divides between the upper reaches of neighbouring parallel valleys

Drawing by Gunnar Schlieder




A cirque B horn C col D tarn E arête F paternoster lake G hanging valley H roche moutonnée I lateral moraine J terminal (end) moraine K kettle L moraine dammed lake M terrace in valley train (outwash)

Figure 32: Erosional and depositional landforms of valley glaciation.

to jagged, knife-edge ridges known as arêtes. Arêtes also form between two cirques facing in opposite directions. The low spot, or saddle, in the arête between two cirques is called a col. A higher mountain often has three or more cirques arranged in a radial pattern on its flanks. Headward erosion of these cirques finally leaves only a sharp peak flanked by near vertical headwall cliffs, which are separated by arêtes. Such glacially eroded mountains are termed horns, the most widely known of which is the Matterhorn in the Swiss Alps.

Horns

*Hanging valleys.* Large valley glacier systems consist of numerous cirques and smaller valley glaciers that feed ice into a large trunk glacier. Because of its greater ice discharge, the trunk glacier has greater erosive capability in its middle and lower reaches than smaller tributary glaciers that join it there. The main valley is therefore eroded more rapidly than the side valleys. With time, the bottom of the main valley becomes lower than the elevation of the tributary valleys. When the ice has retreated, the tributary valleys are left joining the main valley at elevations substantially higher than its bottom. Tributary valleys with such unequal or discordant junctions are called hanging valleys. In extreme cases where a tributary joins the main valley high up in the steep part of the U-shaped trough wall, waterfalls may form after deglaciation, as in Yosemite and Yellowstone national parks in the western United States.

*Paternoster lakes.* Some glacial valleys have an irregular, longitudinal bedrock profile, with alternating short, steep steps and longer, relatively flat portions. Even though attempts have been made to explain this feature in terms of some inherent characteristic of glacial flow, it seems more likely that differential erodibility of the underlying bedrock is the real cause of the phenomenon. Thus the steps are probably formed by harder or less fractured bedrock, whereas the flatter portions between the steps are underlain by softer or more fractured rocks. In some cases, these softer areas have been excavated by a glacier to form shallow bedrock basins. If several of these basins are occupied by lakes along one glacial trough similar to beads on a string, they are called paternoster lakes.

*Roches moutonnées.* These structures are bedrock knobs or hills that have a gently inclined, glacially abraded, and streamlined stoss side (*i.e.,* one that faces the direction from which the overriding glacier impinged) and a steep, glacially plucked lee side. They are generally found where jointing or fracturing in the bedrock allows the glacier to pluck the lee side of the obstacle. In plan view, their long axes are often, but not always, aligned with the general direction of ice movement.

*Rock drumlins.* A feature similar to roches moutonnées, rock drumlins are bedrock knobs or hills completely streamlined, usually with steep stoss sides and gently sloping lee sides. Both roches moutonnées and rock drumlins range in length from several metres to several kilometres and in height from tens of centimetres to hundreds of metres. They are typical of both valley and continental glaciers. The larger ones, however, are restricted to areas of continental glaciation.

Bedrock knobs

**Erosional landforms of continental glaciers.** In contrast to valley glaciers, which form exclusively in areas of high altitude and relief, continental glaciers, including the great ice sheets of the past, occur in high and middle latitudes in both hemispheres, covering landscapes that range from high alpine mountains to low-lying areas with negligible relief. Therefore, the landforms produced by continental glaciers are more diverse and widespread. Yet, just like valley glaciers, they have an area where erosion is the dominant process and an area close to their margins where net deposition generally occurs. The capacity of a continental glacier to erode its substrate has been a subject of intense debate. All of the areas formerly covered by ice sheets show evidence of areally extensive glacial scouring. The average depth of glacial erosion during the Pleistocene probably did not exceed a few tens of metres, however. This is much less than the deepening of glacial valleys during mountain glaciation. One of the reasons for the apparent limited erosional capacity of continental ice sheets in areas of low relief may be the scarcity of tools available to them in these regions. Rocks cannot fall onto a continental ice sheet in the accumulation zone because the entire landscape is buried. Thus, all tools must be quarried by the glacier from the underlying bedrock. With time, this task becomes increasingly difficult as bedrock obstacles are abraded and streamlined. Nonetheless, the figure for depth of glacial erosion during the Pleistocene cited above is an average value, and locally several hundreds of metres of bedrock were apparently removed by the great ice sheets. Such enhanced erosion seems concentrated at points where the glaciers flowed from hard, resistant bedrock onto softer rocks, or where glacial flow was channelized into outlet glaciers (see above).

Glacial scouring

*Streamlined landscape, lakes, and fjords.* As a continental glacier expands, it strips the underlying landscape of the soil and debris accumulated at the preglacial surface due to weathering. The freshly exposed harder bedrock is then eroded by abrasion and plucking. During this process, bedrock obstacles are shaped into streamlined "whaleback" forms, such as roches moutonnées and rock drumlins (see above). The adjoining valleys are scoured into rock-floored basins with the tools plucked from the lee sides of roches moutonnées. The long axes of the hills and valleys are often preferentially oriented in the direction of ice flow. An area totally composed of smooth whaleback forms and basins is called a streamlined landscape.

Streams cannot erode deep basins because water cannot flow uphill. Glaciers, on the other hand, can flow uphill over obstacles at their base as long as there is a sufficient slope on the upper ice surface pointing in that particular direction. Therefore the great majority of the innumerable lake basins and small depressions in formerly glaciated areas can only be a result of glacial erosion. Many of these lakes, such as the Finger Lakes in the state of New York, are aligned parallel to the direction of regional ice flow. Other basins seem to be controlled by pre-glacial drainage systems. Yet, other depressions follow the structure of the bedrock, having been preferentially scoured out of areas underlain by softer or more fractured rock.

A number of the largest freshwater lake basins in the world (*e.g.,* the Great Lakes or the Great Slave Lake and Great Bear Lake in Canada) are situated along the margins of the Precambrian shield of North America. Many researchers believe that glacial erosion was especially effective at these locations because the glaciers could easily abrade the relatively soft sedimentary rocks to the south with hard, resistant crystalline rocks brought from the shield areas that lie to the north. Nonetheless, further research is necessary to determine how much of the deepening of these features can be ascribed to glacial erosion, as opposed to other processes such as tectonic activity and/or pre-glacial stream erosion.

Fjords are found along some steep, high-relief coast lines where continental glaciers formerly flowed into the sea. They are deep, narrow valleys with U-shaped cross sections that often extend inland for tens or hundreds of kilometres and are now partially drowned by the ocean. These troughs are typical of the Norwegian coast, but they also are found in Canada, Alaska, Iceland, Greenland, Antarctica, New Zealand, and southernmost Chile. The floor and steep walls of fjords show ample evidence of glacial erosion. The long profile of many fjords, including alternating basins and steps, is very similar to that of glaciated valleys. Toward the mouth, fjords may reach great depths, as in the case of the Sognafjorden in southern Norway where the maximum water depth is 1,300 metres. At the mouth of a fjord, however, the floor rises steeply and water depths decrease markedly. At the Sognafjorden, the water at this "threshold" is only 150 metres deep, and in many fjords the rock platform is covered by only a few metres of water. The exact origin of fjords is still a matter of debate. While some scientists favour a glacial origin, others believe that much of the relief of fjords is a result of tectonic activity and that glaciers only slightly modified preexisting large valleys. In order to erode Sognafjorden to its present depth, the glacier occupying it during the maximum of the Pleistocene must have been 1,800 to 1,900 metres thick. Such an ice thickness may seem extreme, but even now, during an interglacial period, the

Fjords

Skelton Glacier in Antarctica has a maximum thickness of about 1,450 metres. This outlet glacier of the Antarctic ice sheet occupies a trough, which in places is more than one kilometre below sea level and would become a fjord in the event of a large glacial retreat.

## DEPOSITIONAL LANDFORMS

**Depositional landforms of valley glaciers.** *Moraines.* As a glacier moves along a valley, it picks up rock debris from the valley walls and floor, transporting it in, on, or under the ice. As this material reaches the lower parts of the glacier where ablation is dominant, it is concentrated along the glacier margins as more and more debris melts out of the ice. If the position of the glacier margin is constant for an extended amount of time, larger accumulations of glacial debris (till; see above) will form at the glacier margin. In addition, a great deal of material is rapidly flushed through and out of the glacier by meltwater streams flowing under, within, on, and next to the glacier. Part of this streamload is deposited in front of the glacier close to its snout. There, it may mix with material brought by, and melting out from, the glacier as well as with material washed in from other, nonglaciated tributary valleys. If the glacier then advances or readvances after a time of retreat, it will "bulldoze" all the loose material in front of it into a ridge of chaotic debris that closely hugs the shape of the glacier snout. Any such accumulation of till melted out directly from the glacier or piled into a ridge by the glacier is a moraine. Large valley glaciers are capable of forming moraines a few hundred metres high and many hundreds of metres wide. Linear accumulations of till formed immediately in front of or on the lower end of the glacier are end moraines. The moraines formed along the valley slopes next to the side margins of the glacier are termed lateral moraines. During a single glaciation, a glacier may form many such moraine arcs, but all the smaller moraines, which may have been produced during standstills or short advances while the glacier moved forward to its outermost ice position, are generally destroyed as the glacier resumes its advance. The end moraine of largest extent formed by the glacier (which may not be as extensive as the largest ice advance) during a given glaciation is called the terminal moraine of that glaciation. Successively smaller moraines formed during standstills or small readvances as the glacier retreats from the terminal moraine position are recessional moraines.

Terminal and recessional moraines

*Flutes.* The depositional equivalent of erosional knob-and-tail structures (see above) are known as flutes. Close to the lower margin, some glaciers accumulate so much debris beneath them that they actually glide on a bed of pressurized muddy till. As basal ice flows around a pronounced bedrock knob or a boulder lodged in the substrate, a cavity often forms in the ice on the lee side of the obstacle because of the high viscosity of the ice. Any pressurized muddy paste present under the glacier may then be injected into this cavity and deposited as an elongate tail of till, or flute. The size depends mainly on the size of the obstacle and on the availability of subglacial debris. Flutes vary in height from a few centimetres to tens of metres and in length from tens of centimetres to kilometres, even though very large flutes are generally limited to continental ice sheets.

**Depositional landforms of continental glaciers.** Many of the deposits of continental ice sheets are very similar to those of valley glaciers. Terminal, end, and recessional moraines are formed by the same process as with valley glaciers (see above), but they can be much larger. Morainic ridges may be laterally continuous for hundreds of kilometres, hundreds of metres high, and several kilometres wide. Since each moraine forms at a discreet position of the ice margin, plots of end moraines on a map of suitable scale allow the reconstruction of ice sheets at varying stages during their retreat.

In addition to linear accumulations of glacial debris, continental glaciers often deposit a more or less continuous, thin (less than 10 metres) sheet of till over large areas, which is called ground moraine. This type of moraine generally has a "hummocky" topography of low relief, with alternating small till mounds and depressions. Swamps or lakes typically occupy the low-lying areas. Flutes (see above) are a common feature found in areas covered by ground moraine.

Another depositional landform associated with continental glaciation is the drumlin, a streamlined, elongate mound of sediment. Such structures often occur in groups of tens or hundreds, which are called drumlin fields. The long axis of individual drumlins is usually aligned parallel to the direction of regional ice flow. In long profile, the stoss side of a drumlin is steeper than the lee side. Some drumlins consist entirely of till, while others have bedrock cores draped with till. The till in many drumlins has been shown to have a "fabric" in which the long axes of the individual rocks and sand grains are aligned parallel to the ice flow over the drumlin. Even though the details of the process are not fully understood, drumlins seem to form subglacially close to the edge of an ice sheet, often directly down-ice from large lake basins overridden by the ice during an advance. The difference between a rock drumlin and a drumlin is that the former is an erosional bedrock knob (see above), whereas the latter is a depositional till feature.

Drumlins

**Meltwater deposits.** Much of the debris in the glacial environment of both valley and continental glaciers is transported, reworked, and laid down by water. Whereas glaciofluvial deposits are formed by meltwater streams, glaciolacustrine sediments accumulate at the margins and bottoms of glacial lakes and ponds.

*Glaciofluvial deposits.* The discharge of glacial streams is highly variable, depending on the season, time of day, and cloud cover. Maximum discharges occur during the afternoon on warm, sunny summer days, and minima on cold winter mornings. Beneath or within a glacier, the water flows in tunnels and is generally pressurized during periods of high discharge. In addition to debris washed in from unglaciated highlands adjacent to the glacier, a glacial stream can pick up large amounts of debris along its path at the base of the glacier. For this reason, meltwater streams issuing forth at the snout of a valley glacier or along the margin of an ice sheet are generally laden to transporting capacity with debris. Beyond the glacier margin, the water, which is no longer confined by the walls of the ice tunnel, spreads out and loses some of its velocity. Because of the decreased velocity, the stream must deposit some of its load. As a result, the original stream channel is choked with sediments, and the stream is forced to change its course around the obstacles, often breaking up into many winding and shifting channels separated by sand and gravel bars. The highly variable nature of the sediments laid down by such a braided stream reflects the unstable environment in which they form. Lenses of fine-grained, cross-bedded sands are often interbedded laterally and vertically with stringers of coarse, bouldery gravel. Since the amount of sediment laid down generally decreases with distance from the ice margin, the deposit is often wedge-shaped in cross section, ideally gently sloping off the end moraine formed at that ice position and thinning downstream. The outwash is then said to be "graded to" that particular moraine. In map view, the shape of the deposit depends on the surrounding topography. Where the valleys are deep enough not to be buried by the glaciofluvial sediments, as in most mountainous regions, the resulting elongate, planar deposits are termed valley trains. On the other hand, in low-relief areas the deposits of several ice-marginal streams may merge to form a wide outwash plain, or sandur.

Valley trains and sandurs

If the ice margin stabilizes at a recessional position during glacial retreat, another valley train or sandur may be formed inside of the original one. Because of the downstream thinning of the outwash at any one point in the valley, the recessional deposit will be lower than and inset into the outer, slightly older outwash plain (Figure 33). Flat-topped remnants of the older plain may be left along the valley sides; these are called terraces. Ideally each recessional ice margin has a terrace graded to it, and these structures can be used in addition to moraines to reconstruct the positions of ice margins through time. In some cases where the glacier either never formed moraines or where the moraines were obliterated by the outwash

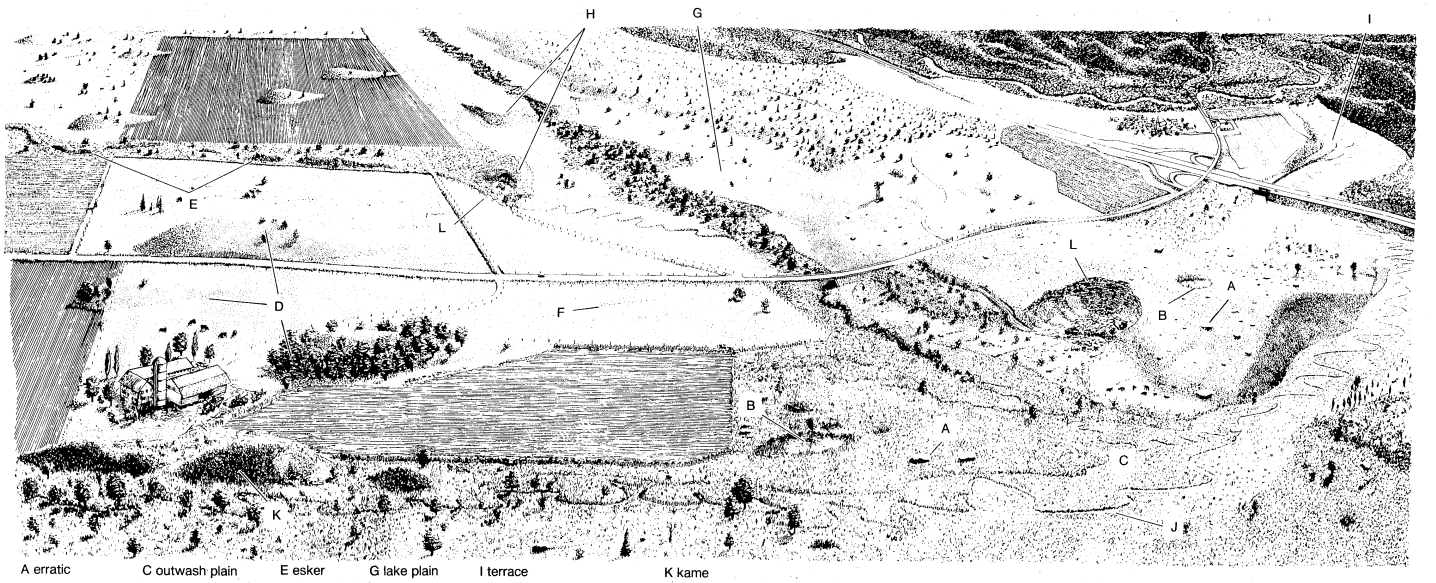



Figure 33: Landforms and features associated with continental glaciers.
Drawing by Gunnar Schlieder

or postglacial erosion, terraces are the only means of ice margin reconstruction.

Streams that flow over the terminus of a glacier often deposit stratified drift in their channels and in depressions on the ice surface. As the ice melts away, this ice-contact stratified drift slumps and partially collapses to form stagnant ice deposits. Isolated mounds of bedded sands and gravels deposited in this manner are called kames. Kame terraces form in a similar manner but between the lateral margin of a glacier and the valley wall. Glacial geologists sometimes employ the term kame moraine to describe deposits of stratified drift laid down at an ice margin in the arcuate shape of a moraine. Some researchers, however, object to the use of the term moraine in this context because the deposit is not composed of till.

In some cases, streams deposit stratified drift in subglacial or englacial tunnels. As the ice melts away, these sinuous channel deposits may be left as long linear gravel ridges called eskers. Some eskers deposited by the great ice sheets of the Pleistocene can be traced for hundreds of kilometres, even though most esker segments are only a few hundred metres to kilometres long and a few to tens of metres high.

Eskers

Kettles, potholes, or ice pits are steep-sided depressions typical of many glacial and glaciofluvial deposits. Kettles form when till or outwash is deposited around ice blocks that have become separated from the active glacier by ablation. Such "stagnant" ice blocks may persist insulated under a mantle of debris for hundreds of years. When they finally melt, depressions remain in their place, bordered by slumped masses of the surrounding glacial deposits. Many of the lakes in areas of glacial deposition are water-filled kettles and so are called kettle lakes. If a sandur or valley train contains many kettles, it is referred to as a pitted outwash plain.

*Glaciolacustrine deposits.* Glacial and proglacial lakes are found in a variety of environments and in considerable numbers. Erosional lake basins have already been mentioned, but many lakes are formed as streams are dammed by the ice itself, by glacial deposits, or by a combination of these factors. Any lake that remains at a stable level for an extended period of time (*e.g.,* hundreds or thousands of years) tends to form a perfectly horizontal, flat, terracelike feature along its beach. Such a bench may be formed by wave erosion of the bedrock or glacial sediments that form the margin of the lake, and it is called a wave-cut bench. On the other hand, it may be formed by deposition of sand and gravel from long-shore currents along the margin of the lake, in which case it is referred to as a beach ridge. The width of these shorelines varies from a few metres to several hundred metres. As the lake level is lowered due to the opening of another outlet or downcutting of the spillway, new, lower shorelines may be formed. Most former or existing glacial lakes (*e.g.,* the Great Salt Lake and the Great Lakes in North America) have several such shorelines that can be used both to determine the former size and depth of now-extinct or shrunken lakes and to determine the amount of differential postglacial uplift because they are now tilted slightly from their original horizontal position.

Where a stream enters a standing body of water, it is forced to deposit its bedload. The coarser gravel and sand are laid down directly at the mouth of the stream as successive, steeply inclined foreset beds. The finer, suspended silt and clay can drift a bit farther into the lake, where they are deposited as almost flat-lying bottomset beds. As the sediment builds out farther into the lake (or ocean), the river deposits a thin veneer of subhorizontal gravelly topset beds over the foreset units. Because the foreset–topset complex often has the shape of a triangle with the mouth of the stream at one apex, such a body of sediment is called a delta. Many gravel and sand pits are located in deltas of former glacial lakes.

The flat-lying, fine-grained bottomset beds of many large former glacial lakes filled in and buried all of the preexisting relief and are now exposed, forming perfectly flat lake plains. Cuts into these sediments often reveal rhythmically interbedded silts and clays. Some of these so-called rhythmites have been shown to be the result of seasonal changes in the proglacial environment. During the warmer summer months, the meltwater streams carry silt and clay into the lakes, and the silt settles out of suspension more rapidly than the clay. A thicker, silty summer layer is thus deposited. During the winter, as the surface of the lake freezes and the meltwater discharge into it ceases, the clays contained in the lake water slowly settle out of suspension to form a thin winter clay layer. Such lacustrine deposits with annual silt and clay "couplets" are known as varves.

Lake plains

## PERIGLACIAL LANDFORMS

In the cold, or periglacial, areas adjacent to and beyond the limit of glaciers, a zone of intense freeze-thaw activity produces periglacial features and landforms. This happens because of the unique behaviour of water as it changes from the liquid to the solid state. As water freezes, its volume increases about 9 percent and can, if confined in a crack or pore space, exert pressures of about 200,000 kilopascals (29,000 pounds per square inch). This is enough to break the enclosing rock. Thus freezing water can be a powerful agent of physical weathering. If multiple freeze-and-thaw cycles occur, the growth of ice crystals fractures and moves material by means of frost shattering and frost

heaving, respectively. In addition, in permafrost regions (see below) where the ground remains frozen all year, characteristic landforms are formed by perennial ice.

Features associated with intense freeze-thaw activity

**Felsenmeers, talus, and rock glaciers.** In nature, the tensional strength of most rocks is exceeded by the pressure of water crystallizing in cracks. Thus repeated freezing and thawing not only forms potholes in poorly constructed roads but is also capable of reducing exposed bedrock outcrops to rubble. Many high mountain peaks are covered with frost-shattered angular rock fragments. A larger area blanketed with such debris is called a felsenmeer, from the German word for "sea of rocks." The broken rock fragments can be transported downslope by flowing water or fall off the cliff from which they were wedged by the ice. Accumulations of this angular debris at the base of steep slopes are known as talus. Due to the steepness of the valley sides of many glacial troughs, talus is commonly found in formerly glaciated mountain regions. Talus cones are formed when the debris coming from above is channelized on its way to the base of the cliff in rock chutes. As the talus cones of neighbouring chutes grow over time, they may coalesce to form a composite talus apron.

In higher mountain regions, the interior of thick accumulations of talus may remain at temperatures below freezing all year. Rain or meltwater percolating into the interstices between the rocks freezes over time, filling the entire pore space. In some cases, enough ice forms to enable the entire mass of rock and ice to move downhill like a glacier. The resulting massive, lobate, mobile feature is called a rock glacier. Some rock glaciers have been shown to contain pure ice under a thick layer of talus with some interstitial ice. These features may be the final retreat stages of valley glaciers buried under talus.

Rock glaciers

**Permafrost, patterned ground, solifluction deposits, and pingos.** Permafrost is ground that remains perennially frozen (see ICE AND ICE FORMATIONS: *Permafrost*). It covers about 20–25 percent of the Earth's land surface today. The "active layer" of soil close to the surface of permafrost regions undergoes many seasonal and daily freeze-thaw cycles. The constant change in the volume of water tends to move the coarser particles in the soil to the surface. Further frost heaving arranges the stones and rocks according to their sizes to produce patterned ground. Circular arrangements of the larger rocks are termed stone rings. When neighbouring stone rings coalesce, they form polygonal stone nets. On steeper slopes, stone rings and stone nets are often stretched into stone stripes by slow downhill motion of the soggy active layer of the permafrost. In other areas, patterned ground is formed by vertical or subvertical polygonal cracks, which are initiated in the soil by contraction during extremely cold winters. During the spring thaw of the active layer, water flows into these cracks, freezes, and expands. This process is repeated year after year, and the ice-filled cracks increase in size. The resulting ice wedges are often several metres deep and a few tens of centimetres wide at the top. Along the sides of ice wedges the soil is deformed and compressed. Because of this disturbance and sediment that may be washed into the crack as the ice melts, relict patterned ground may be preserved during a period of warmer climate long after the permafrost has thawed. Today, relict patterned ground that formed during the last ice age exists more than 1,000 kilometres to the south of the present limit of permafrost.

When the active layer of permafrost moves under the influence of gravity, the process is termed gelifluction. The soft flowing layer is often folded and draped on hillsides and at the base of slopes as solifluction, or gelifluction, lobes.

In some permafrost areas, the ice tends to accommodate the volume increase during freezing by pushing up ice-cored, circular mounds called pingos. These mounds may be several tens of metres high and hundreds of metres in diameter. (E.B.E./Gu.S.)

Pingos

## Caves and karst landscape

Caves are natural openings in the Earth large enough for human exploration. Such cavities, also known as caverns, are formed in many types of rock and by many processes. The largest and most common caves are those formed by chemical reaction between circulating groundwater and bedrock composed of limestone or dolomite. These caves, called solution caves, typically constitute a component of what is known as karst terrain. Named after the Karst region of Yugoslavia located just north of the Adriatic coast, karst terrain in general is characterized by a rough and jumbled landscape of bare bedrock ledges, deranged surface drainage, and sinkholes (funnellike depressions), as well as caves. It should be noted, however, that there is considerable variation among karst areas. Some may have dramatic surface landforms but few caves. By contrast, others may have extensive cave development with little surface expression; for example, the Guadalupe Mountains of New Mexico, the site of Carlsbad Caverns and various other caves, have very few surface karst features.

Association with karst terrain

Karst landscapes are formed by the removal of bedrock (composed in most cases of limestone, dolomite, gypsum, or salt, but in some cases of such normally insoluble rocks as quartzite and granite) in solution through underground routes rather than through surface weathering and surface streams. As a result, much karst drainage is internal. Rainfall flows into closed depressions and down their drains. Further dissolution in the subsurface forms continuous conduits that serve as integrated drains for the rapid movement of underground water. The outlets for the water-carrying conduits often are springs of majestic size. Caves are fragments of such conduit systems and some of them provide access to active streams. These caves may be completely water-filled; others are dry passages left behind by streams that cut to lower levels. Surface streams flowing from areas underlain by insoluble rock often sink when they reach the border of a karst region. These sinking streams form tributaries of the underground drainage system.

### CAVE TYPES

Not all caves are part of karst landscapes. As discussed above, a substantial number of relatively small caves are formed in lava and by the mechanical movement of bedrock. Other caves are formed in glaciers by the melting of ice. Still others are created by the erosive action of water and wind or from the debris of erosive processes; these are sea caves, eolian caves, rock shelters, and talus caves.

Non-solution caves

**Glacier caves.** These are long tunnels formed near the snouts of glaciers between the glacial ice and the underlying bedrock. Meltwater from the surface of a glacier drains downward through crevasses, which are enlarged to form shafts leading to the base of the glacier. Because the inlet water is slightly above the melting point of ice, it gradually melts the ice as it seeps along the base of the glacier.

Glacier caves may reach lengths of several kilometres. Mature caves of this sort are tubular conduits, often with intricately sculptured walls. Some of them have a branching pattern. The floors of glacier caves usually consist of rock. Most glacier caves can be explored only when the surface is frozen; at other times they are filled with water.

**Sea caves, eolian caves, rock shelters, and talus caves.** Sea caves are formed by wave action on fractures or other weaknesses in the bedrock of sea cliffs along coastlines. They may be mere crevices in the cliff or roomy chambers. Some can only be entered by boat at low tide, while others, occurring along beaches, can be walked into. A sea cave may have an opening to the surface at its rear that provides access from the top of the cliff. In some cases, the ceiling entrance serves as a blowhole from which water spouts during times of high tide or rough seas. Sea caves rarely are more than a few hundred metres long.

Eolian caves are chambers scoured by wind action. They are common in desert areas where they are formed in massive sandstone cliffs. Wind sweeping around such a cavity erodes the walls, floor, and ceiling, resulting in a bottle-shaped chamber usually of greater diameter than the entrance. Eolian caves are rarely longer than a few tens of metres.

Rock shelters are produced by bedrock erosion in insoluble rocks. A common setting is where a resistant rock such as a sandstone overlies shale or some other relatively weak rock. Surface weathering or stream action wears away the

shale, cutting it back into the hillside. The sandstone is left behind as a roof to the rock shelter. Rock shelters are minor features as caves, but many are important archaeological or historical sites.

Talus caves are openings formed between boulders piled up on mountain slopes. Most of them are very small both in length and in cross section. Some boulder piles, however, do have explorable interconnected "passages" of considerable length. Some of the largest talus caves occur among granite blocks in New York and New England, where integrated systems of passages between boulders have been mapped to lengths of several kilometres.

**Solution caves.** As previously noted, the largest and most common caves are those formed by dissolution of limestone or dolomite. Limestone is composed mostly of calcium carbonate in the form of the mineral calcite. Dolomite rock consists of calcium magnesium carbonate, the mineral dolomite. Both these carbonate minerals are somewhat soluble in the weak acids formed by carbon dioxide dissolving in groundwater. Water seeping through soils into the bedrock, water collected by sinkholes, and surface streams sinking underground at the margins of karst areas all percolate along fractures in the bedrock and gradually create sizable passages by chemical action. Because the dissolution process takes place deep in the bedrock, it is not necessary that solution caves have entrances. Most entrances are formed by accidental processes such as the downcutting of surface valleys, the collapse of sinkholes, or the emplacement of quarries or road cuts. Accidental processes of passage collapse and passage plugging divide caves into smaller fragments. Because of this, there are many more small caves than large ones. The longest known cave is the Mammoth Cave–Flint Ridge system in south central Kentucky, which had (as of 1986) a surveyed length of nearly 500 kilometres.

Limestone and dolomite caves

Most solution caves form at relatively shallow depths (from a few tens of metres to 1,000 metres) by the action of water rich in carbonic acid ($H_2CO_3$) derived from recent rainfall. Some solution caves, however, appear to have been formed by deep-seated waters such as oil field brines. Sources of acid other than carbonic acid (*e.g.*, sulfuric acid from the oxidation of sulfide minerals or the oxidation of hydrogen sulfide-bearing fluids) may be the dissolving agent for such caves. According to some investigators, Carlsbad Caverns originated from dissolution with sulfuric acid.

Gypsum rock, composed primarily of calcium sulfate dihydrate (the mineral gypsum) is more soluble than limestone. Outcrops of gypsum rock are found at the land surface in arid regions such as West Texas, western Oklahoma, and eastern New Mexico. Caves formed by the dissolution of gypsum are much like limestone caves in the size, shape, and pattern of their passages. The Optimisticheskaya Cave in the Soviet Union is the world's longest gypsum cave, with 153 kilometres of passage.

Gypsum and salt caves

Caves also are formed by the dissolution of salt (the mineral halite). Because it is highly soluble in water, salt outcrops at the land surface only in extremely arid regions. Caves in salt closely resemble limestone caves in passage plan and shape. In most cases, salt caves are small, with passage lengths ranging from a few tens of metres to several hundred metres. Good examples of salt caves occur in Mount Sedom in Israel and in eastern Spain.

## EVOLUTION AND DEMISE OF SOLUTION CAVES

Compared with most geologic phenomena, caves are transient features of the landscape. They form, evolve, and are destroyed over periods of time ranging from a few tens of thousands to a few million years. It is possible to sketch the "life history" of a single cave passage as the sequence from an initiation phase, a series of three critical thresholds, an enlargement phase, a stagnation phase, and a decay phase (Figure 34).

**Initiation phase.** Since limestone is an impermeable rock, groundwater moves mainly through mechanical fractures—joint and bedding-plane partings. Because groundwater seeps slowly through these openings, it becomes nearly saturated with dissolved calcium carbonate, particularly deep in the rock mass. As a result, the ability of the water to further dissolve the limestone is limited, and the fractures thus enlarge very slowly. Calculations show that times on the order of 3,000 to 10,000 years are needed to enlarge a fracture from an initial width of 10 to 50 micrometres to pencil-sized openings five to 10 millimetres wide. When a continuous pathway from the water source to the outlet has been enlarged to five to 10 millimetres width, the initiation phase is complete.

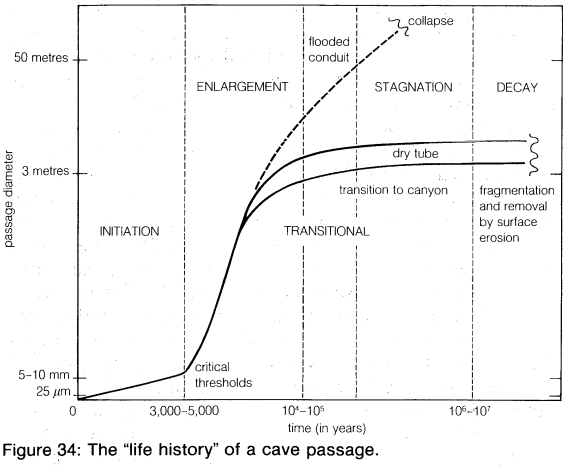



Figure 34: The "life history" of a cave passage.

The five- to 10-millimetre size of the enlarging fracture marks a set of thresholds where new processes come into play. The slow, percolating flow of water is accelerated as the conduit becomes larger, and at the threshold size turbulence appears in the flowing water. The flow pattern is less like percolation through an aquifer and more like flow in a pipe. At the threshold size the opening is large enough and the flow velocities high enough that insoluble sediments can be transported. For the complete development of an underground drainage system, it is necessary that the water-carrying conduits also flush out the soil that washes in through sinkholes, the sediment load of sinking streams, and the insoluble weathering products from the dissolution of the limestone. Another threshold has to do with the rate at which the limestone is dissolved. During the initiation phase when flow velocities are low and the water is nearly saturated, the rate at which limestone is removed is very slow. As velocities increase, unsaturated water moves deep into the bedrock, and the rate of dissolution is greatly increased. The pencil-sized threshold opening marks the boundary between the initial fracture system and the evolving conduit system.

Critical thresholds

**Enlargement phase.** Once a complete pathway has been opened to threshold size, enlargement takes place rapidly as the conduit provides an efficient route for groundwater flow. Enlargement from threshold size to a full-scale cave passage of one to three metres in diameter can be accomplished in 10,000 to 100,000 years, depending on local geology. During the enlargement phase, the conduit may become completely water-filled, in which case the growing passage takes the form of a circular or elliptical conduit as dissolution acts uniformly on the floor, walls, and ceiling. If the water source feeding the conduit is limited, a time will come when there is not enough water to fill the passage. A free air surface then develops and the dissolution of the ceiling will cease, even though the passage will continue to enlarge through dissolution of the lower walls and floor. This transition from pipe flow to open-channel flow results in a change in passage shape from that of an elliptical tube to that of a canyon. Continued solutional erosion causes the canyon to deepen, resulting in canyon passages 30 to 50 metres high and only one metre or less wide.

Transition from pipe flow to open-channel flow

The fate of a cave passage at the end of the enlargement stage depends on what has been happening elsewhere on the land surface and in the drainage basin. If the passage lies deep below the water table, enlargement will continue until the passage becomes too wide for the ceiling bedrock to support its own weight and the passage will ultimately collapse. During the time that the cave passage has been enlarging, surface streams have been downcutting their

beds, and the position of base level and the water table is lowered. If the original water source continues to flow through the cave after the transition to canyon shape, the underground canyon can continue to deepen, keeping its gradient adjusted to the lowering surface streams. Sometimes, however, the conduit passages are simply abandoned. Veneers of insoluble sediment that accumulate on the floors of cave passages tend to protect them from solution. As surface streams downcut, the conduits are left behind and the increased hydraulic gradient causes new passages to form at lower levels. In due course, the flow is completely diverted into these new passages, and the original passages remain air-filled and dry above the descending water table.

**Stagnation and decay phases.** Segments of cave passage abandoned as surface streams downcut can survive for a long time in a stage of stagnation. Truncation of the passages by valley downcutting produces entrances. Caves in the stagnation phase are those most frequently discovered and explored by humans.

Surface erosion continues to dissect the landscape, and hilltops and plateaus are lowered. The underlying cave passages are cut into smaller and smaller fragments. Eventually the denudation of the land surface destroys the last vestiges of the passages, bringing to an end the long history of the cave conduit.

The time scales for the stagnation and decay stages are highly variable, depending on local geologic conditions. Paleomagnetic measurements of the sediments in Mammoth Cave show that the passages at the highest elevations are at least 2,000,000 years old. Studies based on rates of surface weathering in the Appalachian valleys of Pennsylvania indicate that caves at the highest elevation in the residual hills may be 2,000,000 to 3,000,000 years old.

Larger cave systems often have complex patterns of superimposed passages that represent a long history of cave development. The oldest passages, usually but not necessarily those at the highest elevations, may have formed before the glaciations of the Quaternary. The youngest passages may be part of an integrated subsurface drainage system that exists today.

## GEOMORPHIC CHARACTERISTICS OF SOLUTION CAVES

Like many other geologic features concealed beneath the earth, caves are difficult to observe. One cannot really see a cave, even though one may have a point-by-point, cross-sectional view as the cave passage is illuminated during exploration. The horizontal ground plans and vertical profiles of caves must be represented by maps. These in turn are constructed from arduous station-to-station surveys by cave explorers.

Variations in cave-passage plans

Some cave-passage plans take the shape of linear, angulate, or sinuous segments of conduit. These are segments of drainage trunk without tributaries. Other cave-passage plans are branchworks. There may be a well-defined "upstream" direction, with tributary passages joining the trunk. Still other passage plans are networks in which passages are laid out in a "city-block" pattern with many intersecting passages and many closed loops. In terms of flow pattern, a single-conduit type of cave forms where much of the original catchment area was on non-karstic borderlands and the sinking stream injected large quantities of water at a single point. Branchwork caves develop where there are multiple inlets, each at the head of one of the tributary branches. Network caves are formed where flows are controlled by diffuse inlets; flow velocities remain low and solutional erosion takes place along all possible joint openings. A network cave is the underground equivalent of a swamp.

Passage cross-sectional shapes reflect the way the water flowed through the cave and the way in which the water dissolved the bedrock. Passages that formed while completely flooded are dissolved away equally on walls, ceilings, and floors. The result is an elliptical tube. In contrast, a flowing stream with a free air surface can dissolve limestone only in its bed. The result is a canyon-shaped passage. In some caves of this type, the walls are nearly vertical and may measure 30 to 50 metres high, even though the passage may only be one metre wide. Other cave passages are very irregular because of the meanderings of the downcutting stream. There is always competition between the hydraulics of flowing water that works to shape passages into smooth, streamlined forms and the control of passage shape by the structural arrangement of joints, fractures, and bedding-plane partings that initiated the passageway. Joint-controlled passages may be high and narrow, sometimes with irregular walls; such a configuration resulted as the passages were enlarged from the initial joint by slowly percolating water. Passages developed primarily along bedding-plane partings may be low and wide. In general, higher flow velocities favour the hydraulic forms, and slow, percolating flows tend to preserve the shapes of the initial mechanical openings.

Most passages of solution caves are nearly horizontal with gentle average slopes toward the outlet springs. If the caves were formed by pressure flow beneath the water table, the passage profile can be irregular, with both downsloping and upsloping segments. Most cave passages are not graded like surface streams. Only in some alpine environments do caves form with steeply sloping passages. Continuous lowering of the level of groundwater circulation often produces tiers of passages stacked one on top of the other, and these need not be interconnected by an explorable cave. Additional mechanisms are needed to explain the vertical arrangement of some caves that may have an internal relief from tens of metres to more than 1,000 metres. Vertical integration is accomplished by some combination of the following: (1) primary vertical solution in the unsaturated zone above the water table during the same time as conduit dissolution below the water table, (2) dissolution of vertical shafts and solution chimneys (see below) in the unsaturated zone at some time after the development of the conduits, or (3) interconnection of existing dry passages by processes of breakdown and collapse.

Vertical integration

Caves in regions of high relief are frequently developed by inputs of water that move by predominantly vertical paths through the unsaturated zone. Such caves often have a stair-step pattern, with vertical pits and shafts offset by short reaches of horizontal passage. Steeply sloping streamways are common. Some caves of the unsaturated zone are simply pits tens to hundreds of metres deep, which show little horizontal development. Others make up complicated cave systems in which many vertical infeeders join to form master streams that descend to base level as waterfalls plunging down pits. One of the largest such systems is the group of caves on the Huautla Plateau in Mexico. The greatest relief from the highest known entrance of the Sistema Huautla to the lowest point of exploration is 1,252 metres in a cave measuring 33.8 kilometres long (1985 data).

Some conduit systems such as those of the Mammoth Cave area and of the Cumberland Plateau of the Appalachian Mountains develop beneath a protective cap of sandstone, shale, and other relatively non-soluble rocks. As the caprock erodes, the underlying limestone is exposed to the runoff water that drains from the remaining area of the plateau. Such runoff water dissolves away the limestone in the unsaturated zone to form solution chimneys and vertical shafts. Solution chimneys develop along vertical fractures or along bedding planes of vertically bedded limestones. In cross section, they tend to be irregular and elongated along the controlling fracture or bedding plane. Solution chimneys follow the fracture and may be offset or descend at steep angles, depending on the pitch of the guiding fracture. Vertical shafts, by contrast, are controlled by the hydraulic forces of freely flowing water. They are often nearly perfect cylinders with circular cross sections. The walls are vertical and cut across the limestone beds with complete disregard for angle or composition of the beds. Vertical shafts and solution chimneys have no direct relation to the conduit system, especially not to the upper dry levels of the system. They sometimes are connected to present-day active horizontal conduits by drain passages. These drains usually are of small cross section and may extend from hundreds of metres or even several kilometres before connecting with the main drainage conduits. In the unsaturated zone, vertical shafts tend to shear

Solution chimneys and vertical shafts

through high-level passages as though they were not there. When vertical shafts and solution chimneys cut through several tiers of overlying horizontal passage, they provide pathways for exploration and integrate the cave system (Figure 35).

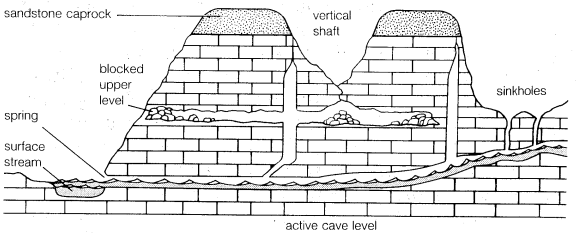



Figure 35: Components of an integrated cave system.

Cave roofs are always in a state of stress. The weight of the ceiling beds causes them to sag slightly, separating the beds along the weaker bedding planes. Each ceiling bed becomes, in effect, a fixed beam spanning the width of the cave passage. There is a strict mechanical relationship between the thickness of a beam, its density, and the width of the span. When the width of the span exceeds a certain critical value, the cave ceiling will collapse under its own weight. Processes such as solution along vertical joints cut the ceiling beams, turning them into cantilevers that have much smaller critical loads. When one ceiling bed falls, support is removed from the bed above and it also may fall. There is thus a process of upward stoping due to ceiling collapse. Upward breakdown and collapse can cause one passage to migrate up into an overlying one.

## SOLUTION CAVE FEATURES

**Solutional sculpturings.** Superimposed on the walls of cave passages are many small solutional sculpturings that record further details of water flow. Pockets of various sizes and kinds are cut back into the walls and ceiling. Some of these have ax-blade shapes and form where water seeping into the cave passage is mixed with the water already in the passage. If the seepage water and the passage water have the correct chemistry, corrosive water forms in the mixing zone and dissolves away the joint-controlled wall and ceiling pockets. Other wall and ceiling pockets are rounded kettle holes or circular cylinders that extend into the solid bedrock of the ceiling with no obvious influence from joints. The ceilings of tropical caves often contain large numbers of the cylindrical cavities, which are used as roosting places by bats. Small secondary channels are carved into the floors or ceilings by flowing water. Floor channels provide evidence of the presence of small later-stage streams that occupied the cave passage after it had been drained of its original flow. Ceiling channels are thought to be the result of upward solutional erosion by cave streams that occurred when the main channel was completely filled with clays, sand, and gravel.

Wall and ceiling pockets

Among the most significant of the solutional sculpturings are the small scooplike depressions known as scallops. Scallops vary in size from a few centimetres to more than one metre. They are asymmetrical in cross section, having a steep wall on the upstream side and a gentler slope on the downstream side. Scallops thus provide information as to the direction of water flow in passages that have been dry for hundreds of thousands of years. The size of a scallop is inversely proportional to the flow velocity of water in the passage. As a consequence, scallops serve not only as paleo-direction indicators but also as paleo-flow meters. Scallops that are a few centimetres wide indicate flow velocities on the order of a few metres per second. The largest scallops, those that are more than one metre wide, indicate flow velocities of a few centimetres per second.

Scallops

The flow velocity of conduit water is sufficient to transport clastic sediment through a cave system. The clastic material is derived from borderlands where it is carried into the karst by sinking streams, from overlying sandstone and shale caprock, from surface soils that are washed underground through sinkholes, and from the insoluble residue of the limestone bedrock. Some of these clastic materials are deposited in caves where they remain as clay, silt, and sand on the cave floors. Some drainage systems carry larger cobble- and boulder-sized materials that are often found in cave streambeds. Most caves have undergone several periods of deposition and excavation, and so remnant beds and pockets of sediment have been left high on cave walls and ledges. These sediments contain iron-bearing magnetic particles, which indicate the position of the Earth's magnetic field at the time when the sediments were deposited. The age of the sedimentary deposits can be determined by measuring the paleomagnetic record in cave sediments and correlating it with the established geomagnetic polarity time scale. Using this method, investigators have ascertained that the age of the sediments in Mammoth Cave is more than 2,000,000 years.

**Depositional materials and features.** There are three broad categories of sedimentary material found in caves: clastic sediments carried in by streams and infiltrated from the surface; blocks, slabs, and fragments of breakdown derived from the local bedrock; and chemical sediments deposited in the cave by percolating waters. The chemical sediments are the most diverse and are responsible for the decorative beauty of many caves.

The most common of the secondary chemical sediments is calcite, calcium carbonate. There also occurs a less common form of calcium carbonate, the mineral aragonite. The second most common cave mineral is gypsum, calcium sulfate dihydrate. Other carbonate, sulfate, and oxide minerals are occasionally found in caves as well. Many of these require that the cave be associated with ore deposits or with other special geologic environments. For this reason, of the more than 200 mineral species known to occur in caves, only about 20 are found widely.

Deposits of cave minerals occur in many forms, their shapes determined by whether they were deposited by dripping, flowing, or seeping water or in standing pools of water. Collectively, these secondary mineral forms are known as speleothems.

Water emerging from a joint in the cave ceiling hangs for a while as a pendant drop. During this time, a small amount of calcium carbonate is deposited in a ring where the drop is in contact with the ceiling. Then the drop falls, and a new drop takes its place, also depositing a small ring of calcium carbonate. In this manner, an icicle-like speleothem called a stalactite is built up. Stalactites vary in shape from thin strawlike features to massive pendants or drapery-like forms (Figure 36). Stalactites have a central canal that carries water from the feeder joint to the stalactite tip. When the drops fall to the floor of the cave, additional mineral matter is deposited and stalagmites are built up. Stalagmites also take on many forms, from slender broom-handle to mound- and pagoda-like shapes (Figure 36). Stalagmites consist of superimposed caps or layers and do not have a central canal. Stalactites may grow so large that they cannot support their own weight; the broken fragments of large stalactites are sometimes found in caves. Stalagmites are not so restricted and can reach heights of tens of metres. Water flowing along ledges and down walls leaves behind sheets of calcite, which build up a massive deposit known as a flowstone.

Stalactites and stalagmites

Most flowstone deposits are composed of calcite, though other minerals occasionally are present. The calcite is usually coarsely crystalline, densely packed, and coloured various shades of tan, orange, and brown. Some of the pigment is from iron oxides carried into the deposit by the seepage water, but the more common colouring agent is humic substances derived from overlying soils. Humic substances are the organic products of plant decay, which are also responsible for the brown colour of some soils and for the tealike colour of some swamp and lake waters. Calcite speleothems may be pure white but appear milky because of many tiny inclusions of water within the structure.

Flowstone

The calcite in speleothems is derived from the overlying limestone near the bedrock/soil interface. Rainwater infiltrating through the soil absorbs carbon dioxide from the carbon dioxide-rich soil and forms a dilute solution of carbonic acid. When this acid water reaches the base of the soil, it reacts with the calcite in the limestone bedrock and takes some of it into solution. The water continues



Figure 36: Interior of a solution cave decorated with stalactites and stalagmites.
William B. White

its downward course through narrow joints and fractures in the unsaturated zone with little further chemical reaction. When the water emerges from the cave roof, carbon dioxide is lost into the cave atmosphere and some of the calcium carbonate is precipitated. The infiltrating water acts as a calcite pump, removing it from the top of the bedrock and redepositing it in the cave below.

Caves provide a very stable environment where temperature and relative humidity may remain constant for thousands of years. The slow growth of crystals is not interrupted, and some speleothems have shapes controlled by the forces of crystal growth rather than by the constraints of dripping and flowing water. Speleothems known as helictites are much like stalactites in that they have a central canal and grow in long tubular forms. They twist and turn in all directions, however, and are not guided by the gravitational pull on pendant water drops. Another variety of speleothem, the anthodite, is a radiating cluster of needlelike crystals. Anthodites are usually composed of aragonite, which has a different habit (*i.e.,* shape of individual crystal grains) than the more common variety of calcium carbonate, calcite. Layered bead or corallike forms occur on cave walls, and complex arrangements of crystals are found in cave pools. Pools of water saturated with calcium carbonate have the remarkable property of surrounding themselves with rimstone dams of precipitated calcite.

Helictites and anthodites

Gypsum and other more water soluble sulfate minerals such as epsomite (magnesium sulfate heptahydrate) and mirabilite (sodium sulfate decahydrate) grow from seepage waters in dry caves. Deposition of the sulfate minerals is due to evaporation of the mineral-bearing solutions. These minerals occur as crusts and in the form of radiating, curving masses of fibrous crystals known as gypsum flowers. Because of their higher solubility, sulfate minerals either do not occur or are destroyed in damp or wet caves.

Gypsum flowers

## KARST TOPOGRAPHY

As previously noted, karst landscapes owe their existence to the removal of bedrock in solution and to the development of underground drainage without the development of surface stream valleys. Within these broad constraints, karst landscapes show much variation and are usually described in terms of a dominant landform. Most important with respect to worldwide occurrence are fluviokarst, doline karst, cone and tower karst, and pavement karst.

Common types of karst landscapes

**Fluviokarst.** In this type of karst landscape, the pattern of surface stream channels and stream valleys are still in evidence, though much of the drainage may be underground. Tributary surface streams may sink underground, and there may be streambeds that carry water only during seasons of high flow or during extreme floods. In addition, the floors of the valleys may be dissected into a sequence of sinkholes.

Consider a normal stream valley that gradually deepens its channel until it cuts into underlying beds of limestone (or dolomite). As the valley cuts deeper and deeper into the carbonate rocks, the stream that flows through it loses water into the limestone through joints and fractures, which begin to enlarge into cave systems. At first, the cave passages will be very small and capable of carrying only a small amount of water. The stream flow on the surface will be reduced but not eliminated. As time passes, the cave passages become larger and capable of carrying more water. There will come a time when they are large enough to take the entire flow of the surface stream during periods of low flow, and during these low-flow periods—typically during summer and fall—the surface stream will run dry. With the passage of more time the cave system continues to enlarge, and more and more of the surface drainage is directed into it. The caves may become large enough to carry even the largest flood flows, and the surface channels will remain dry all year. The surface at this stage is called a dry valley, and it is no longer deepened because no more streams flow through it. Stream banks collapse, channels become overgrown with vegetation, and shallow sinkholes begin to form in the valley floor. Upstream from these "swallow holes" where surface streams are lost to the subsurface, the tributary valleys continue to deepen their channels. These evolve into so-called blind valleys, which end where a stream sinks beneath a cliff. At the top of the cliff is the abandoned floor of the dry valley. In short, fluviokarst is a landscape of active stream valleys, dry valleys, blind valleys, and deranged drainage systems. It is a common type of karst landscape where the soluble carbonate rocks are not as thick as the local relief, so that some parts of the landscape are underlain by carbonate rocks and others by such non-soluble rocks as sandstones or shales.

**Doline karst.** Such karsts are usually rolling plains that have few surface streams and often no surface valleys. Instead, the landscape is pocked with sinkholes, often tens or hundreds of sinkholes per square kilometre. These sinkholes range from barely discernible shallow swales one to two metres wide to depressions hundreds of metres in depth and one or more kilometres in width. As the sinkholes enlarge, they coalesce to form compound sinks or valley sinks. Some sinkholes form by the dissolution of bedrock at the intersections of joints or fractures. Others result from the collapse of cave roofs, and still others form entirely within the soil. The latter, known as cover collapse sinks and cover subsidence sinks, occur where soils are thick and can be washed into the subsurface by the process of soil piping. Soil loss begins at the bedrock interface. An arched void forms, which migrates upward through the soil until finally the roof collapses abruptly to form the sinkhole. These types of sinkhole constitute a serious land-use problem in karst areas and have been responsible for much property damage when they develop beneath streets, parking lots, houses, and commercial buildings.

Predominance of sinkholes

**Cone and tower karst.** This variety of karst landscape occurs mainly in tropical areas. Thick limestones are divided into blocks by a grid of joints and fractures. Solution produces deep rugged gorges along the joints and fractures, dividing the mass of limestone into isolated blocks. Because the water dissolving the gorges drains to the subsurface, the gorges are not integrated into a valley system. In some localities, the intervening blocks are rounded into closely spaced conical hills (cone karst). In others, the deepening gorges reach a base level and begin to widen. Sufficient widening may create a lower-level plain from which the remnants of the limestone blocks stand out as isolated, near-vertical towers (tower karst; Figure 37). The cones and towers themselves are sculptured by solution, so that the rock surface is covered by jagged pinnacles and often punctuated by pits and crevices.

Terrain of deep gorges and isolated limestone blocks

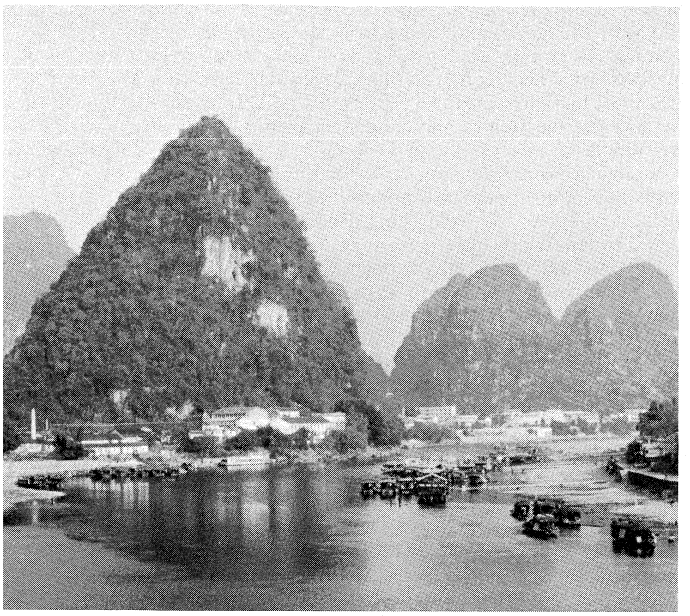


Figure 37: Tower karst in Kuei-lin, China.
G. Prance—Visuals Unlimited

**Pavement karst.** This form of karst develops where bare carbonate rocks are exposed to weathering. The initiation of pavement karst is often due to glaciation, which scrapes off soil and weathered rock material to expose the bare bedrock. Accordingly, pavement karsts occur mainly in high latitudes and alpine regions where glacial activity has been prominent. Solutional weathering of the exposed limestone or dolomite is due both to direct rainfall onto the rock surface and to meltwater derived from winter snowpack.

Karren

Pavement karst is decorated with an array of small landforms created by differential solution. These are collectively known as karren. Karren include solutionally widened joints (kluftkarren, or cleftkarren), small runnels (rinnenkarren, or runnelkarren), small residual pinnacles (spitzkarren, or pinnacle karren), and many other forms.

## GEOGRAPHIC DISTRIBUTION OF KARST TERRAIN

Approximately 15 percent of the Earth's land surface is karst. The distribution of karst is essentially the same as the distribution of carbonate rocks, which means that karst terrain occurs mostly in the great sedimentary basins of the world. It does not occur in the continental shields underlain by granites and related rocks or in volcanic belts, except in certain islands where massive limestones have been deposited on or around old volcanic cones.

Interior Low Plateaus of the United States

The most extensive karst area of the United States occurs in the limestones of Mississippian age (about 325,000,000 to 345,000,000 years old) of the Interior Low Plateaus. Mostly doline karst with some fluviokarst is found from southern Indiana south along both the east and west flanks of the broad fold of the Cincinnati Arch through eastern and central Kentucky and into Tennessee. Karst also occurs in the limestones of Ordovician age (about 430,000,000 to 500,000,000 years old) that lie exposed on the inner Bluegrass structural dome in Kentucky and on the Nashville Dome in Tennessee. In south central Kentucky is the Mammoth Cave area with the world's longest known cave and many other large cave systems. The Mississippian karst of Kentucky, Tennessee, and Indiana is quite remarkable because the many long cave systems and large areas of doline karst occur in a layer of limestone slightly more than 150 metres thick. Extensive karst also is developed on the limestones that ring the Ozark Dome in Missouri and northern Arkansas. Large caves and areas of fluviokarst and doline karst are found there.

Other notable karst regions of eastern North America include the Appalachians (specifically the Valley and Ridge and Great Valley provinces as well as the Cumberland and Allegheny plateaus) and Florida, where a raised platform of carbonate rocks has large areas of doline karst and extensive internal drainage through a major limestone aquifer. Bermuda and the Bahama Islands also are underlain by young limestones that are highly "karstified." Much of this karst was drowned by rises in sea level at the end of the Pleistocene glaciation. Caves containing stalactites and stalagmites are found at depths of tens of metres below present sea level.

The southwestern United States has very diverse karst regions. For example, West Texas, western Oklahoma, and eastern New Mexico have extensive areas of doline karst in gypsum with many small caves. The Edwards Plateau in south central Texas has a subdued surface karst and numerous small caves. The Capitan reef limestone in southeastern New Mexico contains Carlsbad Caverns and other deep and large volume caves.

The Canadian Rockies

The Rocky Mountains have many small areas of alpine karst in Colorado, Wyoming, Utah, and Montana. These are mostly pavement karst with relatively small caves. The Rockies of Canada contain some of that country's longest and deepest caves as well as extensive areas of alpine karst.

Some of the most spectacular examples of tropical karst occur in Central America and the Caribbean. The islands of the Greater Antilles (Cuba, Jamaica, Hispaniola, and Puerto Rico) are underlain by massive limestones up to 1,000 metres thick. Regions of cone and tower karst have developed in these limestones. The karst of Mexico varies from the streamless, low-relief plain of the Yucatán Peninsula to the high plateaus of the interior with their large dolines and deep vertical caves. Cone and tower karst occurs in the southern part of Mexico and in Belize and Guatemala. Many caves have been reported in Venezuela and Colombia. Little is known of karst in the other countries of South America. Much of the continent is occupied by the Guiana Shield and the Andes Mountains.

The Grande Causse of France

Because of its diversity of geologic and climatic settings, Europe has many different types of karst terrain. In the south, the Pyrenees exhibit spectacular alpine karst on both the Spanish and French sides. The high-altitude pavement karst contains many deep shafts. The Pierre Saint-Martin System, for example, is 1,342 metres deep and drains a large area of the mountain range. Southern France, notably the Grande Causse, has some of the most spectacular karst in Europe, with deep gorges, numerous caves, and much sculptured limestone. In the Alps are massive folded and faulted limestones and dolomites that underlie alpine karst terrain from France to Yugoslavia. In France, the Vercors Plateau is pavement karst featuring many deep caves, including the Berger Shaft—one of the deepest in Europe. The Hölloch Cave, the world's third longest at 133 kilometres, is found in the Swiss Alps. Individual limestone massifs capped with karst plateaus and abounding with deep caves occur in the Austrian Alps.

Karst is more of a local affair in northern Europe with relatively small caves in West Germany and Scandinavia. Some caves have been formed since the Pleistocene glaciation in Norway, as has some high-latitude pavement karst.

England and Ireland have extensive karst areas. The karst of Wales contains the longest caves in England, while the Yorkshire karst has complex vertical caves. Many parts of Ireland are underlain by limestone, and an area called the Burren in County Clare not only has the most caves but also some of the most extensive low-altitude pavement karsts.

The Adriatic Karst

Most of the countries of eastern Europe have karst areas, but special attention must be paid to Yugoslavia. The entire southern margin of the country from Slovenia to Montenegro and from the Adriatic Coast 50 kilometres into the interior is karst. In addition to areas of fluviokarst, doline karst, and pavement karst, the karst of Yugoslavia is unique for its large number of poljes. These are closed depressions with flat and alluviated bottoms that may be as much as 60 kilometres in diameter. Many of these depressions are elongate parallel to the geologic structure and to the Adriatic coastline. Although isolated poljes have been identified elsewhere, their large numbers in the Adriatic Karst are attributable to a system of active faults as well as to intense solution activity in nearly 9,000 metres of carbonate rock.

Much of the Mediterranean region—Greece, Turkey, Lebanon, Israel, and parts of the Arabian peninsula—are arid karst. The region had much more rainfall during the ice ages of the Quaternary, and so karst landscapes developed. Today, a combination of arid climatic conditions and overgrazing has reduced many parts of the region to bare rock, an arid-climate form of pavement karst. This is effectively a fossil karst that preserves a record of earlier climatic conditions. The karst regions extend eastward through parts of Iraq to the Zagros Mountains of Iran.

Relatively little karst has been described in Africa. Deep shafts and many caves occur in the Atlas Mountains in the northern part of the continent. Some caves have been described in Zaire, and caves are known in South Africa where sinkhole collapse in the Transvaal Dolomite due to dewatering by gold mining has been a serious environmental problem.

Asia is a vast region where many types of karst occur. In the Soviet Union, important karst areas are found in the Caucasus and Ural mountains. There is an important area of gypsum karst in the Ukraine, where very large network caves of gypsum occur. Karst covers about 2,000,000 square kilometres in China, but most renowned is the tower karst of Kweichow, Kwangsi, Yunnan, and Hunan provinces. The Chinese tower karst is developed on folded and faulted rocks unlike most other regions of cone and tower karst, which occur on thick horizontal strata. Isolated vertical-walled towers more than 200 metres high are found along river floodplains in those provinces.

Tower karst of China

Karst regions occur in the South Pacific. In Australia there are caves and some scattered sinkholes along the Nullarbor Plain. Additional karst areas occur in the eastern part of the continent. Many of the Pacific islands are coral reefs that have become karst to varying extents. Extensive cone and tower karst is found in New Guinea, Java, Borneo, and the Malay Peninsula. (W.B.Wh.)

## Landforms produced by coastal processes

The coastal environment of the world is made up of a wide variety of landforms manifested in a spectrum of sizes and shapes ranging from gently sloping beaches to high cliffs. The type of landform that is present along any coast is the result of a combination of processes, sediments, and the geology of the coast itself.

Coastal landforms are best considered in two broad categories; erosional and depositional. In fact, the overall nature of any coast may be described in terms of one or the other of these categories. It should be noted, however, that each of the two major landform types may occur on any given reach of coast.

### FACTORS AND FORCES IN THE FORMATION OF COASTAL FEATURES

The landforms that develop and persist along the coast are the result of a combination of processes acting upon the sediments and rocks present in the coastal zone. The most prominent of these processes involves waves and the currents that they generate, along with tides. Other factors that significantly affect coastal morphology are climate and gravity.

**Waves.** The most obvious of all coastal processes is the continual motion of the waves moving toward the beach. Waves vary considerably in size over time at any given location and also vary markedly from place to place. Waves interact with the ocean bottom as they travel into shallow water; as a result, they cause sediment to become temporarily suspended and available for movement by coastal currents. The larger the wave, the deeper the water in which this process takes place and the larger the particle that can be moved. Even small waves that are only a few tens of centimetres high can pick up sand as they reach the shore. Larger waves can move cobbles and rock material as large as boulders.

Suspension of sediment for transport by currents

Generally, small waves cause sediment—usually sand—to be transported toward the coast and to become deposited on the beach. Larger waves, typically during storms, are responsible for the removal of sediment from the coast and its conveyance out into relatively deep water.

Waves erode the bedrock along the coast largely by abrasion. The suspended sediment particles in waves, especially pebbles and larger rock debris, have much the same effect on a surface as sandpaper does. Waves have considerable force and so may break up bedrock simply by impact.

**Longshore currents.** Waves usually approach the coast at some acute angle rather than exactly parallel to it. Because of this, the waves are bent (or refracted) as they enter shallow water, which in turn generates a current along the shore and parallel to it. Such a current is called a longshore current, and it extends from the shoreline out through the zone of breaking waves. The speed of the current is related to the size of the waves and to their angle of approach. Under rather quiescent conditions, longshore currents move only about 10–30 centimetres per second; however, under stormy conditions they may exceed one metre per second. The combination of waves and longshore current acts to transport large quantities of sediment along the shallow zone adjacent to the shoreline.

Because longshore currents are caused by the approaching and refracting waves, they may move in either direction along the coast, depending on the direction of wave approach. This direction of approach is a result of the wind direction, which is therefore the ultimate factor in determining the direction of longshore currents and the transport of sediment along the shoreline.

Although a longshore current can entrain sediment if it moves fast enough, waves typically cause sediment to be picked up from the bottom, and the longshore current transports it along the coast. In some locations, there is quite a large volume of net sediment transport along the coast because of a dominance of one wind direction—and therefore wave direction—over another. This volume may be on the order of 100,000 cubic metres per year. Other locations may experience more of a balance in wave approach, which causes the longshore current and sediment transport in one direction to be nearly balanced by the same process in the other direction.

**Rip currents.** Another type of coastal current caused by wave activity is the rip current (incorrectly called rip tide in popular usage). As waves move toward the beach, there is some net shoreward transport of water. This leads to a slight but important upward slope of the water level (setup), so that the absolute water level at the shoreline is a few centimetres higher than it is beyond the surf zone. This situation is an unstable one, and water moves seaward through the surf zone in an effort to relieve the instability of the sloping water. The seaward movement is typically confined to narrow pathways. In most cases, rip currents are regularly spaced and flow at speeds of up to several tens of centimetres per second. They can carry sediment and often are recognized by the plume of suspended sediment moving out through the surf zone. In some localities rip currents persist for months at the same site, whereas in others they are quite ephemeral.

**Tides.** The rise and fall of sea level caused by astronomical conditions is regular and predictable. There is a great range in the magnitude of this daily or semi-daily change in water level. Along some coasts the tidal range is less than 0.5 metre, whereas in the Bay of Fundy in southeastern Canada the maximum tidal range is just over 16 metres. A simple but useful classification of coasts is based solely on tidal range without regard to any other variable. Three categories have been established: micro-tidal (less than two metres), meso-tidal (two to four metres), and macro-tidal (more than four metres). Micro-tidal coasts constitute the largest percentage of the world's coasts, but the other two categories also are widespread.

Tidal ranges

The role of tides in molding coastal landforms is twofold: (1) tidal currents transport large quantities of sediment and may erode bedrock, and (2) the rise and fall of the tide distributes wave energy across a shore zone by changing the depth of water and the position of the shoreline.

Tidal currents transport sediment in the same way that longshore currents do. The speeds necessary to transport the sediment (typically sand) are only generated under certain conditions—usually in inlets, at the mouths of estuaries, or any other place where there is a constriction in the coast through which tidal exchange must take place.

Tidal currents on the open coast, such as along a beach or rocky coast, are not swift enough to transport sediment. The speed of tidal currents in constricted areas, however, may exceed two metres per second, especially in inlets located on a barrier island complex. The speed of these tidal currents is dictated by the volume of water that must pass through the inlet during a given flood or ebb-tide cycle. This may be either six or 12 hours in duration, depending on whether the local situation is semidiurnal (12-hour cycle) or diurnal (24-hour cycle). The volume of water involved, called the tidal prism, is the product of the tidal range and the area of the coastal bay being served by the inlet. This means that though there may be a direct relationship between tidal range and tidal-current speed, it is also possible to have very swift tidal currents on a coast where the tidal range is low if the bay being served by the inlet is quite large. This is a very common situation along the coast of the Gulf of Mexico where the range is typically less than one metre but where there are many large coastal bays.

The rise and fall of the tide along the open coast has an indirect effect on sediment transport, even though currents capable of moving sediment are not present. As the tide comes in and then retreats along a beach or on a rocky coast, it causes the shoreline to move accordingly. This movement of the shoreline changes the zone where waves and longshore currents can do their work. Tidal range in combination with the topography of the coast is quite important in this situation. The greater the tidal range, the more effect this phenomenon has on the coast. The slope of a beach or other coastal landform also is important, however, because a steep cliff provides only a nominal change in the area over which waves and currents can do their work even in a macro-tidal environment. On the other hand, a broad, gently sloping beach or tidal flat may experience a change in the shoreline of as much as one kilometre during a tidal cycle in a macro-tidal setting. Examples of this situation occur in the Bay of Fundy and along the West German coast of the North Sea.

**Other factors and processes.** Climate is an extremely important factor in the development of coastal landforms. The elements of climate include rainfall, temperature, and wind.

The role of climate

Rainfall is important because it provides runoff in the form of streams and also is a factor in producing and transporting sediment to the coast. This fact gives rise to a marked contrast between the volume and type of sediment carried to the coast in a tropical environment and those in a desert environment.

Temperature is important for two quite different reasons. It is a factor in the physical weathering of sediments and rocks along the coast and in the adjacent drainage basins. This is particularly significant in cold regions where the freezing of water within cracks in rocks causes the rocks to fragment and thereby yield sediment. Some temperate and arctic regions have shore ice up to several months each year. Under these conditions there is no wave impact, and the coast becomes essentially static until the ice thaws or breaks up during severe storms. Such conditions prevail for three to four months along much of the coast of the Great Lakes in North America.

Wind is important primarily because of its relationship to waves. Coasts that experience prolonged and intense winds also experience high wave-energy conditions. Seasonal patterns in both wind direction and intensity can be translated directly into wave conditions. Wind also can be a key factor in directly forming coastal landforms, particularly coastal dunes. The persistence of onshore winds throughout much of the world's coast gives rise to sand dunes in all places where enough sediment is available and where there is a place for it to accumulate.

The effect of gravity

Gravity, too, plays a major role in coastal processes. Not only is it indirectly involved in processes associated with wind and waves but it also is directly involved through downslope movement of sediment and rock as well. This role is particularly evident along shoreline cliffs where waves attack the base of the cliffs and undercut the slope, resulting in the eventual collapse of rocks into the sea or their accumulation as debris at the base of the cliffs.

## LANDFORMS OF EROSIONAL COASTS

There are two major types of coastal morphology: one is dominated by erosion and the other by deposition. They exhibit distinctly different landforms, though each type may contain some features of the other. In general, erosional coasts are those with little or no sediment, whereas depositional coasts are characterized by abundant sediment accumulation over the long term. Both temporal and geographic variations may occur in each of these coastal types.

Characteristics of erosional coasts

Erosional coasts typically exhibit high relief and rugged topography. They tend to occur on the leading edge of lithospheric plates, the west coasts of both North and South America being excellent examples (see above *Tectonic landforms*). Glacial activity also may give rise to erosional coasts, as in northern New England and in the Scandinavian countries. Typically, these coasts are dominated by exposed bedrock with steep slopes and high elevations adjacent to the shore. Although these coasts are erosional, the rate of shoreline retreat is slow due to the resistance of bedrock to erosion. The type of rock and its lithification are important factors in the rate of erosion.

**Sea cliffs.** The most widespread landforms of erosional coasts are sea cliffs (Figure 38). These very steep to vertical bedrock cliffs range from only a few metres high to hundreds of metres above sea level. Their vertical nature is the result of wave-induced erosion near sea level and the subsequent collapse of rocks at higher elevation. Cliffs that extend to the shoreline commonly have a notch cut into them where waves have battered the bedrock surface.

At many coastal locations there is a thin, narrow veneer of sediment forming a beach along the base of sea cliffs. This sediment may consist of sand, but it is more commonly composed of coarse material—cobbles or boulders. Beaches of this kind usually accumulate during relatively low wave-energy conditions and are removed during the stormy season when waves are larger. The coasts of California and Oregon contain many places where this situation prevails. The presence of even a narrow beach along a rocky coast provides the cliffs protection against direct wave attack and slows the rate of erosion.

Peter Arnold, Inc.

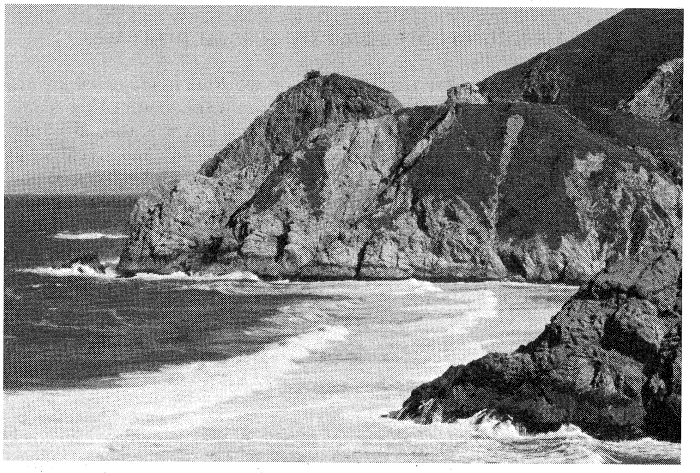


Figure 38: Erosional coast exemplified by rugged cliffs in San Mateo County, Calif.

**Wave-cut platforms.** At the base of most cliffs along a rocky coast one finds a flat surface at about the mid-tide elevation. This is a benchlike feature called a wave-cut platform, or wave-cut bench. Such surfaces may measure from a few metres to hundreds of metres wide and extend to the base of the adjacent cliff. They are formed by wave action on the bedrock along the coast. The formation process can take a long time, depending on the type of rock present. The existence of extensive wave-cut platforms thus implies that sea level did not fluctuate during the periods of formation. Multiple platforms of this type along a given reach of coast indicate various positions of sea level.

**Sea stacks.** Erosion along rocky coasts occurs at various rates and is dependent both on the rock type and on the wave energy at a particular site. As a result of

the above-mentioned conditions, wave-cut platforms may be incomplete, with erosional remnants on the horizontal wave-cut surface. These remnants are called sea stacks, and they provide a spectacular type of coastal landform. Some are many metres high and form isolated pinnacles on the otherwise smooth wave-cut surface. Because erosion is a continual process, these features are not permanent and will eventually be eroded, leaving no trace of their existence.

Isolated rock pinnacles

**Sea arches.** Another spectacular type of erosional landform is the sea arch, which forms as the result of different rates of erosion typically due to the varied resistence of bedrock. These archways may have an arcuate or rectangular shape, with the opening extending below water level. The height of an arch can be up to tens of metres above sea level (Figure 39).

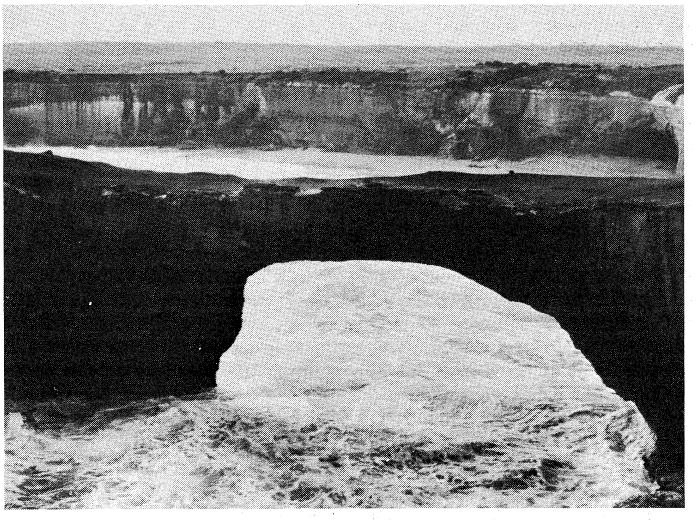


Figure 39: London Bridge, a sea arch along the southern coast of Victoria, Australia. It rises about 30 metres above sea level.

It is common for sea arches to form when a rocky coast undergoes erosion and a wave-cut platform develops. Continued erosion can result in the collapse of an arch, leaving an isolated sea stack on the platform. Still further erosion removes the stack, and eventually only the wave-cut platform remains adjacent to the eroding coastal cliff.

## LANDFORMS OF DEPOSITIONAL COASTS

Characteristics of depositional coasts

Coasts adjacent to the trailing edge of lithospheric plates tend to have widespread coastal plains and low relief. The Atlantic and Gulf coasts of the United States are representative. Such coasts may have numerous estuaries and lagoons with barrier islands or may develop river deltas. They are characterized by an accumulation of a wide range of sediment types and by many varied coastal environments. The sediment is dominated by mud and sand; however, some gravel may be present, especially in the form of shell material.

Depositional coasts may experience erosion at certain times and places due to such factors as storms, depletion of sediment supply, and rising sea level. The latter is a continuing problem as the mean annual temperature of the Earth rises and the ice caps melt. Nevertheless, the overall, long-range tendency along these coasts is that of sediment deposition.

All of the processes discussed at the beginning of this section are in evidence along depositional coasts. Waves, wave-generated currents, and tides significantly influence the development of depositional landforms. In general, waves exert energy that is distributed along the coast essentially parallel to it. This is accomplished by the waves themselves as they strike the shore and also by the longshore currents that move along it. In contrast, tides tend to exert their influence perpendicular to the coast as they flood and ebb. The result is that the landforms that develop along some coasts are due primarily to wave processes while along other coasts they may be due mainly to tidal processes. Some coasts are the result of near equal balance between tide and wave processes. As a consequence, investigators speak of wave-dominated coasts, tide-dominated coasts, and mixed coasts.

Sand beaches and long barrier islands

A wave-dominated coast is one that is characterized by well-developed sand beaches typically formed on long barrier islands with a few widely spaced tidal inlets. The barrier islands tend to be narrow and rather low in elevation. Longshore transport is extensive, and the inlets are often small and unstable. Jetties are commonly placed along the inlet mouths to stabilize them and keep them open for navigation. The Texas and North Carolina coasts of the United States are excellent examples of this coastal type.

Tide-dominated coasts are not as widespread as those dominated by waves. They tend to develop where tidal range is high or where wave energy is low. The result is a coastal morphology that is dominated by funnel-shaped embayments and long sediment bodies oriented essentially perpendicular to the overall coastal trend. Tidal flats, salt marshes, and tidal creeks are extensive. The West German coast of the North Sea is a good example of such a coast.

Mixed coasts are those where both tidal and wave processes exert considerable influence. These coasts characteristically have short stubby barrier islands and numerous tidal inlets. The barriers commonly are wide at one end and narrow at the other. Inlets are fairly stable and have large sediment bodies on both their landward and seaward sides. The Georgia and South Carolina coasts of the United States typify a mixed coast.

Primary large-scale coastal types

**General coastal morphology.** Depositional coasts can be described in terms of three primary large-scale types: (1) deltas, (2) barrier island/estuarine systems, and (3) strand-plain coasts. The latter two have numerous features in common.

*Deltas.* An accumulation of sediment at the mouth of a river extending beyond the trend of the adjacent coast is called a delta. Deltas vary greatly in both size and shape, but they all require that more sediment is deposited at the river mouth than can be carried away by coastal processes. A delta also requires a shallow site for accumulation—namely, a gently sloping continental shelf.

The size of a delta is typically related to the size of the river, specifically to its discharge. The shape of a delta, on the other hand, is a result of the interaction of the river with tidal and wave processes along the coast. A classification utilizing each of these three factors as end members provides a good way of considering the variation in delta morphology (Figure 40). River-dominated deltas are those where both wave and tidal current energy on the coast is low and the discharge of water and sediment are little affected by them. The result is an irregularly shaped delta with numerous digitate distributaries.The Mississippi Delta is a good example of a river-dominated delta.

Waves may remove much of the fine deltaic sediment and smooth the outer margin of the delta landform as

From W.E. Galloway, *Deltas, Models for Exploration* (1975); Houston Geological Society

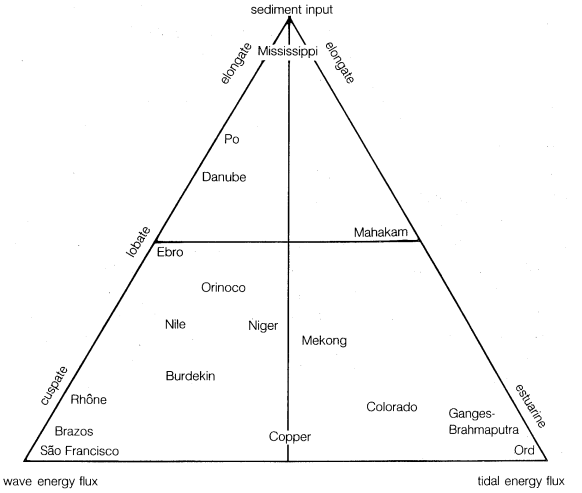



Figure 40: Classification of river deltas based on the three dominant processes that control delta morphology (see text).

well. This results in a smooth, cuspate delta that has few distributaries. The São Francisco Delta in Brazil is such a delta. Some wave-dominated deltas are strongly affected by longshore currents, and the river mouth is diverted markedly along the coast. The Sénégal Delta on the west coast of Africa is an example.

Tide-dominated deltas tend to be developed in wide, funnel-shaped configurations with long sand bodies that fan out from the coast. These sand bodies are oriented with the strong tidal currents of the delta. Tidal flats and salt marshes also are common. The Ord Delta in northern Australia and the Ganges-Brahmaputra Delta in Bangladesh are representative of such a deltaic type.

*Barrier island/estuarine systems.* Many depositional coasts display a complex of environments and landforms that typically occur together. Irregular coasts have numerous embayments, many of which are fed by streams. Such

stuaries

embayments are called estuaries, and they receive much sediment due to runoff from an adjacent coastal plain. Seaward of the estuaries are elongate barrier islands that generally parallel the shore. Consisting mostly of sand, they are formed primarily by waves and longshore currents. These barrier islands are typically separated from the mainland and may have lagoons, which are long, narrow, coastal bodies of water situated between the barrier and the mainland.

Most barrier islands contain a well-developed beach, coastal dunes, and various environments on their landward side, including tidal flats, marshes, or washover fans. Such coastal barriers are typically interrupted by tidal inlets, which provide circulation between the various coastal bays and the open marine environment. These inlets also are important pathways for organisms that migrate between coastal and open marine areas as well as for pleasure and commercial boat traffic.

*Strand-plain coasts.* Some wave-dominated coasts do not contain estuaries and have no barrier island system. These coasts, however, do have beaches and dunes, and may even have coastal marshes. The term strand plain has been applied to coasts of this sort. Examples include parts of western Louisiana and eastern Texas. In most respects, they are similar in morphology to barrier islands but lack inlets.

**Beaches and coastal dunes.** There are several specific landforms representative of coastal environments that are common to each of the three major categories described above. Especially prominent among these are beaches and dunes. They are the primary landforms on barrier islands, strand-plain coasts, and many deltas, particularly the wave-dominated variety.

*Beaches.* A consideration of the beach must also include the seaward adjacent nearshore environment because the two are intimately related. The nearshore environment extends from the outer limit of the longshore bars that are usually present to the low-tide line. In areas where longshore bars are absent, it can be regarded as coincident with the surf zone. The beach extends from the low-tide line to the distinct change in slope and/or material landward of the unvegetated and active zone of sediment accumulation. It may consist of sand, gravel, or even mud, though sand is the most common beach material.

The beach profile typically can be divided into two distinct parts: (1) the seaward and relatively steep sloping

oreshore

foreshore, which is essentially the intertidal beach, and

nd

(2) the landward, nearly horizontal backshore (Figure 41).

ackshore

Beach profiles take on two different appearances, depending on conditions at any given time. During quiescent wave conditions, the beach is said to be accretional, and both the foreshore and backshore are present. During storm conditions, however, the beach experiences erosion, and the result is typically a profile that shows only the seaward sloping foreshore. Because the beach tends to repair itself during nonstorm periods, a cyclic pattern of profile shapes is common.

The nearshore zone is where waves steepen and break, and then re-form in their passage to the beach, where they break for the last time and surge up the foreshore. Much sediment is transported in this zone, both along the shore and perpendicular to it. During storms the waves tend to be steep, and erosion of the beach occurs with sediment transported offshore. The intervening calmer conditions permit sediment to be transported landward and rebuild the beach. Because wave conditions may change daily, the nature of the profile and the sediment on the foreshore portion of the beach may also change daily. This is the zone of continual change on the beach.

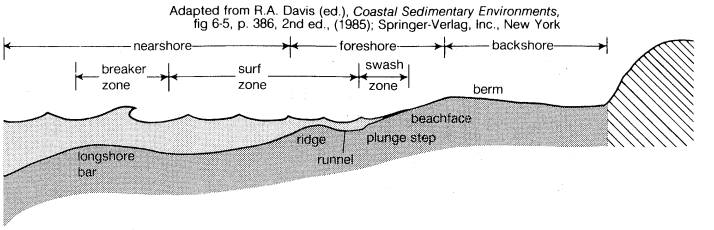



Figure 41: Schematic diagram of a typical beach and nearshore profile.

The backshore of the beach is not subjected to wave activity except during storm conditions. It is actually in the supra-tidal zone—*i.e.,* the zone above high tide where inundation by water is caused not by regular astronomical tides but rather by storm-generated tides. During nonstorm conditions the back-beach is relatively inactive except for wind action, which may move sediment. In most cases, there is an onshore component to the wind, and sediment is carried from the back-beach landward, typically forming dunes. Any obstruction on the back-beach, such as vegetation, pieces of driftwood, fences, or even trash discarded by people, results in wind-blown sand accumulation.

There are variations in beach forms along the shore as well as in those perpendicular to the shore. Most common is the rhythmic topography that is seen along the foreshore. A close look at the shoreline along most beaches will show that it is not straight or gently curved but rather that it displays a regularly undulating surface much like a low-amplitude sine curve. This is seen both on the plan view of the shoreline and the topography of the foreshore. The spacing is regular along a given reach of coast, but it may vary from place to place or from time to time at a given place. At some locations, concentrations of gravel or shells may develop, forming beach cusps (more or less triangular deposits that point seaward) during some wave conditions.

Beach cusps

Although there is a common trend to the beach profile, some variation exists both because of energy conditions and because of the material making up the beach. Generally speaking, a beach that is accumulating sediment and experiencing low-energy conditions tends to have a steep foreshore, whereas the same beach would have a relatively gentle foreshore during storm conditions when erosion is prevalant. The grain size of beach sediment also is an important factor in the slope of the foreshore. In general, the coarser the constituent grains, the steeper the foreshore. Examples include the gravel beaches of New England, as contrasted to the gently sloping sand beaches of the Texas coast.

*Coastal dunes.* Immediately landward of the beach are commonly found large, linear accumulations of sand known as dunes. (For coverage of dunes in arid and semi-arid regions, see above *Sand dunes.*) They form as the wind carries sediment from the beach in a landward direction and deposits it wherever an obstruction hinders further transport. Sediment supply is the key limiting factor in dune development and is the primary reason why some coastal dunes, such as those on the west Florida peninsula, are quite small, whereas others in such areas as the Texas coast and the Florida panhandle have large dunes.

Small wind-shadow dunes, or coppice mounds, actually may form on the backshore of the beach. If sediment continues to be supplied and beach erosion does not destroy them, these small sand accumulations will become foredunes, the seaward-most line of coastal dunes. It is

Foredunes

in this fashion that a coast progrades, or grows seaward. Many barrier-island or strand-plain coasts exhibit numerous, essentially parallel dune ridges testifying to this type of growth.

The sediment in dunes tends to be fine to medium sand

that is quite well sorted. Shell debris or other material is uncommon unless it is the same size or mass as the dune sand. There are various types of vegetation that grow on the dune surface and stabilize it. These grasses and vines often can be seen on the backshore portion of beaches as well. Dunes lacking vegetation are usually active and exhibit various signs of sand mobility. Most widespread are the nearly ubiquitous ripples that cover the dune surface. Large lobes of sand or even an entire dune may also move as wind blows across the dune. This activity results in cross stratification of the dune in large sweeping patterns of wedge-shaped packages of sand. (R.A.D.)

# OTHER TYPES OF LANDFORMS

## Impact craters

There are features on the continents that have been created by the impact of cosmic bodies (meteorites, asteroids, or comets) on the Earth's surface. In fact, the collision of a Mars-sized body with the Earth early in the history of the solar system may have caused the formation of the Moon and the continents themselves, or rather the ocean/continent dichotomy of the Earth's outer layers.

Impact craters differ from one another principally in terms of size, progressing from small, simple cup-shaped depressions originally with raised rims, through hollows having a central hill (or peak), to those with more complicated central peaks and basins, and finally to multi-ringed basins up to 1,000 kilometres in diameter. The surfaces of the Moon, Mercury, and certain regions of Mars and Venus, as well as most of the satellites of the solar system, are dominated by impact craters. They are much rarer on the Earth, which has about 100 ranging in size from a few tens of metres to 160 kilometres. These have been identified with varying degrees of certainty. The reason for the relative scarcity of impact craters on the Earth is that the processes that formed its surface—largely those of weathering, erosion, and mountain building—have undoubtedly removed the majority of the older structures, such as are observed on the surfaces of the other planets mentioned above.

There is abundant evidence that impact cratering was not only an important surface process in planetary history but that large impact events produced effects that were crustal in scale. The formation of multi-ring basins on the early Moon, for example, is just as important a process in defining the tectonic framework of that body as plate-tectonic phenomena are on the Earth (see above *Tectonic landforms*). Evidence from several planets indicates that the effects of very large-scale impacts go beyond the simple formation of an impact structure and serve to localize increased internal geologic activity over an extended period of geologic time. Although no longer occurring with the same frequency and magnitude as during the early solar system, large-scale impact events have continued to affect the local geology of the inner planets.

Effects of large-scale impact events

The Moon travels through essentially the same orbital space as the Earth. Since the Moon has not been subjected to the surficial geomorphic processes and tectonic activity that prevail on Earth, the lunar surface serves as an invaluable repository of information about the rate and effects of large body impacts that are likely to have occurred on the Earth. Telescopic observations, orbiting remote-sensing spacecraft, and the Apollo manned explorations of the Moon conducted by the United States during the early 1970s have provided the necessary data to establish surface impact rates. These are quite comparable to rates derived independently from astronomical studies for fluxes of asteroidal and cometary objects in near-Earth space. (For details pertaining to this and related questions, see SOLAR SYSTEM: *Asteroids, Comets,* and *Meteoroids, meteors, and meteorites.*)

Some important conclusions for the Earth follow. Asteroidal objects as large as 20 kilometres in diameter probably have struck the planet during the last few billion years, and bodies measuring 10 kilometres across apparently may collide with it every 50,000,000 to 100,000,000 years. Cometary nuclei of similar size may have nearly comparable rates of collision. A 10-kilometre stony object with a density of about three grams per cubic centimetre (0.12 pound per cubic inch) striking the terrestrial surface at a velocity of 25 kilometres per second (15.5 miles per second) would have kinetic energy in excess of 100,000,000 megatons—far greater than that contained in the world's total nuclear arsenal. On land, the impact of such an object would produce a transient crater 60 to 70 kilometres in diameter; the subsequent collapse of the basin walls would result in a final crater having a diameter of possibly 100 to 125 kilometres. Moreover, an impact of this kind would probably exert a planetwide influence on biological evolution because it could trigger mass extinctions of entire species, as perhaps in the case of the dinosaurs 65,000,000 years ago.

### GENERAL CHARACTERISTICS OF IMPACT CRATERS ON THE TERRESTRIAL SURFACE

In general, the larger the structure, the longer some of it is preserved but also the less certain the interpretation as an impact structure. Young craters have circular topographic highs around a cup-shaped depression. Older structures may completely lack discernible crater rims, yet they may still retain some degree of circularity, have distinctive internal deformation features, or contain remnants of original rock fragments (breccia) that formed during the impact event. Radial, circular, and annular drainage patterns are common.

Older, more eroded examples of craters are called astroblemes, or more commonly cryptoexplosion structures. The latter designation is a modification of the term cryptovolcanic structure, which is no longer used because investigators have found no evidence for volcanism but much evidence for impact at many such sites. Thus, in a broader sense, an impact crater can be defined as any structure now at or near the surface that contains evidence of a shock-producing impact in which the disturbance by deformation and fragmentation of rock or soil is generally circular and which fades out rapidly in intensity of deformation upon reaching diameter-to-depth ratios of about three to one.

Cryptoexplosion structures

Complex impact structures can be characterized by the presence or absence of peaks and rings, with subdivision into central peak craters, central peak basins, peak ring basins, and multi-ring basins. In general, the sequence from simple structures to multi-ring basins corresponds to increasing diameter or impact energy. The transition diameter of the morphological change from simple to complex form varies between planets probably because of differences in gravity. Most known and suspected terrestrial impact structures more than about three kilometres in diameter contain central uplifts, whereas the smaller ones do not. On the Moon, where gravity is one-sixth that on Earth, craters with central peaks are usually more than 15 to 20 kilometres in diameter. The central peaks make up roughly 10 percent of the width of the total circular deformed zone. Although there is evidence that planetary gravity exerts some control over the value of the transition diameter, it is apparent that parameters such as target rock characteristics also have an effect. This is evident on the Earth where the transition from simple to complex form occurs at four to five kilometres in crystalline rocks but at two to three kilometres in sedimentary rocks. Most of what is known about these various types is based on studies of the lunar surface. (C.R.S.)

### FORMATION OF IMPACT CRATERS

**Mechanics of the cratering process.** The cratering mechanics of simple craters are fairly well understood (Figure 42). On impact, the bulk of the kinetic energy of the projectile is transferred to the planetary surface, where a shock wave is formed. A shock wave is a transient pulse

Generation of shock waves

that moves at velocities higher than those of elastic (seismic or sonic) waves in the same material or medium and that produces an almost instantaneous rise in pressure behind its advancing front. Ahead of this pressure rise, the medium is not yet affected in any way. Behind it, the medium is immediately compressed, leading to a decrease in volume (increase in density) until decompression, or rarefaction, waves allow progressive relaxation. Pressure magnitudes in shock waves exceed the dynamic elastic limits of the transmitting materials. For rocks, this limit falls between 20 and 100 kilobars (1 kilobar equals 1,000 bars, or approximately 14,700 pounds per square inch). By comparison, nuclear explosions and impacts generate pressure waves whose initial amplitudes can reach megabars (1,000,000 bars) in the materials undergoing shock.

An important property of shock waves involves a possible change of phase upon encountering a free surface or another medium of different density and wave propagating characteristics. An initial compression wave will thus be converted to a rarefaction wave, which, if reflected over the previous wave path, can "unload" the state of compression or place the medium under tension (*i.e.,* cause it to stretch or tear). As a shock wave diverges spherically from a point source, its expansion into an ever larger volume also reduces its magnitude, so that the pressure at the moving front continually decreases with increasing radial distance from its source.

Simple craters

**Stages of formation.** The sequence of events involved in the formation of a simple impact crater (shown schematically in Figure 42) is as follows. Immediately after an incoming object (*e.g.,* a meteorite) strikes the ground, shock waves are imparted both to the rocks and to the object itself. The shock front in the rocks outruns the penetrating meteorite and forms an expanding cavity. The rocks behind the front are strongly compressed and generally are set in radial motion outward from the region of penetration. Depending on the properties of the rocks present, varying fractions of this target material are vaporized, melted, crushed, fragmented, or fractured. Ultimately, different segments of the volume of rock actually excavated are mixed together and dispersed. The initially intense shock attenuates as it diverges outward. The net effect is that the degree of shock damage attributable to compression diminishes with increasing radial distance from the line of penetration. Only a small proportion of the total volume excavated undergoes strong to intense shock pressures, and many of these rocks experience little or no permanent damage.

Excavation process

The actual excavation process requires rarefaction waves. These waves form as the advancing compression waves become isolated, or "detached," from the excavation flow and move onward in the target as a fast-moving shock wave. The shock pressure declines to zero behind the front, but the particle velocity does not. This soon sets up rarefactions that effectively place the rock medium under tension, resulting in a general disruption of the component units. Because these now fragmented pieces were already in motion, many retain enough momentum to carry them forward and out of the developing crater at low angles as ejecta. Fragments that move out of the crater at high angles fall back as crater-filling breccia deposits. Some of the rock material is carried downward without ever leaving the crater to form concentrations of fragmental debris along its base. Beneath these rocks lies a rupture zone consisting of broken and displaced rock that underlies the breccia fill. Compression and rarefaction waves become too weak to efficiently excavate material below the upper boundary of the rupture zone.

Within the rupture zone below the true crater floor and within its walls, the rocks may be folded, faulted, and intricately deformed. Especially in the rim, which marks the outer limit of surface excavation, layered units may be folded over completely to form structures that are similar in complexity to some of the great nappes (folded rock masses moved by thrust faulting) of the Alps. The overturning results when adjacent units below the uplifting free surface are peeled back as they move in close succession. Massive rocks do not fold back but tend to undergo uplift along fractures. All crater-related deformation, however, decreases abruptly away from the junction between the breccia and rupture zone, so beyond this interface the transition from shocked or disturbed rock to ordinary, unmodified country rock is rapid.

During the initial compression stage, materials behind

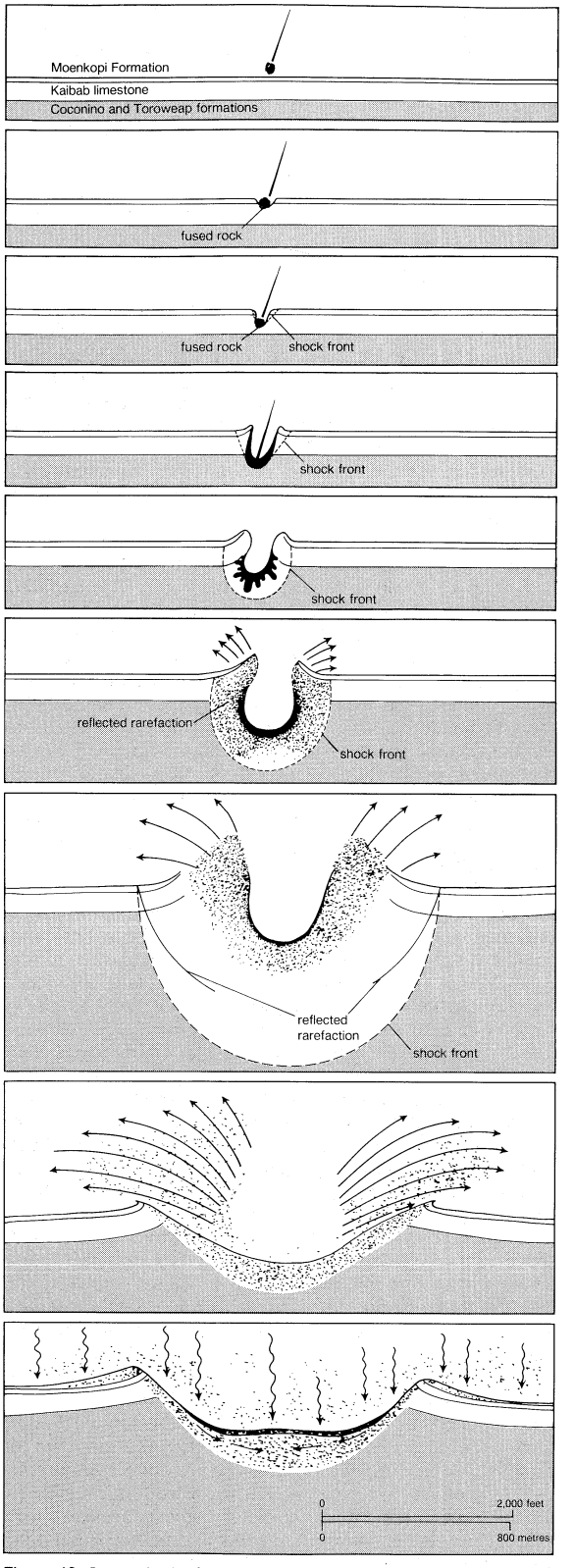



Figure 42: *Stages in the formation of a simple crater.* Shown here is the sequence of events thought to have occurred when an incoming iron meteorite struck the layered rock structures of the Colorado Plateau near Winslow, Ariz., and produced Meteor Crater (see text).

After E.M. Shoemaker in B.M. Middlehurst and G.P. Kuiper (eds.), *The Solar System,* vol. 4 (1963); University of Chicago Press

the shock front experience pressures so high (many hundreds of kilobars) that their dynamic compressive strengths are greatly exceeded. Under these conditions rocks behave much like fluids.

As the spherical shock front advances outward in larger impacts, the compressed shell of material behind it mixes with the meteorite and ejects steadily increasing amounts at the onset of the excavation stage. The complex geometric patterns of particle movements that follow progressive relaxation by rarefaction waves result in the lateral deflection of flowing materials from their earlier radial paths. Below the line of impact, particle movements still proceed mainly downward, but they deviate increasingly sideward until a tangential flow from the crater at low angles dominates the region of the expanding rim. Ejection and crater growth represent a continuous, orderly process involving a steady flow of materials along flow lines. Cratering ceases when stresses become too weak to break up materials.

**Variations in structure.** In its original state, a simple crater is characterized by a well-defined rim, a bowl-shaped cross section, extensive fallback deposits that thicken toward the centre, and small to moderate amounts (1 to 2 percent) of impact melt concentrated along the central base (Figure 42). Its initial diameter to depth ratio is about five to one. When the crater diameter is greater than four to five kilometres, the rupture zone beneath the central base tends to uplift along shear lines.

Complex craters

Large complex craters have broad, dome-like uplifts, which are sometimes preserved in topographic expression as central peaks (Figure 43). Rocks within this uplifted portion may be strongly shocked and intricately contorted. The mechanism for uplift is related to the size of the impact, properties of the materials, and planetary gravity. It results from the plasticlike rebound of the materials upward under the centre of the impact (Figure 44, uplift stage). Normally, in layered materials the uplift contains stratigraphic units from depths below the general level of the crater base, which indicates the real upward movement of rocks.

Some large craters with central peaks may have been caused by comet impact rather than meteorite impact. When the nucleus of a low-density comet strikes the ground, the dominant energy release remains near the surface. As a result, lateral excavation is more effective than in the case of penetrating iron or stony meteorites.

Generally, complex craters have their breccia-melt deposits in an annular depression between the rim and central peak.

The passage of the transient pressure waves through the rocks and their constituent minerals induces distinctive

By courtesy of the National Aeronautics and Space Administration

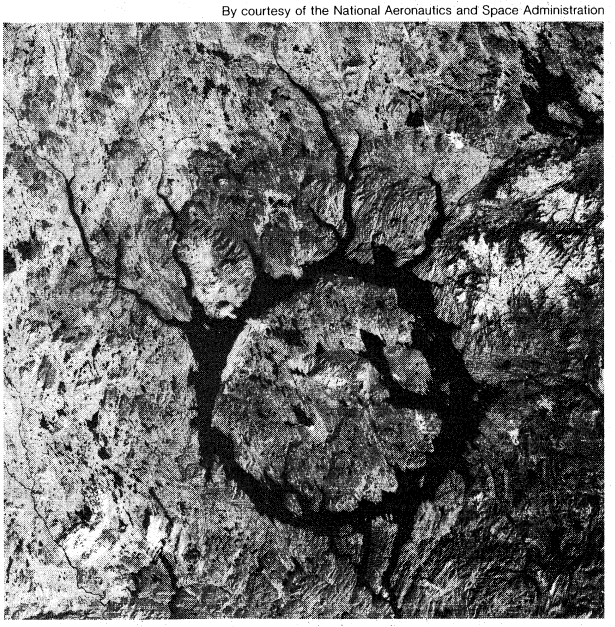


Figure 43: Satellite photograph of the Manicouagan Crater, a complex impact structure, in Quebec, Canada.

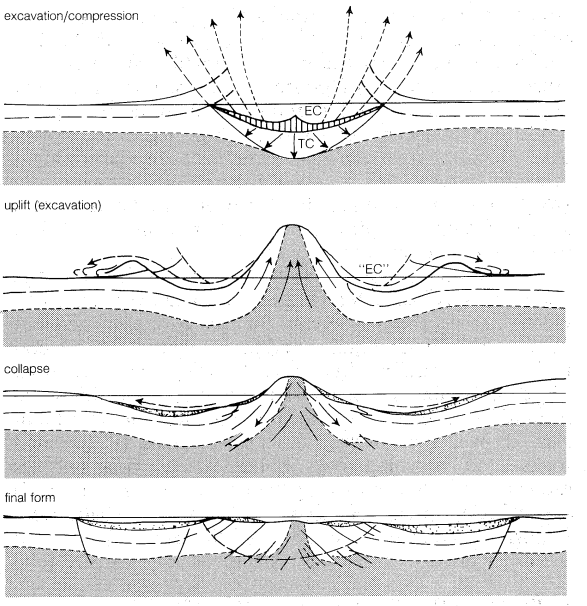



Figure 44: *Stages in the formation of a complex impact crater.* During excavation/compression, the depression created by the penetrating meteorite reaches maximum size, indicated by the extent of the excavated cavity (EC) and transient cavity (TC). The area shown with stripes represents the flow field of fragmented and melted rock that materializes below the level of farthest penetration by the meteorite. The small dashed lines mark the limit of fracturing and brecciation. During uplift, the transient cavity floor undergoes maximum rebound. The resulting dome-like structure subsequently collapses, accompanied by overthrusting and deposition of rock debris in the surrounding depression. In some cases the uplift is preserved as a central peak, while in others it collapses below the original surface level, as shown here.

Adapted from R.A.F. Grieve, P.B. Robertson, and M.R. Dence, "Constraints on the Formation of Ring Impact Structures, Based on Terrestrial Data," *Multi-ring Basins, Proceedings of Lunar and Planetary Sciences* (1981)

effects of very high pressures and temperatures and rapid strain rates that alter their composition and properties. Such effects are described as shock metamorphism. They include the formation of high-pressure polymorphs (different crystalline forms of the same composition) that normally are not stable in the Earth's upper crust; fragmented, fractured, and granulated rock masses; unusual intracrystalline markings; solid-state and selective melting and brief vesiculation (bubbling) caused by pressures ranging up to 500–750 kilobars (comparable to pressures at depths of 1,400–1,800 kilometres within the Earth). The effective temperatures may rise above 2,000° C (3,600° F), and the strain rates are millions of times faster than those operating during mountain building. For these reasons, shock pressures can be determined from rock samples, and the original size of ancient, eroded craters can be estimated. At 20–30 kilobars, cleavage planes in mica start to bend and kink as though folded like layered strata. At higher pressures, more individual mica crystals begin to become distorted, and the intensity of deformation increases in all flakes thus affected. When the shock pressures in the passing waves exceed values of 50–75 kilobars (the dynamic elastic limit of many granitic rocks), changes unknown even in equivalent rock types buried deep within the crust begin to appear.

Shock metamorphism

Another such change is the development of shatter cones. These are peculiar external fracture surfaces in which close-spaced "grooves" seemingly radiate outward from the apex of a cone somewhat like the hair of a horse's tail. Many such cones usually cluster together. Most have apexes that point in a common direction, which, when separate cone-bearing layers are rotated to their original position by photographic analysis, converge toward a single origin that appropriately coincides with the explosion centre of the initial impact. In 1947 Robert S. Dietz of the United States proposed that these features were indicative of impact—*i.e.,* that they originated from shock waves. Subsequent research has proved this hypothesis to be correct.

Shatter cones

In quartz grains, sets of lamellae-like planes form within crystals over a pressure range from about 100 to 400 kilobars. These planar features presumably result from the crushing or slipping of crystal portions along preferred planes of weakness as the grains are compressed rapidly into much smaller volumes. In the higher pressure ranges over which the quartz planar features are produced, high-pressure forms of silica called coesite and stishovite, develop. Other shock-induced polymorphic transformations, as, for example, diamond from other forms of carbon, have been discovered in impact structures.

At still higher pressures (450–600 kilobars), the effect is one of rapidly rising temperature from the conversion of mechanical to thermal energy that leads first to incipient melting of individual grains and then eventually to general melting and variable mixing of the whole rock material.

Transformation of rock material

The net result is the formation of a crater and the transformation of some of the local rocks into impact lithologies: impact melt rocks, shocked breccias (fragmented rocks), and deformed rocks. Evidence from rock samples returned from the lunar highlands, 90 percent of which are of impact origin, suggests that impact lithologies were important geologic constituents of the early crusts of all the inner planets of the solar system. (N.M.S./C.R.S.)

POPULATION OF TERRESTRIAL IMPACT CRATERS

It has been estimated that over the past billion years the land surface of the Earth has been subjected to about 130,000 impacts that have produced craters one kilometre or more in diameter. For the reasons mentioned above, there simply remains no record of most of these impacts. Approximately 100 major craters have been identified worldwide. Numerous small craters have been detected as well. Some are concentrated in clusters and are so small that they cannot be considered landforms. Various small craters also have been associated with specific meteorites.

Terrestrial impact craters range in age from Precambrian to Recent, but the record is heavily biased toward the Recent (see above). Examples of simple bowl-shaped craters, complex central peak structures, central peak basins, and peak-ring basins have been found on the Earth. Examples of multi-ring basins are rare, however. This is a result of the relatively young surface of the Earth and the sharp decrease in the number of impacting bodies in recent geologic time, as compared with the period when large multi-ring basins were formed on the Moon. Based on the lunar record scaled to terrestrial conditions, the Earth, which is more than 4,000,000,000 years old, may have had 25 to 50 basins larger than 1,000 kilometres in diameter. It is possible that the Sudbury crater in Ontario, Canada, and the Vredefort crater in South Africa, both of which measure about 140 kilometres across, originally had a multi-ring form. They both, however, are Precambrian in age and are highly degraded or modified.

Scarcity of multi-ring basins on the Earth

Possibly the best candidates for a well-preserved, multi-ring structure on Earth are the Popigai crater in the Soviet Union, which measures roughly 100 kilometres in diameter and apparently has three rings in the crystalline rocks of its floor; or the Acraman crater in South Australia, which seems to have an inner depressed area about 30 kilometres in diameter, an intermediate depression or ring approximately 90 kilometres in diameter, and possibly an outer ring about 160 kilometres in diameter. Also, investigators have located what appears to be a large crater with central peaks in Wilkes Land, Antarctica. This depression is more than 848 metres deep and 243 kilometres across. (C.R.S.)

## Biogenic landforms

Topographic features that can be attributed to the activity of organisms are diverse in both kind and scale. Organisms contribute to the genesis of most topography involving rock weathering; they play an auxiliary role, as demonstrated by bacterial and lichen activity, the effects of root wedging, and solutional erosion made possible by humic acid produced by rapid organic decay. The latter is responsible for much tropical karst (see above *Caves and karst landscape*).

On an entirely different level are features that constitute what may be termed micro topography. Some of these are produced by individual creatures or groups of such creatures, as, for example, the cylindrical mud towers that stand 40–50 centimetres high atop crayfish burrows in the southern part of the United States; badger and bear den burrows; elephant waterholes on the veld (grasslands of Africa); and quarries and open-pit mines dug by humans. Other topographic features are attributable to colonial organisms. In various parts of the world such as the semiarid plains of the Western Sahara, colonies of termites build large conical mounds that reach several metres high. The interaction of corals, algae, and bryozoa is largely responsible for the framework of features known as organic reefs, which abound in tropical marine settings. Some of these reefs have given rise to entire insular land areas many kilometres in diameter. The largest example is the Great Barrier Reef of Australia, which covers an area of about 207,000 square kilometres. Though nearly submerged today, it was an island during the Pleistocene glaciation.

Micro topographic features

With the possible exception of the Great Barrier Reef, all major biogenic landforms produced in recent times are attributable to the activities of humankind. The construction of modern superhighways involves some of the most extensive terrain changes on Earth, having in some cases resulted in the removal of mountains or at least large portions thereof. Many human effects are not necessarily tied to particular construction projects. On a subtler level, the removal of fluids from the ground, principally water and petroleum, have dropped water tables and reduced pore pressure so greatly that extensive areas have experienced subsidence, collapse, and shrinkage. Terrain changes due to groundwater removal are extremely severe in such regions as the southwestern United States or the area near Mexico City. To the foregoing human effects on topography must be added the bomb craters left by war that are ever so slowly being obliterated from Europe and Asia, and the erosional gullying of terrain where uncontrolled deforestation has been allowed. Finally, there are the engineering modifications of waterways and coasts practiced nowhere more intensively than in the United States and Europe. River flow patterns have been drastically altered, usually by channel straightening, and the construction of large dams has converted entire valleys, gorges, and canyons into lakes. In fact, dams are among the largest biogenic landforms produced. (H.F.G.)

Impact of human activities

BIBLIOGRAPHY

*Landform evolution and theoretical views:* The seminal works of WILLIAM MORRIS DAVIS, especially his 1899 essay "The Geographical Cycle" and his 1905 essay "The Geographical Cycle in an Arid Climate," both reprinted in *Geographical Essays* (1909, reprinted 1954), paved the way for WALTHER PENCK, *Morphological Analysis of Land Forms: A Contribution to Physical Geology* (1953, reprinted 1972; originally published in German, 1924); and for LESTER C. KING, "Canons of Landscape Evolution," *Bulletin of the Geological Society of America,* 64:721–752 (1953). Davis' thinking continued to dominate geomorphic theory in such textbooks as A.K. LOBECK, *Geomorphology: An Introduction to the Study of Landscapes* (1939); and WILLIAM D. THORNBURY, *Principles of Geomorphology,* 2nd ed. (1968). The history of geomorphic theory up to the time of Davis is told in RICHARD J. CHORLEY, ANTONY J. DUNN, and ROBERT P. BECKINSALE, *The History of the Study of Landforms,* 2 vol. (1964–73). CUCHLAINE A.M. KING (ed.), *Landforms and Geomorphology: Concepts and History* (1976), contains a historical collection of key papers in the development of geomorphic ideas from 1802 to 1972. Introductory essays on the many aspects of the field may be found in ALISTAIR PITTY (ed.), *Geomorphology: Themes and Trends* (1985).

The first comprehensive attempt to synthesize geomorphic thinking on a systems level that incorporates modern tectonic, climatic, and process aspects is H.F. GARNER, *The Origin of Landscapes: A Synthesis of Geomorphology* (1974). An analysis of these and other geomorphic theories to date appeared in WILLIAM N. MELHORN and RONALD C. FLEMAL (eds.), *Theories of Landform Development* (1975, reissued 1980). More restricted aspects of geomorphology are dealt with in LUNA B. LEOPOLD, M. GORDON WOLMAN, and JOHN P. MILLER, *Fluvial Processes in Geomorphology* (1964); G.H. DURY (ed.), *Essays in Geomorphology* (1966); J. TRICART and A. CAILLEUX, *Introduction to Climatic Geomorphology* (1972; originally published in French, 1965); ROBERT V. RUHE, *Geomorphology: Geomorphic Processes and Surficial Geology* (1975); C.R. TWIDALE, *Analysis*

*of Landforms* (1975, reprinted 1976); R.J. RICE, *Fundamentals of Geomorphology* (1977); DALE F. RITTER, *Process Geomorphology,* 2nd ed. (1986); and RITA GARDNER and HELEN SCOGING (eds.), *Mega-Geomorphology* (1983).

(H.F.G.)

*Tectonic landforms:* (*Tectonic processes*): An excellent introduction to geology, somewhat dated but very well illustrated, is ARTHUR HOLMES, *Principles of Physical Geology,* 2nd rev. ed. (1965), with a particularly good treatment of rift valleys. WILLIAM J. PERRY, DIETRICH H. RODER, and DAVID R. LAGESON (comps.), *North American Thrust-Faulted Terranes* (1984), is a collection of technical papers describing segments of folded and thrusted mountain belts in North America and the mechanics of such deformation. Introductory articles on volcanism at hot spots and island arcs include K. BURKE and T. WILSON, "Hot Spots on the Earth's Surface," *Scientific American,* 235(2):46–57 (August 1976); and BRUCE D. MARSH, "Island-Arc Volcanism," *American Scientist,* 67(2):161–172 (March-April 1979). A discussion of the forces that support mountain ranges and how some ranges are constructed can be found in PETER MOLNAR, "The Structure of Mountain Ranges," *Scientific American,* 255(1):70–79 (July 1986).

(*Mountain systems*): Detailed introductions to the European Alps are E.R. OXBURGH, *The Geology of the Eastern Alps* (1968), concentrating on Austria; and R. TRÜMPY, *An Outline of the Geology of Switzerland,* vol. 1 of *Geology of Switzerland: A Guide-Book,* ed. by the SCHWEIZERISCHE GEOLOGISCHE KOMMISSION (1980). Introductory articles on other mountain belts of the world are found in *Scientific American,* including the following: DON L. ANDERSON, "The San Andreas Fault," 225(5):52–68 (November 1971); PETER MOLNAR and PAUL TAPPONNIER, "The Collision Between India and Eurasia," 236(4):30–41 (April 1977), describing the evolution of mountain belts in eastern Asia; and DAVID L. JONES *et al.,* "The Growth of Western North America," 247(5):70–84 (November 1982), describing the accretion of exotic terrains to that area. Two useful articles in *Geological Society of America Bulletin* are TANYA ATWATER, "Implications of Plate Tectonics for the Cenozoic Tectonic Evolution of Western North America," 81(12):3513–35 (December 1970), a classic article that changed the way geologists view that region; and JOHN F. DEWAY *et al.,* "Plate Tectonics and the Evolution of the Alpine System," 84(10)3137–80 (October 1973), a summary of motions of lithospheric plates and mountain building in the Mediterranean area. See also PETER MOLNAR, "The Geologic History and Structure of the Himalaya," *American Scientist,* 74(2):144–154 (March-April 1986), describing how the structure of the Himalayas allows the chain to achieve its great height.

(*Rift valleys*): B.H. BAKER, P.A. MOHR, and L.A.J. WILLIAMS, *Geology of the Eastern Rift System of Africa* (1972), provides a summary of the eastern branch of the East African Rift System. More specific treatment is found in B.H. BAKER and J. WOHLENBERG, "Structure and Evolution of the Kenya Rift Valley," *Nature,* 229(5286):538–542 (Feb. 19, 1971).

(P.H.M.)

*Structural landforms:* Landforms resulting from erosion and deposition are analyzed in RICHARD J. CHORLEY, STANLEY A. SCHUMM, and DAVID E. SUGDEN, *Geomorphology* (1984); and DALE F. RITTER, *Process Geomorphology,* 2nd ed. (1986).

(*Stream valleys and canyons*): Descriptions can be found in geomorphology textbooks, including M.J. SELBY, *Earth's Changing Surface: An Introduction to Geomorphology* (1985); and ARTHUR L. BLOOM, *Geomorphology: A Systematic Analysis of Late Cenozoic Landforms* (1978). The relevant fluvial phenomena are treated in DAVID KNIGHTON, *Fluvial Forms and Processes* (1984); and the relevant hillslope phenomena in M.J. SELBY, *Hillslope Materials and Processes* (1982). More specialized topics include the following: VICTOR R. BAKER, *The Channels of Mars* (1982); THEODORE OBERLANDER, *The Zagros Streams: A New Interpretation of Transverse Drainage* (1965); JULIUS BÜDEL, *Climatic Geomorphology* (1982; originally published in German, 1977); and GEORGE F. ADAMS (ed.), *Planation Surfaces: Peneplains, Pediplains, and Etchplains* (1975).

(*Playas, pans, and saline flats*): A useful collection of scientific papers is JAMES T. NEAL (ed.), *Playas and Dried Lakes: Occurrence and Development* (1975). The playas of Death Valley, California, are described by CHARLES B. HUNT, *Death Valley: Geology, Ecology, and Archaeology* (1975).

(V.R.B.)

(*Sand dunes*): Descriptions and geographic distribution of forms produced by wind deposition are provided in a still-useful classic text, R.A. BAGNOLD, *The Physics of Blown Sand and Desert Dunes* (1941, reprinted 1984). More information is given in K.W. GLENNIE, *Desert Sedimentary Environments* (1970); RONALD U. COOKE and ANDREW WARREN, *Geomorphology in Deserts* (1973); EDWIN D. MCKEE, CAROL S. BREED, and STEVEN G. FRYBERGER, "Desert Sand Seas," ch. 2 in *Skylab Explores the Earth* (1977), pp. 5–48, NASA document sp 380; CAROL S. BREED, MAURICE J. GROLIER, and JOHN F. MCCAULEY, "Morphology and Distribution of Common Sand Dunes on Mars: Comparison with the Earth," *Journal of Geophysical Research,* 84(B14):8183–8204 (Dec. 30, 1979); EDWIN D. MCKEE (ed.), *A Study of Global Sand Seas* (1979); CAROL S. BREED *et al.,* "Eolian (Wind-Formed) Landforms," in TERAH L. SMILEY *et al.* (eds.), *Landscapes of Arizona: The Geological Story* (1984), pp. 359–413; and RONALD GREELEY and JAMES D. IVERSEN, *Wind as a Geological Process: On Earth, Mars, Venus, and Titan* (1985).

(W.J.Br./C.S.Br.)

(*Glacial and periglacial landforms*): The classic text on glacial geology is RICHARD FOSTER FLINT, *Glacial and Quaternary Geology* (1971), encyclopaedic coverage including an extensive bibliography. Recent hypotheses and observations on glacial erosion and deposition are included in DAVID DREWRY, *Glacial Geologic Processes* (1986), even though the coverage of glacial landforms is not complete. DAVID E. SUGDEN and BRIAN S. JOHN, *Glaciers and Landscape: A Geomorphological Approach* (1976, reprinted 1984), is an excellent detailed introduction to glacial landforms and the processes that shaped them. More theoretical emphasis can be found in CLIFFORD EMBLETON and CUCHLAINE A.M. KING, *Glacial Geomorphology,* 2nd ed. (1975), and *Periglacial Geomorphology,* 2nd ed. (1975). A.L. WASHBURN, *Geocryology: A Survey of Periglacial Processes and Environments* (1979), contains numerous explanatory photographs and diagrams. A collection of articles is found in CUCHLAINE A.M. KING (ed.), *Periglacial Processes* (1976).

(E.B.E./Gu.S.)

(*Caves and karst topography*): Two introductory texts are ALFRED BÖGLI, *Karst Hydrology and Physical Speleology* (1980; originally published in German, 1978), emphasizing caves; and J.N. JENNINGS, *Karst Geomorphology,* 2nd rev. ed. (1985), covering both caves and surface landforms—a recommended introduction to karst. Discussion of cave geology, hydrology, mineralogy, and biology can be found in T.D. FORD and C.H.D. CULLINGFORD (eds.), *The Science of Speleology* (1976), with a good chapter on volcanic caves. CAROL A. HILL and PAOLO FORTI, *Cave Minerals of the World* (1986), is a systematic description of minerals and speleothems. PAUL COURBON and CLAUDE CHABERT, *Atlas des grandes cavités mondiales* (1986), provides a definitive world list of long and deep caves, with maps and descriptive texts. Country-by-country review chapters on karst geology and surface landforms are contained in M. HERAK and V.T. STRINGFIELD (eds.), *Karst: Important Karst Regions of the Northern Hemisphere* (1972), concentrating mostly on Europe, the U.S.S.R., and the United States.

(W.B.Wh.)

(*Coastal processes*): An introduction to coastal geomorphology is provided by ERIC C.F. BIRD, *Coasts* (1984). Other useful texts include FRANCIS P. SHEPARD and HAROLD R. WANLESS, *Our Changing Coastlines* (1971); PAUL D. KOMAR, *Beach Processes and Sedimentation* (1976); J.L. DAVIES, *Geographical Variation in Coastal Development,* 2nd ed. (1980); and RICHARD A. DAVIS, JR., *Coastal Sedimentary Environments,* 2nd rev. and expanded ed. (1985). ERIC C.F. BIRD and MAURICE L. SCHWARTZ (eds.), *The World's Coastline* (1985), is a comprehensive round-the-world treatment of coastal features.

(R.A.D.)

(*Impact craters*): Two general books on the subject are RONALD GREELEY, *Planetary Landscapes* (1985); and BRUCE MURRAY, MICHAEL C. MALIN, and RONALD GREELEY, *Earthlike Planets: Surfaces of Mercury, Venus, Earth, Moon, Mars* (1981). Collected papers from scientific meetings include BEVAN N. FRENCH and NICHOLAS M. SHORT (eds.), *Shock Metamorphism of Natural Materials* (1968); and LEON T. SILVER and PETER H. SCHULTZ (eds.), *Geological Implications of Impacts of Large Asteroids and Comets on the Earth* (1982), in which see especially RICHARD A.F. GRIEVE, "The Record of Impact on Earth: Implications for a Major Cretaceous/Tertiary Impact Event," pp. 25–37.

(C.R.S.)

# Copernicus

A major contribution to Western thought was the publication in 1543 of *De revolutionibus orbium coelestium, libri VI* (Eng. trans., *On the Revolutions of the Celestial Spheres,* 1952; Latin reprint, 1965) by Copernicus, Polish astronomer, who is noted for the Copernican theory of the heavens. By attributing to the Earth a daily motion around its own axis and a yearly motion around the stationary Sun, Copernicus developed an idea that had far-reaching implications for the rise of modern science. Henceforth, the Earth could no longer be considered the centre of the cosmos; rather, as one celestial body among many, it became subject to mathematical description.

By courtesy of the Museum of Jagiellonian University, Kraków, Poland




Copernicus, 17th-century copy of a 16th-century self-portrait. In the Museum of Jagiellonian University, Kraków, Poland.

Education

Nicolaus Copernicus (in Polish Mikołaj Kopernik, in German-Prussian dialect Niklas Koppernick) was born on February 19, 1473, at Toruń, near the Vistula River in eastern Poland, where his father was a merchant of social standing. In 1491 Copernicus entered the University of Kraków, where he became interested in the study of astronomy; he probably returned home in 1494 (or 1496). His maternal uncle, Lucas Waczenrode, newly elected bishop of Ermeland, wanted him to enter the canonry of Frauenburg in order to secure lifelong financial independence. While waiting for a vacancy to occur, he was sent by his uncle in 1497 for further training to the University of Bologna, where he associated himself with the German students.

For three and a half years Copernicus studied the Greek language, mathematics, and the writings of Plato; he also became further acquainted with the astronomical thought of the day. In Bologna he also made his first recorded observation of the heavens, an occultation (overlapping, or eclipse) of the star Aldebaran by the Moon on March 9, 1497; the light of the former was shut off by the Moon. The same year he was elected (by proxy) a canon of Frauenburg. He travelled to Rome in 1500 for the great jubilee celebration and may have given informal lectures in mathematics there. In 1501 he briefly visited Frauenburg to claim his post on the cathedral staff, returning promptly to Italy under special leave of absence to continue his studies at the University of Padua. There, enrolled with other Polish students, he studied both law and medicine. Except for a short interruption in 1503, when he was granted the degree of doctor of canon law by the University of Ferrara, he spent almost four years in Padua.

On returning to Poland in 1503, he visited Kraków and later acted as adviser to his uncle until the latter's death in 1512. Copernicus settled permanently at Frauenburg, where he acted as representative of the cathedral chapter, his medical skill being used particularly in aid of the indigent.

As a result of his studies in Kraków and Padua, Copernicus may be said to have mastered all the knowledge of the day in mathematics, astronomy, medicine, and theology. Copernicus appears to have planned a systematic program of astronomical work. Although he did not make extensive observations, he did enough to enable him to recalculate the major components of the supposed orbits of the Sun, Moon, and planets around the Earth. He published 27 such observations made during the years 1497–1529, and a few others have been found entered in books in his private library. He also published for his uncle in 1509 a Latin translation of Greek verses of Theophylactus, a Byzantine poet of the 7th century AD, and from 1519 to 1528 prepared an exposition of the principles of currency reform for certain Polish provinces; the latter was not published in Warsaw, however, until 1816.

Copernicus' fame as an industrious student of astronomy rapidly increased, and in 1514 he was invited to give his opinion on calendar reform, which was then being considered by the Lateran Council, a general meeting of the church authorities. He refused to express any firm views, for he felt that the positions of the Sun and Moon were not known with sufficient accuracy to permit a proper reassessment.

Dissatisfaction with the Ptolemaic system

Yet, as his studies progressed, Copernicus became increasingly dissatisfied with the Ptolemaic system of astronomy. He was not alone in this dissatisfaction; indeed, he himself said that the many divergent views prevalent in his day gave him cause for profound thought. Ptolemy's system, which contained not only original work but also a synthesis of the views of previous Greek philosophers, was basically geocentric and circular in conception. By the 16th century this geocentric interpretation of the heavens had become firmly entrenched in astronomical thought, virtually as an article of faith. Although certain Greek philosophers had suggested, as far back as the 3rd century BC, that the Sun—and not the Earth—was the centre of the universe, their ideas had not been widely accepted. Difficulties had arisen when ancient astronomers sought to account for the accumulated observations of the Sun, Moon, and planets. Accordingly, Ptolemy in the 2nd century AD had devised an elaborate geocentric model of the heavens composed of large circles, called deferents, and small circles, called epicycles. Each planet rode on the circumference of an epicycle, the centre of which revolved on the deferent. Ptolemy used this system to account for observed irregularities of the planets, such as changes in brightness, and particularly for their puzzling retrogressive motions, when they seemed to stop and move backward and forward in a loop. Moreover, to account for observed variations in velocity, Ptolemy introduced the equant, which was an imaginary point in space where uniform, circular speed would indeed be observed. This system enabled astronomers to account for the phenomena and to make predictions. As observations in succeeding centuries became more accurate, however, it became increasingly difficult to compute the future positions of the heavenly bodies, and much of the flexibility and elegance of the Ptolemaic system was thereby lost.

Copernicus concluded that, in view of the many circles and their displacements from the center of the Earth that the Ptolemaic system required to account for the observed

motions of heavenly bodies, a simpler, alternative explanation might be possible. In consequence, he read the works of many original Greek authors and found that, indeed, heliocentric ideas had been suggested. The idea of a moving Earth seemed absurd at first, but, when Copernicus applied this assumption, the result was an aesthetically superior, although not much simpler, system, even though, as might be expected, he still believed that the planets moved with uniform circular motion. After many years of mathematical calculations, he became convinced that his new idea was true, yet he made no attempt to publish.

From about 1510 to 1514, Copernicus prepared a short manuscript to summarize his new idea, *De hypothesibus motuum coelestium a se constitutis commentariolus* ("A Commentary on the Theories of the Motions of Heavenly Objects from Their Arrangements"), which he privately circulated among friends in 1514. Its main points were that the apparent daily motion of the stars, the annual motion of the Sun, and the retrogressive behaviour of the planets result from the Earth's daily rotation on its axis and yearly revolution around the Sun, which is stationary at the centre of the planetary system. The Earth, therefore, is not the centre of the universe but only of the Moon's orbit. As the years passed, he developed his argument with diagrams and mathematical calculations. Lectures on the principles expounded in the *Commentariolus* were given in Rome in 1533 before Pope Clement VII, who approved, and a formal request to publish was made to Copernicus in 1536. But he continued to hesitate. It was only through the efforts of his friends—in particular, his pupil and disciple Georg Joachim Rhäticus, who studied with him for two years—that he finally published his work. In 1540 Rhäticus was permitted to take the completed manuscript to Nürnberg, Germany, for printing. Because of opposition from Martin Luther, Philipp Melanchthon, and other reformers, Rhäticus left Nürnberg and went to Leipzig, where he passed on the task of publication to Andreas Osiander. Apparently fearing criticism of a treatise that proposed an annual motion of the Earth around a stationary Sun, Osiander, on his own responsibility, inserted a preface emphasizing that the hypothesis of a stationary Sun was only a convenient means for simplifying planetary computations.

The Copernican system

A careful examination of the text makes it clear, however, that Copernicus had really come to believe in the heliocentric system—rather, heliostatic, since he placed the Sun at some distance from the centre—as a true picture of the universe. He wrote *On the Revolutions of the Celestial Spheres,* in six sections, as a mathematical reinterpretation of Ptolemy. He wished to provide an alternative computational scheme that would make possible more accurate predictions that would be used in calendar reform and eclipses, and that would, at the same time, explain the troublesome variations of brightness, retrogressions, and velocity with a simpler geometric system of points and circles.

In the first section, Copernicus gave some basic mathematical rules, countered the old arguments about the fixity of the Earth, and discussed the order of the planets from the Sun. He could no longer accept the old arrangement—Earth, Moon, Mercury, Venus, Sun, Mars, Jupiter, and Saturn—since this had been a consequence of a geocentric system. He found it necessary to adapt it to his heliocentric system and adopted the following order from the stationary Sun: Mercury, Venus, Earth with the Moon orbiting around it, Mars, Jupiter, and Saturn. In the second section, Copernicus applied the basic mathematical rules of the previous section to the apparent motions of the stars and planets, and attributed the motion of the Sun to the motion of the Earth. The third section contains a mathematical description of the Earth's motion, including the precession of the equinoxes, which is caused by the gyration of the Earth's axis. Sections four, five, and six deal with the motions of the Moon and of the five remaining planets.

In his heliocentric theory, Copernicus found himself able to describe the movements of the Moon and planets in a more elegant way than Ptolemy in his geocentric system. To fit the observations, Ptolemy had been forced to offset the centres of regular motion a slight way from the centre of the Earth, and this Copernicus believed to conflict with the basic rule of true circular motion. In *De revolutionibus* the centres all lay at the centre of the Sun, but, because Copernicus still adopted circular motions at an unvarying speed, his system proved to be virtually as complex as Ptolemy's. Nevertheless, Copernicus believed that his system was aesthetically more satisfying and that it was a true picture of the divinely ordained cosmos.

A copy of the great work is believed to have been brought to Copernicus at Frauenburg on the last day of his life, May 24, 1543.

Assessment

The Copernican system appealed to a large number of independent-minded astronomers and mathematicians. Its attraction was not only because of its elegance but also, in part, because of its break with traditional doctrines: in particular, it opposed Aristotle, who had argued cogently for the fixity of the Earth; furthermore, it provided an alternative to Ptolemy's geocentric universe. In Western Christendom both of these views had been elevated almost to the level of religious dogma; to many thoughtful observers, however, they stifled development and were overdue for rejection.

Scientifically, the Copernican theory demanded two important changes in outlook. The first change had to do with the apparent size of the universe. The stars always appeared in precisely the same fixed positions, but if the Earth were in orbit round the Sun, they should display a small periodic change. Copernicus explained that the starry sphere was too far distant for the change to be detected. His theory thus led to the belief in a much larger universe than previously conceived and, in England, where the theory was openly accepted with enthusiasm, to the idea of an infinite universe with the stars scattered throughout space. The second change concerned the reason why bodies fall to the ground. Aristotle had taught that they fell to their "natural place," which was the centre of the universe. But because, according to the heliocentric theory, the Earth no longer coincided with the centre of the universe, a new explanation was needed. This re-examination of the laws governing falling bodies led eventually to the Newtonian concept of universal gravitation.

The dethronement of the Earth from the centre of the universe caused profound shock. No longer could the Earth be considered the epitome of creation, for it was only a planet like the other planets. No longer was the Earth the centre of all change and decay with the changeless universe encompassing it. And the belief in a correspondence between man, the microcosm, as a mirror of the surrounding universe, the macrocosm, was no longer valid. The successful challenge to the entire system of ancient authority required a complete change in man's philosophical conception of the universe. This is what is rightly termed "the Copernican Revolution."

(C.A.R.)

**BIBLIOGRAPHY.** The general reader should consult three short biographies: JAN ADAMCZEWSKI and EDWARD J. PISZEK, *Nicolaus Copernicus and His Epoch* (1974), with emphasis on the Polish setting; ANGUS ARMITAGE, *Sun, Stand Thou Still* (1947; reissued as *The World of Copernicus,* 1956); and JÓZEF RUDNICKI, *Nicholas Copernicus* (1943); the last was written for the 400th celebration of the publication of *De revolutionibus orbium coelestium* and is particularly well illustrated. FRED HOYLE, *Nicolaus Copernicus: An Essay on His Life and Work* (1973), is mainly an examination of his scientific work; ANGUS ARMITAGE, *Copernicus: The Founder of Modern Astronomy* (1938, reissued 1962), is a more detailed biography. For an analysis of the impact of Copernicus' ideas, see THOMAS S. KUHN, *The Copernican Revolution* (1957). *De revolutionibus* has been translated by CHARLES G. WALLIS and published in "Great Books of the Western World," vol. 16 (1952). The specialized reader might refer to EDWARD ROSEN, *Three Copernican Treatises,* 3rd rev. ed. (1971). For a thorough discussion of Copernicus' work, see J.L.E. DREYER, *A History of Astronomy from Thales to Kepler,* 2nd ed. (1953); and ALEXANDRE KOYRÉ, *From the Closed World to the Infinite Universe* (1957, reissued 1968). ROBERT S. WESTMAN (ed.), *The Copernican Achievement* (1976), contains essays on technical aspects of his theories. For the reception of the Copernican theory in England, see FRANCIS R. JOHNSON, *Astronomical Thought in Renaissance England* (1937, reprinted 1968).

(C.A.R./Ed.)

# The Cosmos

If one looks up on a clear night, one sees that the sky is full of stars. During the summer months in the Northern Hemisphere, a faint band of light stretches from horizon to horizon, a swath of pale white cutting across a background of deepest black. For the early Egyptians, this was the heavenly Nile, flowing through the land of the dead ruled by Osiris. The ancient Greeks likened it to a river of milk. Astronomers now know that the band is actually composed of countless stars in a flattened disk seen edge on. The stars are so close to one another along the line of sight that the unaided eye has difficulty discerning the individual members. Through a large telescope, astronomers find myriads of like systems sprinkled throughout the depths of space. They call such vast collections of stars galaxies, after the Greek word for milk, and call the local galaxy to which the Sun belongs the Milky Way Galaxy or simply the Galaxy.

Every visible star is a sun in its own right. Ever since this realization first dawned in the collective mind of humanity, it has been speculated that many stars other than the Sun also have planetary systems encircling them. The related issue of the origin of the solar system, too, has always had special fascination for speculative thinkers, and the quest to understand it on a firm scientific basis has continued into the present day.

Some stars are intrinsically brighter than the Sun; others, fainter. Much less light is received from the stars than from the Sun because the stars are all much farther away. Indeed, they appear densely packed in the Milky Way only because there are so many of them. The actual separations of the stars are enormous, so large that it is conventional to measure their distances in units of how far light can travel in a given amount of time. The speed of light (in a vacuum) equals $3 \times 10^{10}$ cm/sec (centimetres per second); at such a speed, it is possible to circle the Earth seven times in a single second. Thus in terrestrial terms the Sun, which lies 500 light-seconds from the Earth, is very far away; however, even the next closest star, Proxima Centauri, at a distance of 4.3 light-years ($4.1 \times 10^{18}$ cm), is 270,000 times farther yet. The stars that lie on the opposite side of the Milky Way from the Sun have distances that are on the order of 100,000 light-years, which is the typical diameter of a large spiral galaxy.

If the kingdom of the stars seems vast, the realm of the galaxies is larger still. The nearest galaxies to the Milky Way system are the Large and Small Magellanic Clouds, two irregular satellites of the Galaxy visible to the naked eye in the Southern Hemisphere. The Magellanic Clouds are relatively small (containing roughly $10^9$ stars) compared to the Galaxy (with some $10^{11}$ stars), and they lie at a distance of about 200,000 light-years. The nearest large galaxy comparable to the Galaxy is the Andromeda galaxy (also called M31 because it was the 31st entry in a catalog of astronomical objects compiled by the French astronomer Charles Messier in 1781), and it lies at a distance of about 2,000,000 light-years. The Magellanic Clouds, the Andromeda galaxy, and the Milky Way system all are part of an aggregation of two dozen or so neighbouring galaxies known as the Local Group. The Galaxy and M31 are the largest members of this group.

The Galaxy and M31 are both spiral galaxies, and they are among the brighter and more massive of all spiral galaxies. The most luminous and brightest galaxies, however, are not spirals but rather supergiant ellipticals (also called cD galaxies by astronomers for historical reasons that are not particularly illuminating). Elliptical galaxies have roundish shapes rather than the flattened distributions that characterize spiral galaxies, and they tend to occur in rich clusters (those containing thousands of members) rather than in the loose groups favoured by spirals.

The brightest member galaxies of rich clusters have been detected at distances exceeding several thousand million light-years from the Earth. The branch of learning that deals with phenomena at the scale of many millions of light-years is called cosmology—a term derived from combining two Greek words, *kosmos,* meaning "order," "harmony," and "the world," and *logos,* signifying "word" or "discourse." Cosmology is, in effect, the study of the universe at large. A dramatic new feature, not present on small scales, emerges when the universe is viewed in the large—namely, the cosmological expansion. On cosmological scales, galaxies (or, at least, clusters of galaxies) appear to be racing away from one another with the apparent velocity of recession being linearly proportional to the distance of the object. This relation is known as the Hubble law (after its discoverer, the American astronomer Edwin Powell Hubble). Interpreted in the simplest fashion, the Hubble law implies that roughly $10^{10}$ years ago, all of the matter in the universe was closely packed together in an incredibly dense state and that everything then exploded in a "big bang," the signature of the explosion being written eventually in the galaxies of stars that formed out of the expanding debris of matter. Strong scientific support for this interpretation of a big bang origin of the universe comes from the detection by radio telescopes of a steady and uniform background of microwave radiation. The cosmic microwave background is believed to be a ghostly remnant of the fierce light of the primeval fireball reduced by cosmic expansion to a shadow of its former splendour but still pervading every corner of the known universe.

The simple (and most common) interpretation of the Hubble law as a recession of the galaxies over time through space, however, contains a misleading notion. In a sense, as will be made more precise later in the article, the expansion of the universe represents not so much a fundamental motion of galaxies within a framework of absolute time and absolute space, but an expansion of time and space themselves. On cosmological scales, the use of light-travel times to measure distances assumes a special significance because the lengths become so vast that even light, traveling at the fastest speed attainable by any physical entity, takes a significant fraction of the age of the universe, roughly $10^{10}$ years, to travel from an object to an observer. Thus, when astronomers measure objects at cosmological distances from the Local Group, they are seeing the objects as they existed during a time when the universe was much younger than it is today. Under these circumstances, Albert Einstein taught in his theory of general relativity that the gravitational field of everything in the universe so warps space and time as to require a very careful reevaluation of quantities whose seemingly elementary natures are normally taken for granted.

The observed expansion of the universe immediately raises the spectre that the universe is evolving, that it had a beginning and will have an end. The steady state alternative, postulated by a British school of cosmologists in 1948, is no longer considered viable by most astronomers. Yet, the notion that the Cosmos had a beginning, while common in many theologies, raises deep and puzzling questions for science, for it implies a creation event—a creation not only of all the mass-energy that now exists in the universe but also perhaps of space-time itself.

The issue of how the universe will end seems, at first sight, more amenable to conventional analysis. Because the universe is currently expanding, one may ask whether this expansion will continue into the indefinite future or whether after the passage of some finite time, the expansion will be reversed by the gravitational attraction of all of the matter for itself. The procedure for answering this question seems straightforward: either measure directly the rate of deceleration in the expansion of the galaxies to extrapolate whether they will eventually come to a halt,

or measure the total amount of matter in the universe to see if there is enough to supply the gravitation needed to make the universe bound. Unfortunately astronomers' assaults on both fronts have been stymied by two unforeseen circumstances. First, it is now generally conceded that all earlier attempts to measure the deceleration rate have been affected by evolutionary effects of unknown magnitude in the observed galaxies that invalidate the simple interpretations. Second, it is now generally recognized that there may be an unknown amount of "hidden mass" in the Cosmos, which cannot be seen by conventional astronomical techniques but which contributes very substantially to the gravitation of the universe.

The hope is that, somehow, quantum physics will ultimately supply theoretical answers (which can then be tested observationally and experimentally) to each of these difficulties. The ongoing effort in particle physics to find a unified basis for all the elementary forces of nature has yielded promising new ways to think about the most fundamental of all questions regarding astronomical origins; it has offered a tentative prediction concerning the deceleration rate of the universe; and it has offered a plethora of candidates for the hidden mass of the universe.

This article traces the historical development of modern conceptions of the Cosmos and summarizes the prevailing theories of its origin and evolution. Humanity has traveled a long road since self-centred societies imagined the creation of the Earth, the Sun, and the Moon as the main act, with the formation of the rest of the universe as almost an afterthought. Today, it is known that the Earth is only a small ball of rock in a Cosmos of unimaginable vastness and that the birth of the solar system was probably only one event among many that occurred against the backdrop of an already mature universe.

Yet, as humbling as the lesson has been, it has also unveiled a remarkable fact, one which endows the minutest particle in this universe with a rich and noble heritage. Events hypothesized to have occurred in the first few minutes of the creation of the universe turn out to have had profound influence on the birth, life, and death of galaxies, stars, and planets. Indeed, there is a direct, though tortuous, lineage from the forging of the matter of the universe in a primal furnace of incredible heat and light to the gathering on Earth of atoms versatile enough to serve as a chemical basis of life. The intrinsic harmony of the resultant worldview has great philosophical and aesthetic appeal and explains perhaps the modern resurgence of public interest in this subject.

For detailed information on the structure and evolution of the major components of the Cosmos, see GALAXIES, STARS AND STAR CLUSTERS, NEBULA, and SOLAR SYSTEM. The present article considers only aspects of these topics that satisfy one of three criteria: (1) they bear on the general issue of astronomical origins; (2) they are important to an integrated picture of how the universe evolved; or (3) they play an important role in forming humanity's growing vision of the miraculous unity that is the Cosmos.

The article is divided into the following sections:


History of humanity's perception of the universe 763
  Earliest conceptions 763
  Astronomical theories of the ancient Greeks 764
  The system of Aristotle and its impact on medieval thought 764
  The Copernican revolution 764
  Perceptions of the 20th century 765
Components of the universe 766
  Planetary systems 766
    The Sun
    Planets and their satellites
    Asteroids, meteoroids, comets, and interplanetary dust
    Origin of the solar system
    Extrasolar planetary systems
  Stars and the chemical elements 770
    Main-sequence structure of the stars
    The end states of stars
    The evolution of stars
    Interstellar clouds
    Star formation
  Galaxies 773
    The Milky Way Galaxy
    Classification of galaxies
    Dynamics of ellipticals and spirals
    Interacting galaxies
    Galaxy formation
  Quasars and related objects 777
    Extragalactic radio sources
    Quasars
    Black-hole model for active galactic nuclei
    Observational tests
  Other components 780
    Cosmic rays and magnetic fields
    Microwave background radiation
    Intergalactic gas
    Low-energy neutrinos
    Gravitational waves
Large-scale structure and expansion of the universe 783
  Clustering of galaxies 783
    The Local Group
    Neighbouring groups and clusters
    Superclusters
    Statistics of clustering
  Gravitational theories of clustering 785
    Modes of gravitational instability
    Top-down and bottom-up theories
  Unorthodox theories of clustering and galaxy formation 786
  The extragalactic distance scale and Hubble's constant 787
Cosmological models 787
  Early cosmological ideas 787
  Gravitation and the geometry of space-time 788
  Relativistic cosmologies 789
    Einstein's model
    De Sitter's model
    Friedmann–Lemaître models
    The Einstein–de Sitter universe
    Bound and unbound universes and the closure density
    The age of the universe
    Global observational tests
    The ultimate fate of the universe
  The hot big bang 791
    Primordial nucleosynthesis
    The deuterium abundance
  The very early universe 792
    Matter-antimatter asymmetry
    Superunification and the Planck era
    Inflation
  The steady state theory and other alternative cosmologies 794
Summary 794
Bibliography 794


## History of humanity's perception of the universe

### EARLIEST CONCEPTIONS

All scientific thinking on the nature of the Cosmos can be traced to the distinctive geometric patterns formed by the stars in the night sky. Even prehistoric people must have noticed that, apart from a daily rotation (which is now understood to arise from the spin of the Earth), the stars did not seem to move with respect to one another: the stars appear "fixed." Early nomads found that knowledge of the constellations could guide their travels, and they developed stories to help them remember the relative positions of the stars in the night sky. These stories became the mythical tales that are part of most cultures.

When nomads turned to farming, an intimate knowledge of the constellations served a new function—an aid in timekeeping, in particular, for keeping track of the seasons. People had noticed very early that certain celestial objects did not remain stationary relative to the "fixed" stars; instead, during the course of a year, they moved forward and backward in a narrow strip of the sky that contained 12 constellations constituting the signs of the zodiac. Seven such wanderers were known to the ancients:

the Sun, Moon, Mercury, Venus, Mars, Jupiter, and Saturn. Foremost among the wanderers was the Sun: day and night came with its rising and setting, and its motion through the zodiac signaled the season to plant and the season to reap. Next in importance was the Moon: its position correlated with the tides and its shape changed intriguingly over the course of a month. The Sun and Moon had the power of gods; why not then the other wanderers? Thus probably arose the astrological belief that the positions of the planets (from the Greek word *planetes,* "wanderers") in the zodiac could influence worldly events and even cause the rise and fall of kings. In homage to this belief, Babylonian priests devised the week of seven days, whose names even in various modern languages (for example, English, French, or Norwegian) can still easily be traced to their origins in the seven planet-gods.

### ASTRONOMICAL THEORIES OF THE ANCIENT GREEKS

The apex in the description of planetary motions during classical antiquity was reached with the Greeks, who were of course superb geometers. Like their predecessors, Greek astronomers adopted the natural picture, from the point of view of an observer on Earth, that the Earth lay motionless at the centre of a rigidly rotating celestial sphere (to which the stars were "fixed"), and that the complex to-and-fro wanderings of the planets in the zodiac were to be described against this unchanging backdrop. They developed an epicyclic model that would reproduce the observed planetary motions with quite astonishing accuracy. The model invoked small circles on top of large circles, all rotating at individual uniform speeds, and it culminated about AD 140 with the work of Ptolemy, who introduced the ingenious artifact of displaced centres for the circles to improve the empirical fit. Although the model was purely kinematic and did not attempt to address the dynamical reasons for why the motions were as they were, it laid the groundwork for the paradigm that nature is not capricious but possesses a regularity and precision that can be discovered from experience and used to predict future events.

tolemy
nd the
picyclic
iodel

The application of the methods of Euclidean geometry to planetary astronomy by the Greeks led to other schools of thought as well. Pythagoras (*c.* 570–? BC), for example, argued that the world could be understood on rational principles ("all things are numbers"); that it was made of four elements—earth, water, air, and fire; that the Earth was a sphere; and that the Moon shone by reflected light. In the 4th century BC Heracleides, a follower of Pythagoras, taught that the spherical Earth rotated freely in space and that Mercury and Venus revolved about the Sun. From the different lengths of shadows cast in Syene and Alexandria at noon on the first day of summer, Eratosthenes (*c.* 276–194 BC) computed the radius of the Earth to an accuracy within 20 percent of the modern value. Starting with the size of the Earth's shadow cast on the Moon during a lunar eclipse, Aristarchus of Samos (*c.* 310–230 BC) calculated the linear size of the Moon relative to the Earth. From its measured angular size, he then obtained the distance to the Moon. He also proposed a clever scheme to measure the size and distance of the Sun. Although flawed, the method did enable him to deduce that the Sun is much larger than the Earth. This deduction led Aristarchus to speculate that the Earth revolves about the Sun rather than the other way around.

he
eliocentric
ypothesis
f
ristarchus

Unfortunately, except for the conception that the Earth is a sphere (inferred from the Earth's shadow on the Moon always being circular during a lunar eclipse), these ideas failed to gain general acceptance. The precise reasons remain unclear, but the growing separation between the empirical and aesthetic branches of learning must have played a major role. The unparalleled numerical accuracy achieved by the theory of epicyclic motions for planetary motions lent great empirical validity to the Ptolemaic system. Henceforth, such computational matters could be left to practical astronomers without the necessity of having to ascertain the physical reality of the model. Instead, absolute truth was to be sought through the Platonic ideal of pure thought. Even the Pythagoreans fell into this trap; the depths to which they eventually sank may be judged from the story that they discovered and then tried to conceal the fact that the square root of 2 is an irrational number (*i.e.,* cannot be expressed as a ratio of two integers).

### THE SYSTEM OF ARISTOTLE AND ITS IMPACT ON MEDIEVAL THOUGHT

The systematic application of pure reason to the explanation of natural phenomena reached its extreme development with Aristotle (384–322 BC), whose great system of the world later came to be regarded as the synthesis of all worthwhile knowledge. Aristotle argued that humans could not inhabit a moving and rotating Earth without violating commonsense perceptions. Moreover, in his theory of impetus, all terrestrial motion, presumably including that of the Earth itself, would grind to a halt without the continued application of force. He took for granted the action of friction because he would not allow the seminal idealization of a body moving through a void ("nature abhors a vacuum"). Thus, Aristotle was misled into equating force with velocity rather than, as Sir Isaac Newton was to show much later, with (mass times) acceleration. Celestial objects were exempt from dynamical decay because they moved in a higher stratum whereby a perfect sphere was the natural shape of heavenly bodies and uniform rotation in circles was the natural state of their motion. Indeed, primary motion was derived from the outermost sphere, the seat of the unchangeable stars and of divine power. No further explanation was needed beyond the aesthetic one. In this scheme, the imperfect motion of comets had to be postulated as meteorological phenomena that took place within the imperfect atmosphere of the Earth.

The great merit of Aristotle's system was its internal logic, a grand attempt to unify all branches of human knowledge within the scope of a single self-consistent and comprehensive theory. Its great weakness was that its rigid arguments rested almost entirely on aesthetic grounds; it lacked a mechanism by which empirical knowledge gained from experimentation or observation could be used to test, modify, or reject the fundamental principles underlying the theory. Aristotle's system had the underlying philosophical drive of modern science without its flexible procedure of self-correction that allows the truth to be approached in a series of successive approximations.

With the fall of the Roman Empire in AD 476, much of what was known to the Greeks was lost or forgotten—at least to Western civilizations. (Hindu astronomers still taught that the Earth was a sphere and that it rotated once daily.) The Aristotelian system, however, resonated with the teachings of the Roman Catholic Church during the Middle Ages, especially in the writings of St. Thomas Aquinas in the 13th century, and later, during the period of the Counter-Reformation in the 16th and early 17th century, it ascended to the status of religious dogma. Thus did the notion of an Earth-centred universe become gradually enmeshed in the politics of religion. Also welcome in an age that insisted on a literal interpretation of the Scriptures was Aristotle's view that the living species of the Earth were fixed for all time. What was not accepted was Aristotle's argument on logical grounds that the world was eternal, extending infinitely into the past and the future even though it had finite spatial extent. For the church, there was definitely a creation event, and infinity was reserved for God, not space or time.

### THE COPERNICAN REVOLUTION

The Renaissance brought a fresh spirit of inquiry to the arts and sciences. Explorers and travelers brought home the vestiges of classical knowledge that had been preserved in the Muslim world and the East, and in the 15th century Aristarchus' heliocentric hypothesis again came to be debated in certain educated circles. The boldest step was taken by the Polish astronomer Nicolaus Copernicus, who hesitated for so long in publication that he did not see a printed copy of his own work until he lay on his deathbed in 1543. Copernicus recognized more profoundly than anyone else the advantages of a Sun-centred planetary system. By adopting the view that the Earth circled the Sun, he could qualitatively explain the to-and-fro wanderings of the planets much more simply than Ptolemy. For example, at certain times in the motions of the Earth

and Mars about the Sun, the Earth would catch up with Mars's projected motion, and then that planet would appear to go backward through the zodiac. Unfortunately in his Sun-centred system, Copernicus continued to adhere to the established tradition of using uniform circular motion, and if he adopted only one large circle for the orbit of each planet, his calculated planetary positions would in fact be quantitatively poorer in comparison with the observed positions of the planets than tables based on the Ptolemaic system. This defect could be partially corrected by providing additional smaller circles, but then much of the beauty and simplicity of Copernicus' original system would be lost. Moreover, though the Sun was now removed from the list of planets and the Earth added, the Moon still needed to move around the Earth.

It was Galileo who exploited the power of newly invented lenses to build a telescope that would accumulate indirect support for the Copernican viewpoint. Critics had no rational response to Galileo's discovery of the correlation of Venus' phases of illumination with its orbital position relative to the Sun, which required it to circle that body rather than the Earth. Nor could they refute his discovery of the four brightest satellites of Jupiter (the so-called Galilean satellites), which demonstrated that planets could indeed possess moons. They could only refuse to look through the telescope or refuse to see what their own eyes told them.

Galileo also mounted a systematic attack on other accepted teachings of Aristotle by showing, for example, that the Sun was not perfect but had spots. Besieged on all sides by what it perceived as heretical stirrings, the church forced Galileo to recant his support of the heliocentric system in 1633. Confined to house arrest during his last years, Galileo would perform actual experiments and thought experiments (summarized in a treatise) that would refute the core of Aristotelian dynamics. Most notably, he formulated the concept that would eventually lead (in the hands of René Descartes) to the so-called first law of mechanics—namely, that a body in motion, freed from friction and from all other forces, would move, not in a circle, but in a straight line at uniform speed. The frame of reference for making such measurements was ultimately the "fixed stars." Galileo also argued that, in the gravitational field of the Earth and in the absence of air drag, bodies of different weights would fall at the same rate. This finding would eventually lead (in the hands of Einstein) to the principle of equivalence, a cornerstone of the theory of general relativity.

Galileo's refutation of Aristotelian dynamics

It was the German astronomer Johannes Kepler, a contemporary of Galileo, who would provide the crucial blow that assured the success of the Copernican revolution. Of all the planets whose orbits Copernicus had tried to explain with a single circle, Mars had the largest departure (the largest eccentricity, in astronomical nomenclature); consequently, Kepler arranged to work with the foremost observational astronomer of his day, Tycho Brahe of Denmark, who had accumulated over many years the most precise positional measurements of this planet. When Kepler finally gained access to the data upon Tycho's death, he painstakingly tried to fit the observations to one curve after another. The work was especially difficult because he had to assume an orbit for the Earth before he could self-consistently subtract the effects of its motion. Finally, after many close calls and rejections, he hit upon a simple, elegant solution—an ellipse with the Sun at one focus. The other planets also fell into place. This triumph was followed by others, notable among which was Kepler's discovery of his so-called three laws of planetary motion. The empirical victory secure, the stage was set for Newton's matchless theoretical campaigns.

Two towering achievements paved the way for Newton's conquest of the dynamical problem of planetary motions: his discoveries of the second law of mechanics and of the law of universal gravitation. The second law of mechanics generalized the work of Galileo and Descartes on terrestrial dynamics, asserting how bodies generally move when they are subjected to external forces. The law of universal gravitation generalized the work of Galileo and the English physicist Robert Hooke on terrestrial gravity, asserting that two massive bodies attract one another with a force directly proportional to the product of their masses and inversely proportional to the square of their separation distance. By pure mathematical deduction, Newton showed that these two general laws (whose empirical basis rested in the laboratory) implied, when applied to the celestial realm, Kepler's three laws of planetary motion. This brilliant coup completed the Copernican program to replace the old worldview with an alternative that was far superior, both in conceptual principle and in practical application. In the same stroke of genius, Newton unified the mechanics of heaven and Earth and initiated the era of modern science.

The contributions of Newton

In formulating his laws, Newton asserted as postulates the notions of absolute space (in the sense of Euclidean geometry) and absolute time (a mathematical quantity that flows in the universe without reference to anything else). A kind of relativity principle did exist ("Galilean relativity") in the freedom to choose different inertial frames of reference—*i.e.,* the form of Newton's laws was unaffected by motion at constant velocity with respect to the "fixed stars." However, Newton's scheme unambiguously sundered space and time as fundamentally separate entities. This step was necessary for progress to be made, and it was such a wonderfully accurate approximation to the truth for describing motions that are slow compared to the speed of light that it withstood all tests for more than two centuries.

In 1705 the English astronomer Edmond Halley used Newton's laws to predict that a certain comet last seen in 1682 would reappear 76 years later. When Halley's comet returned on Christmas night 1758, many years after the deaths of both Newton and Halley, no educated person could ever again seriously doubt the power of mechanistic explanations for natural phenomena. Nor would anyone worry again that the unruly excursions of comets through the solar system would smash the crystalline spheres that earlier thinkers had mentally constructed to carry planets and the other celestial bodies through the heavens. The attention of professional astronomers now turned increasingly toward an understanding of the stars.

In the latter effort, the British astronomer William Herschel and his son John led the assault. The construction of ever more powerful reflecting telescopes allowed them during the late 1700s and early 1800s to measure the angular positions and apparent brightnesses of many faint stars. In an earlier epoch, Galileo had turned his telescope to the Milky Way and saw that it was composed of countless individual stars. Now the Herschels began an ambitious program to gauge quantitatively the distribution of the stars in the sky. On the assumption (first adopted by the Dutch mathematician and scientist Christiaan Huygens) that faintness is a statistical measure of distance, they inferred the enormous average separations of stars. This view received direct confirmation for the nearest stars through parallax measurements of their distances from the Earth. Later, photographs taken over a period of many years also showed that some stars changed locations across the line of sight relative to the background; thus, astronomers learned that stars are not truly fixed, but rather have motions with respect to one another. These real motions—as well as the apparent ones due to parallax, first measured by the German astronomer Friedrich Bessel in 1838—were not detected by the ancients because of the enormous distance scale of the stellar universe.

## PERCEPTIONS OF THE 20TH CENTURY

The statistical studies based on these new perceptions continued into the early 20th century. They culminated with the analysis by the Dutch astronomer Jacobus Cornelius Kapteyn who, like William Herschel before him, used number counts of stars to study their distribution in space. It can be shown for stars with an arbitrary but fixed mixture of intrinsic brightnesses that—in the absence of absorption of starlight—the number $N$ of stars with apparent brightness, energy flux $f$, larger than a specified level $f_0$, is given by $N = Af_0^{-3/2}$, where $A$ is a constant, if the stars are distributed uniformly in Euclidean space (space satisfying the principles of Euclidean geometry).

The statistical studies of J.C. Kapteyn

The number $N$ would increase with decreasing limiting apparent brightness $f_0$, because one is sampling, on average, larger volumes of space when one counts fainter sources. Kapteyn found that the number $N$ increased less rapidly with decreasing $f_0$ than the hypothetical value $Af_0^{-3/2}$; this indicated to him that the solar system lay near the centre of a distribution of stars, which thinned in number with increasing distance from the centre. Moreover, Kapteyn determined that the rate of thinning was more rapid in certain directions than in others. This observation, in conjunction with other arguments that set the scale, led him in the first two decades of the 20th century to depict the Milky Way Galaxy (then confused with the entire universe) as a rather small, flattened stratum of stars and gaseous nebulas in which the number of stars decreased to 10 percent of their central value at a distance in the plane of about 8,500 light-years from the galactic centre.

In 1917 the American astronomer Harlow Shapley mounted a serious challenge to the Kapteyn universe. Shapley's study of the distances of globular clusters led him to conclude that their distribution centred on a point that lay in the direction of the constellation Sagittarius and at a distance that he estimated to be about 45,000 light-years (50 percent larger than the modern value). Shapley was able to determine the distance to the globulars through the calibration of the intrinsic brightnesses of some variable stars found in them. (Knowing the period of the light variations allowed Shapley to infer the average intrinsic brightness. A measurement of the average apparent brightness then allowed, from the $1/r^2$ law of brightness, a deduction of the distance $r$.) According to Shapley, the galactic system was much larger than Kapteyn's estimate. Moreover, the Sun was located not at its centre but rather at its radial outskirts (though close to the midplane of a flattened disk). Shapley's dethronement of the Sun from the centre of the stellar system has often been compared with Copernicus' dethronement of the Earth from the centre of the planetary system, but its largest astronomical impact rested with the enormous physical dimensions ascribed to the Galaxy. In 1920 a debate was arranged between Shapley and Heber D. Curtis to discuss this issue before the National Academy of Sciences in Washington, D.C.

'he
hapley-
'urtis
ebate

The debate also addressed a second controversy—the nature of the so-called spiral nebulas. Shapley and his adherents held that these objects were made up of diffuse gas and were therefore similar to the other gas clouds known within the confines of the Milky Way Galaxy. Curtis and others, by contrast, maintained that the spirals consisted of stars and were thus equivalent to independent galaxies coequal to the Galaxy. A parallel line of thought had been proposed earlier by the philosophers Immanuel Kant and Thomas Wright and by William Herschel. The renewed argument over the status of the spirals grew in part out of an important development that occurred around the turn of the 20th century: the astronomical incorporation of the methods of spectroscopy both to study the physical nature of celestial bodies and to obtain the component of their velocities along the line of sight. By analyzing the properties of spectral lines in the received light (*e.g.*, seeing if the lines were produced by absorption or emission and if the lines were broad or narrow), or by analyzing the gross colours of the observed object, astronomers learned to distinguish between ordinary stars and gaseous nebulas existing in the regions between stars. By measuring the displacement in wavelength of the spectral lines with respect to their laboratory counterparts and assuming the displacement to arise from the Doppler effect, they could deduce the velocity of recession (or approach). The spirals posed interpretative difficulties on all counts: they had spectral properties that were unlike either local collections of stars or gaseous nebulas (because of the unforeseen roles of dust and different populations of stars in the arms, disk, and central bulge of a spiral galaxy); and, as had been shown by the American astronomer Vesto Slipher, they generally possessed recession velocities that were enormous compared to those then known for any other astronomical object.

The formal debate between Shapley and Curtis ended inconclusively, but history has proved Shapley to be mostly right on the issue of the off-centre position of the solar system and the large scale of the Galaxy, and Curtis to be mostly right on the issue of the nature of the spirals as independent galaxies. As demonstrated in the work of the Swiss-born U.S. astronomer Robert J. Trumpler in 1930. Kapteyn (and Herschel) had been misled by the effects of the undiscovered but pervasive interstellar dust to think that the stars in the Milky Way thinned out with distance much more quickly than they actually do. The effect of interstellar dust was much less important for Shapley's studies because the globular clusters mostly lie well away from the plane of the Milky Way system.

The decisive piece of evidence concerning the extragalactic nature of the spirals was provided in 1923–24 by Hubble, who succeeded in resolving one field in the Andromeda galaxy (M31) into a collection of distinct stars. Some of the stars proved to be variables of a type similar to those found by Shapley in globular clusters. Measurements of the properties of these variables yielded estimates of their distances. As it turned out, the distance to M31 put it well outside the confines of even Shapley's huge model of the Galaxy, and M31 therefore must be an independent system of stars (and gas clouds).

Hubble's research on extragalactic systems

Hubble's findings inaugurated the era of extragalactic astronomy. He himself went on to classify the morphological types of the different galaxies he found: spirals, ellipticals, and irregulars. In 1926 he showed that, apart from a "zone of avoidance" (region characterized by an apparent absence of galaxies near the plane of the Milky Way caused by the obscuration of interstellar dust), the distribution of galaxies in space is close to uniform when averaged over sufficiently large scales, with no observable boundary or edge. The procedure was identical to that used by Kapteyn and Herschel, with galaxies replacing stars as the luminous sources. The difference was that this time the number count $N$ was proportional to $f_0^{-3/2}$, to the limits of the original survey. Hubble's finding provided the empirical justification for the so-called cosmological principle, a term coined by the English mathematician and astrophysicist Edward A. Milne to describe the assumption that at any instant in time the universe is, in the large, homogeneous and isotropic—*i.e.*, statistically the same in every place and in every direction. This represented the ultimate triumph for the Copernican revolution.

Cosmological principle

It was also Hubble who interpreted and quantified Slipher's results on the large recessional velocities of galaxies—they correspond to a general overall expansion of the universe. The Hubble law, enunciated in 1929, marked a major turning point in modern thinking about the origin and evolution of the Cosmos. The announcement of cosmological expansion came at a time when scientists were beginning to grapple with the theoretical implications of the revolutions taking place in physics. In his theory of special relativity, formulated in 1905, Einstein had effected a union of space and time, one that fundamentally modified Newtonian perceptions of dynamics, allowing, for example, transformations between mass and energy. In his theory of general relativity, proposed in 1916, Einstein effected an even more remarkable union, one that fundamentally altered Newtonian perceptions of gravitation, allowing gravitation to be seen, not as a force, but as the dynamics of space-time. Taken together, the discoveries of Hubble and Einstein gave rise to a new worldview. The new cosmology gave empirical validation to the notion of a creation event; it assigned a numerical estimate for when the arrow of time first took flight; and it eventually led to the breathtaking idea that everything in the universe could have arisen from literally nothing (see below).

## Components of the universe

### PLANETARY SYSTEMS

Although it is commonly believed that planetary systems are plentiful in the universe, the only example known with certainty is the solar system. The solar system is conventionally taken to contain the Sun, the nine planets and their satellites, asteroids, comets, interplanetary dust, and interplanetary particles and fields largely associated

with the solar wind. Humanity's knowledge of these objects has expanded greatly due to space exploration. Combined with centuries of intense astronomical observation and theoretical calculation, data transmitted by spacecraft have shed considerable insight into the relation between the solar system and the rest of the universe, the problem of the origin of the Earth and the other planets, and the question of the likelihood of comparable planetary systems around other stars.

**The Sun.** At the centre of the solar system lies the Sun. Energetically and dynamically, it is the dominant influence, and no deep understanding of the system can be achieved without an appreciation of the basic properties of the Sun. Moreover, because stars constitute the basic building blocks of the Cosmos and the Sun is in many respects a typical star, it provides a link in the chain of being between the planets and the rest of the universe.

The mass and composition of the Sun

The mass of the Sun can be measured from its gravitational pull on the planets and equals $1.99 \times 10^{33}$ g (grams); thus, it is 1,000 times more massive than Jupiter and 330,000 times more massive than the Earth. As a fraction of its mass, the atmospheric composition of the Sun is probably 72 percent hydrogen and 26 percent helium, with the remaining 2 percent composed of elements heavier than hydrogen and helium. Because there is little mixing between atmosphere and deep interior, this composition is believed to be the one that the Sun was born with. A gas with approximately the solar mix of elements is said to have cosmic abundances because a similar composition is found spectroscopically for most other stars as well as for the medium between the stars.

The percentage of helium in the Sun's central regions is believed to be larger than the cosmic proportion because of the conversion of hydrogen into helium by thermonuclear fusion. Except for the emission of neutrinos, however, the interior of the Sun is too opaque to radiation of all types to be directly observable. Neutrinos produced in nuclear reactions at the centre of the Sun could in principle provide a tracer of the rate of such reactions, because neutrinos interact relatively little with matter. Unfortunately recent attempts to measure solar neutrinos via radio-chlorine techniques have found levels that are only about one-third the best theoretical predictions. The source of the discrepancy is uncertain.

The observed rate of release of radiant energy by the Sun equals $3.86 \times 10^{33}$ erg/sec (ergs per second). The particles of radiation (photons) stream more or less freely from a layer called the photosphere, which, in the Sun, is at a temperature of about 5,800 K (roughly 5,500° C or 10,000° F). The distribution of wavelengths is characteristic of a thermal body radiating at such a temperature; therefore, in accordance with Planck's law, it peaks in the yellow part of the visible spectrum. The solar luminosity is enormous, but it is much less than it would be if the photons in the hot interior of the Sun could also stream freely. However, the high opacity of the material regulates the actual outward progress of the photons to a slow stately diffusion. Indeed, the blockage of diffusive heat is so severe in the envelope of the Sun that the component layers are unstable to the development of convection currents. The presence of convection cells in the Sun can be seen in high-resolution photographs that show its outer layers to have a granular appearance.

The observed radius of the Sun equals $6.95 \times 10^{10}$ centimetres and is understood to be the result of a balance of forces between the Sun's self-gravity and the interior pressure of its hot gases, which exist in a nearly fully ionized state (a plasma of positive ions and free electrons) in the deep interior. The plasma in the core of the Sun is compressed to temperatures (about $10^7$ K) that are sufficient to provide a rate of thermonuclear reactions that just offsets the slow diffusive loss of radiative heat. Thus, the Sun constitutes a controlled fusion reactor capable of sustaining its present steady loss of radiant energy for a full 9,000,000,000 years before all of its initial supply of hydrogen fuel in the core has been converted into helium. From the radioactive dating of meteorites, it has been estimated that the solar system is 4,600,000,000 years old. If this is the age of the Sun, then it is roughly midway through the phase of stable core hydrogen fusion—*i.e.*, the "main-sequence" phase of stellar evolution.

Controlled hydrogen fusion in the Sun's core

The equator of the Sun turns once every 25 days and its poles once every 35 days. This differential rotation may couple with the Sun's convection zone to produce a dynamo action that amplifies magnetic fields. The basic idea is that magnetic fields carried upward (or downward) by convection currents are twisted and amplified by the differential rotation. "Ropes" of high field strength buoy to the surface where they pop out as loops into the corona of the Sun. The corona is an extended region containing very rarefied gas that lies above the photosphere and a transition region called the chromosphere; the temperature of the corona is about 2,000,000 K. The anchor points of the ropes of high magnetic flux in the photosphere correspond to sunspots, regions where the gas is cooler than the average photospheric temperature of 5,800 K. Thus, these spots appear relatively dark against the bright yellow background of the general photosphere.

Sunspots and other manifestations of magnetic activity

Sunspots appear, migrate about the solar surface, and disappear as the plasma to which they are anchored moves under the influence of rotation and convection. The average number of sunspots increases and decreases more or less regularly in an 11-year cycle; however, there have been prolonged minimas in history. It has been proposed that such minimas correlate with changing climate conditions on the Earth, although the precise mechanisms for effecting such changes remain unclear.

Other manifestations of magnetic activity arise because of the motion of the flux ropes. It is believed that flares occur on those occasions when two flux ropes of opposite polarity are pressed against each other, and the opposing magnetic fields annihilate in a catastrophic event of magnetic reconnection. The energy stored in the field is thought to go into accelerating fast particles (solar cosmic rays) and into heating the ambient gas, which, being rarefied, has very little heat capacity. Magnetic activity of this type may be what maintains the corona at much higher temperatures than the photosphere.

Pictures of the solar corona taken during one of the U.S. manned Skylab missions (1973) showed that hot coronal gas trapped in closed loops of field lines becomes dense enough to emit appreciable amounts of X rays. In contrast, coronal holes lacking X-ray emission correspond to regions where the magnetic field is too weak to keep the gas trapped, and the hot gas has burst open the magnetic-field configuration, expanding away from the surface of the Sun as part of a general solar wind.

The presence of a solar wind blowing through interplanetary space was first deduced from observations of the ion tails of comets made during the 1950s. With the advent of Earth-orbiting satellites, the particles and fields carried by the solar wind could be measured directly. When the wind blows past the Earth, it contains on average about five particles per cubic centimetre (mostly protons, the nuclei of hydrogen atoms) moving at about 500 km/sec (kilometres per second), but these numbers fluctuate greatly depending on the phase of the solar magnetic cycle and the presence or absence of recent flare activity.

**Planets and their satellites.** Clues to the question of how the planets were formed lie in the regularities of their orbital motions, their satellite systems, and their chemical compositions. Compared to their sizes, the separations of planets from each other are enormous; and, apart from a diffuse solar wind and minor debris, interplanetary space is remarkably empty. Consequently, the planets have been well isolated dynamically and chemically since their birth, and the present configuration of the solar system provides hints of the initial conditions, notwithstanding the more than 4,000,000,000 years of subsequent evolution.

With the exception of Mercury and Pluto, the orbits of the planets are all nearly circular; they lie within a few degrees of the same plane; and they have the same direct sense of revolution as the rotation of the Sun. Since these facts were first noted, they have suggested to philosophers and scientists such as Kant and Pierre-Simon Laplace that the planets of the solar system must have originally formed from a flat nebular disk that revolved about the primitive Sun. The exceptions, Mercury and Pluto, are not trou-

blesome. They both have resonant interactions (Mercury with the Sun and Pluto with Neptune), which may have modified considerably their original orbital characteristics.

The spins of the planets are more problematic. The spins of most of the planets are direct, with the equators lying in roughly the same plane as both their orbits about the Sun and the orbits of their (larger) satellites about them. However, Venus spins in a retrograde direction, Uranus' spin is tilted over on its side with respect to its orbit, and Triton orbits Neptune in the opposite sense to the planet's spin. In reconstructing the details of the formation of the solar system, astronomers are therefore undecided whether to take these exceptions more seriously than the rule.

Differences between terrestrial and Jovian planets

In the inner planetary system where the terrestrial planets—Mercury, Venus, Earth, and Mars—reside, the distance between successive planets is relatively small in comparison with the outer planetary system where the Jovian planets—Jupiter, Saturn, Uranus, and Neptune—reside. Moreover, the terrestrial planets are small and rocky or ironlike, while the Jovian planets (also called the giant planets) are large and gaseous or icy. Neither the terrestrial nor the Jovian planets exhibit the chemical elements in their cosmic proportions, but the latter, particularly Jupiter and Saturn, approach these proportions to a much closer degree. This implies that the process of planet building, unlike the mechanism of star formation, probably involves forces other than gravity, for gravitation is universal and does not distinguish between different elements if they are in a gaseous form. Condensation (*i.e.*, the separation of solid phases of matter from gaseous phases if the temperature drops to sufficiently low values) suggests itself as an important process.

From this point of view, the terrestrial planets have managed only to gather into their bodies mostly materials containing elements heavier than hydrogen and helium—materials such as silicate rocks and metallic iron or nickel, which can condense as solids from a gaseous phase even at relatively high temperatures (between 1,200 and 2,000 K). In contrast, Uranus and Neptune have not only accumulated rocky and metallic compounds but also ices of water, ammonia, and methane that can condense from nebular gas only at much lower temperatures (between 100 and 200 K). Jupiter and Saturn succeeded additionally in capturing substantial amounts of hydrogen and helium (in their envelopes). Since hydrogen and helium do not solidify at plausible nebular pressures unless the temperature is even lower than in the coldest regions of interstellar space, this suggests that in the two largest planets of the solar system gravitation did play a role in the direct acquisition of massive amounts of these gases.

Pluto, which is small and icy and resides in the outer planetary system relatively well separated from Neptune, is not readily classifiable in the scheme outlined above. The discrepancy is not disruptive, however, because Pluto is a comparatively minor body. Possible explanations for Pluto are that it is an escaped moon of Neptune or that it is a planet whose growth was stunted by being at the outermost fringe of the original solar nebula.

The terrestrial and Jovian planets possess another systematic difference: the former generally have no rings or satellites, while the latter each have a set of rings and several satellites. Here, Earth and Mars are exceptions to the rule. Earth has of course one satellite, the Moon; Mars has two, Phobos and Deimos. Of these exceptions, the more difficult case to explain has long remained the Moon because it is an unusually large object for a satellite. Indeed, the Moon is only somewhat smaller than the largest and most massive satellites in the solar system: Jupiter's Ganymede, Saturn's Titan, and Neptune's Triton. In comparison, Phobos and Deimos are tiny objects that may well have been captured after Mars had already formed.

Like Earth, Pluto has a satellite, Charon. Discovered in 1978, Charon is not much smaller than Pluto. It has been speculated that Pluto may be spinning rapidly enough to allow Charon to have fissioned from it. A similar hypothesis seems less plausible for the Earth–Moon system because if the Moon and the angular momentum that it carries were to be put into the Earth, the combined body would still not be rotating fast enough to fission.

The satellite and ring systems of the giant planets, particularly those of Jupiter and Saturn, resemble miniature planetary systems. As an analogy, one may say that moons and rings are to the giant planets what the planets and the asteroid belt are to the Sun. The moons of the giant planets can be classified as either regular or irregular. The regular satellites generally have nearly circular orbits and revolve in the same direction as that of the parent planet's rotation and in the same plane as the equator of the planet. The irregular satellites violate one or more of the above rules. In addition, they generally tend to be small bodies and to lie at large distances from the central planet. The regular satellites may have formed from protoplanetary disks that encircled the planet in the same manner as a protostellar disk encircled the Sun in the nebular hypothesis. The most likely explanation for the irregular satellites is that they are captured bodies.

Regular and irregular satellites

The thin flat rings that encircle Jupiter, Saturn, Uranus, and Neptune (actually Neptune's rings seem to be only partial arcs) are composed of innumerable small solid bodies. Each piece of the ring is in a nearly perfect circular orbit about the central planet. Theory suggests that noncircular motions are damped by mutual inelastic collisions of the particulate matter to very small values. These collisions would have led to gradual agglomeration into larger bodies had the rings not lain in such close proximity to the planet (*i.e.*, within the Roche limit). The strong tidal forces that exist inside the Roche limit of a planet are believed to be capable of tearing apart loosely bound aggregates of particulate matter and thereby preventing their agglomeration into moons. It is unclear, however, whether planetary rings are the natural debris left over from an earlier period of satellite formation in a protoplanetary disk that extended almost to the planet's surface, or whether they arose from the more recent breakup and erosion (by micrometeoroid bombardment) of some larger parent body. There does exist some evidence from dynamical studies of the gravitational interactions of the rings and satellites of Saturn that the rings may be appreciably younger than the solar system in general.

**Asteroids, meteoroids, comets, and interplanetary dust.** In addition to the Sun and its wind and the nine planets and their satellites, the solar system contains a large number of minor bodies. The most conspicuous of these are the asteroids and comets. Smaller bodies also exist—meteoroids, micrometeoroids, and interplanetary dust—but these probably are fragments of the larger asteroids and comets. Indeed, there is a continuous distribution of minor bodies in the solar system, from dust particles with radii of only a fraction of a micrometre to asteroids (or minor planets) with radii of several hundred kilometres.

Asteroids are rocky or iron-bearing bodies that can be found orbiting the Sun in great numbers in a belt between Mars and Jupiter. Nearly all of the total mass of the asteroids, about $10^{-3}$ that of the Earth, is contained in the largest examples such as Ceres, Pallas, and Vesta, but the largest numbers have radii of one to 10 kilometres (the lower limit being more a matter of nomenclature than of measurement). A few bodies, as, for example, Chiron, lie outside the belt between Mars and Jupiter. The exceptions, however, are relatively few. The theoretical understanding of this observational result lies in computer simulations that show that an asteroid placed almost anywhere else in the solar system besides the known asteroid belt would be unstable to gravitational interactions with the planets. If the early solar system were littered with asteroid-sized bodies, the creation of the planets would have swept interplanetary space relatively clean except for the debris that happened to have orbits fit for survival.

Meteoroids are chunks of asteroids or comets that have Earth-crossing orbits. One theory for the production of meteoroids has them originating from the shattering of two asteroids that collide violently in space. Some of the pieces may subsequently undergo resonant interactions with Jupiter, which bring them into Earth-crossing orbits in $10^4$ to $10^5$ years. A meteoroid entering the Earth's atmosphere heats up during the passage and becomes a meteor, a fiery "shooting star." If the mass of the meteor exceeds one kilogram, it can survive the flight and land

on the ground as a meteorite. Meteorites come in three basic compositions: stones, stony irons, and irons. Radioactive dating of meteorites establishes that they have a narrow range of ages. The time since their parent bodies first solidified equals about 4,600,000,000 years, which is generally regarded as the age of the entire solar system.

Asteroids and meteoroids as possible building blocks of the terrestrial planets

The cratering records on the airless (and therefore erosion-free) Moon and Mercury are consistent with a very heavy period of meteoritic impacts during the first several hundred million years of the history of the solar system, with the bombardment tailing off dramatically about 4,000,000,000 years ago. This picture suggests that primitive asteroids and meteoroids may have been the building blocks ("planetesimals") of the terrestrial planets (and perhaps also the cores of the giant planets), and that the present-day asteroids failed to be gathered into another full-fledged planet because their random velocities are so high (probably due to the past action of Jupiter's gravitational perturbations) as to cause them generally to shatter rather than agglomerate when they collide.

Comets also are cosmic debris, probably planetesimals that originally resided in the vicinity of the orbits of Uranus and Neptune rather than in the warmer regions of the asteroid belt. Thus, the nuclei of comets are icy balls of frozen water, methane, and ammonia, mixed with small pieces of rock and dust, instead of the largely volatile-free stones and irons that typify asteroids. In the most popular theory, icy planetesimals in the primitive solar nebula that wandered close to Uranus or Neptune but not close enough to be captured by them were flung to great distances from the Sun, some to be lost from the solar system and others to populate what was to become

Oort cloud

a great cloud of cometary bodies, perhaps $10^{12}$ in number. The existence of such a cloud was first hypothesized by the Dutch astronomer Jan Hendrik Oort.

In the original version of the theory, the Oort cloud extended tens of thousands of times farther from the Sun than the Earth, a significant fraction of the way to the nearest stars. Random encounters with passing stars would periodically throw some of the comets into new orbits, plunging them back toward the heart of the solar system. As a comet nears the Sun, the ices begin to evaporate, loosening the trapped dust and forming a large coma that completely surrounds the small nucleus, which is the ultimate source of all the material. The solar wind blows back the evaporating gas into an ion tail, and radiation pressure pushes back the small particulate solids into a dust tail. Each solid particle is now an independently orbiting satellite of the Sun, and the accumulation of millions of such passages by many comets contributes to the total quantity of dust particles and micrometeoroids found in interplanetary space.

The total mass contained in all the comets is highly uncertain. Modern estimates range from one to 100 Earth masses. Part of the uncertainty concerns the reality of a hypothesized massive "inner Oort cloud" of comets, which would exist at distances from the Sun $10^3$ to $10^4$ times that of the orbit of the Earth. At such distances, the comets would not be much perturbed by typical passing stars nor by the gravity of the planets of the solar system, and the comets could reside in the inner Oort cloud for long periods of time without detection. It has been speculated, however, that a rare close passage by another star (possibly an undetected companion of the Sun) may send a shower of these comets streaming toward the inner solar system. If enough large cometary nuclei in such showers happen to strike the Earth, the clouds of dust and ash that they would raise might be sufficient to trigger mass biological extinctions. An event of this kind appears especially promising for explaining the relatively sudden disappearance of the dinosaurs from the Earth.

**Origin of the solar system.** Modern versions of the

Nebular hypothesis

nebular hypothesis all begin with the collapse of a rotating interstellar cloud that is destined to form the solar system. The tendency to conserve angular momentum causes the falling gas to spin faster and flatten, eventually forming a central concentration (protosun) surrounded by a rotating disk of matter. Calculations show that there may be a prolonged phase of infall that continues to build up a disk of increasing mass and size. There also may be some accretion of the material in the disk onto the star, the process transferring mass inward and angular momentum outward, which helps to explain why the Sun presently contains 99.9 percent of the total mass of the solar system but only 2 percent of the total angular momentum.

Because the chemical compositions of the planets as a function of increasing radial distance from the Sun follow a pattern that corresponds to sequential condensation from a gaseous state, cosmochemists like to postulate for simplicity that the solar nebula was initially hot and purely gaseous. Small pieces of solids are then imagined to condense from the gas in the disk as the latter slowly cooled from high temperatures, with the coolest final temperatures being reached at the greatest distance from the centre. The process is akin to soot forming out of a smoking candle flame. Astronomical observations, however, show that dust grains of approximately the correct composition already exist in the interstellar medium, and theoretical calculations indicate that the refractory cores of the grains would survive introduction into most of the primitive solar nebula. The icy mantles that coat the grain cores would, however, be evaporated away in the inner solar system. It is possible therefore that the systematics of the observed planetary compositions reflect, not a condensation sequence, but rather an evaporation sequence.

In any case, whether the dust particles form by chemical condensation from the nebular gas or exist from the start, there seems little doubt that they would grow rapidly by various agglomeration processes and dissipatively settle into a thin layer of particulate matter in the midplane of the disk. Planetesimals of the sizes of asteroids and the nuclei of comets accumulate in this thin layer and grow further by gravitational processes into full-sized planets. The formation of the planets under these dissipative circumstances would explain why their orbits are nearly coplanar and circular.

Insofar as the planets first grow by the accumulation of solids, it is interesting to note that observations indicate all four Jovian planets to have rocky and icy cores containing 15–20 Earth masses. Around these cores Jupiter and Saturn have hydrogen and helium envelopes amounting to about 300 and 80 Earth masses, respectively. This suggests, as theoretical calculations bear out, that 15–20 Earth masses represents a critical mass above which a growing planet in the solar nebula will begin to gravitationally gather nebular gas faster than it will accumulate solids. Indeed, once a protoplanet becomes massive enough, it can efficiently eject solid bodies as well as capture them. (The ones catapulted out by Jupiter and Saturn, in the way that the U.S. Pioneer and Voyager spacecraft to these planets were, are likely to escape the system altogether.) In this manner, by either capturing or ejecting much of the material around them, did Jupiter and Saturn become large and grow to occupy large spacings.

Why Uranus and Neptune did not also gather massive gaseous envelopes is somewhat of a mystery. Perhaps they, like the terrestrial planets, did not finish forming until after the solar wind began to blow and swept interplanetary space clean of gas and small solid particles. Many of the icy planetesimals flung out by Uranus and Neptune probably now populate the Oort cloud of comets; otherwise, only the asteroid belt remains to indicate the magnitude of the debris that used to litter the solar system.

There exist other substantive arguments that Jupiter, at least, finished forming before the terrestrial planets did. A massive body at Jupiter's orbit may so stir up the orbits of the planetesimals in the asteroid belt as to prevent them from accumulating into a larger body (see above). A fully formed Jupiter may also have stunted the growth of nearby Mars, explaining why the latter is so much smaller than either Venus or the Earth. (Mercury is small presumably because it is so close to the Sun; the ratio of its rocky mantle to iron–nickel core is less than that of Venus, Earth, or Mars.)

The most speculative application of the idea that the giant planets sent fairly large bodies careening through the early solar system has to do with the Moon's origin. In one version of the event, a Mars-sized body crashed into

the primitive Earth. The iron core sank to the centre of the molten Earth, and the rocky mantle went into orbit and re-accreted into the Moon. The formation of the Moon from the mantle would then explain why the lunar landing missions revealed the Moon to be much poorer in iron than the Earth.

**Extrasolar planetary systems.** Direct astronomical detection of planets around other stars is made difficult by the glare of the central star. At least two independent lines of evidence now exist, however, that are indirectly related to the existence of extrasolar planetary systems.

First, it is known from studies of gas clouds where stars are currently forming in the Galaxy that such regions generally rotate too quickly to collapse to a single normal star without any companions. There are many examples where the excess angular momentum has apparently been absorbed in the birth of a nearby orbiting star. Indeed, binary stars are known to be the most common outcome of the star-formation process. It is, nevertheless, encouraging that infrared searches for faint companions around apparently single stars have discovered numerous examples of objects that lie intermediate to the least massive normal star and a giant planet such as Jupiter.

Second, infrared searches with Earth-orbiting and ground-based telescopes have also found flattened distributions of particulate solids encircling young stars that resemble the dusty nebular disk long hypothesized for the origin of the solar system. In a few cases, radio-spectroscopic observations at millimetre wavelengths have detected gaseous molecular material coextant with the solid particulates. These observations lend strong support to the view that the creation of planetary systems is likely to be a common by-product of the process of star formation.

Detection of nebular disks around young stars

## STARS AND THE CHEMICAL ELEMENTS

Stars are the great factories of the universe. They gradually transform the raw material that emerged from the big bang into an array of versatile chemical elements that makes possible the birth of planets and their inhabitants. The empirical evidence for the vital role that stars play in nucleosynthesis lies in the spectroscopic analysis of the atmospheric compositions of different generations of stars. The oldest stars, which belong to globular clusters, possess very little in the way of elements heavier than hydrogen and helium—in some cases less than 1 percent of the value possessed by the Sun. On the other hand, the youngest stars, which have ages on the order of $10^6$ years, have heavy elements in even slightly greater abundance than the Sun. Astronomers give these results explicit recognition by designating stars with high heavy-element abundance as Population I stars; those with low heavy-element abundance are said to be Population II stars.

Synthesis of heavy elements in stellar interiors

The accepted interpretation of the abundance differences of Populations I and II is that stars synthesize heavy elements in their interiors. In the process of dying, some stars spew great quantities of this processed material into the gas clouds occupying the regions between the stars. The enriched matter then becomes incorporated into a new generation of forming stars, each successive generation having on average a greater proportion of heavy elements (and helium) than the last. During the 20th century astronomers have obtained considerable insight into why these processes should be the natural outcome of the structure and evolution of stars.

**Main-sequence structure of the stars.** The same general principles that determine the structure of the Sun apply more broadly to all normal stars: (1) Hydrostatic equilibrium—for a star to be mechanically in equilibrium, the internal pressure must balance the weight of the material on top. (2) Energy transfer—photons diffusively carry energy outward from a hot interior; if the luminosity to be carried exceeds the capacity of photon diffusion, convection ensues. (3) Energy balance—for a star to be thermally in equilibrium, the energy carried outward by radiative diffusion or convection must be balanced by an equal release of nuclear energy; if the rate of thermonuclear fusion is inadequate, gravitational contraction of the central regions will result, usually accompanied by an expansion of the outer layers.

Much of the time of the luminous stages of a star's life is spent on the main sequence, when it stably fuses hydrogen into helium in its core. The fusion process in a star slightly more massive than the Sun is somewhat different from that in a star of one solar mass or less. In high-mass stars, hydrogen fusion occurs at high temperatures using preexisting nuclei of carbon and nitrogen as catalysts and, in the process, converting much of the carbon into nitrogen. In low-mass stars, hydrogen fusion occurs by direct fusion of the hydrogen nuclei or their reaction products. The end product, however, is the same: the conversion of four hydrogen nuclei into one helium nucleus, with the release of the nuclear binding energy as a source of heat for the star.

The time that a low-mass star spends on the main sequence differs drastically from that of a high-mass star. On the main sequence, a low-mass star spends its nuclear resource thriftily; a high-mass star, prodigiously. Hence, core hydrogen exhaustion for low-mass stars is delayed in comparison to high-mass stars. The main-sequence lifetime of a star half as massive as the Sun is about $3 \times 10^{10}$ years, whereas that of a star 50 times more massive than the Sun would be roughly $3 \times 10^6$ years.

Since the lifetime of a high-mass star is much less than the age of the Milky Way Galaxy (roughly $10^{10}$ years) and since such stars exist during the present epoch in the Galaxy, the formation of high-mass stars must be an ongoing process. This is borne out by observations of the Galaxy and external galaxies, where bright blue stars are always found near giant clouds of gas and dust—the sites of both high-mass and low-mass star formation.

Study of star clusters

One of the most important tests for the theory of stellar structure and evolution comes from the examination of star clusters. Star clusters are gravitationally bound stellar groups that occur in two basic types: globular clusters, which typically are rich systems containing perhaps 1,000,000 members distributed in a compact spherical volume with a strong concentration toward the centre; and open clusters, which typically are poor systems containing 1,000 members or less distributed loosely throughout an irregular volume. Globular cluster stars belong to Population II, while open cluster stars belong to Population I.

A given cluster usually occupies a small angular extent in the sky; thus, the individual member stars can effectively be assumed to lie at the same distance. The distance of one open cluster relative to another can be obtained by comparing the apparent brightnesses of their main-sequence stars on the assumption that main-sequence stars of a given spectral type have the same intrinsic brightness. (In actual practice, corrections need to be made for obscuration by interstellar dust; indeed, it was from the study of open clusters that Robert Trumpler first derived irrefutable evidence for the general presence of interstellar extinction.) The absolute distance to all observable open clusters can then be obtained by measuring the distance to one of them. The astronomical standard for this purpose is the Hyades cluster, whose distance can be ascertained by other techniques (among them, the so-called moving cluster method). The distances to globular clusters are not obtained in this fashion, because they are generally too far away for their faint main-sequence stars to be measured accurately. Fortunately, globulars contain numerous variable stars (RR Lyrae variables), and the known absolute brightnesses of these objects allow a determination of their distances, as was first demonstrated by Shapley.

All astronomical observations of a star cluster indicate that its members formed from the same parent cloud. Thus, the stars in a cluster have the same age and initial compositions; the only notable difference between them is their masses. Since stars of different masses evolve at different rates, it should be possible to see a progression of evolutionary states as stars of increasing mass are considered. The effect is indeed seen, and the comparison of theoretical predictions with astronomical observations of star clusters yields one of the most satisfactory success stories of modern astrophysics. Such studies allow estimates of the ages of star clusters. The oldest turn out to be the globular clusters; they have ages estimated by various investigators between 10,000,000,000 and 18,000,000,000

Ages of star clusters

years. Within the errors of the determinations, the ages of globulars are consistent with the expansion age of the universe—roughly $10^{10}$ years—obtained from Hubble's law. Thus, the globular cluster stars in the Galaxy must constitute some of the oldest stars in the Cosmos.

On the main sequence, a high-mass star is not only much more luminous than a low-mass star but it also appears much bluer because its surface temperature is a few tens of thousands degrees instead of a few thousand degrees. The difference in surface temperature manifests itself not only in broadband colours but also in the pattern of atomic absorption lines that appear in spectroscopic diagnostics of the star. Astronomers use the Latin letters OBAFGKM to classify stars of different spectral types, with O stars having the hottest surface temperatures and M stars the coolest. The Sun is a G star. This classification scheme applies to all stars, not merely to those on the main sequence. To distinguish stars on the main sequence from those in different evolutionary states, astronomers introduced the concept of luminosity class. These categories are designated by Roman numerals from I to V, with I corresponding to supergiants and V to dwarfs. Main-sequence stars are dwarfs because stars have their smallest sizes as luminous objects when they shine by hydrogen fusion in the core, and a small star (dwarf or subgiant) of a given spectral type—*i.e.,* surface temperature—radiates less than a large star (giant or bright giant or supergiant) of the same spectral type. Stars smaller than main-sequence stars are known (white dwarfs, neutron stars, or black holes), but they are very faint and are not normal stars. Consequently, they are not assigned classifications in the normal scheme.

Predominance of main-sequence stars

About 90 percent of the luminous stars in a galaxy at any given time are on the main sequence. Most of the mass of a galaxy is contained in low-mass stars, but the small number of high-mass stars contributes a disproportionate fraction of the total light, especially at blue wavelengths. Most of the light at red wavelengths comes from evolved stars because all stars tend to become redder as they evolve from the main sequence (*i.e.,* as their surfaces expand and cool). In addition, low-mass stars also tend to brighten as they age.

**The end states of stars.** The attempt of stars to achieve mechanical and thermal balance during their luminous lifetime leads inexorably to their demise. The fundamental reason is simple, at least in outline. Because a normal star is composed of ordinary compressible gases, it has to be hot inside to sustain the thermal pressure that resists the inward pull of its self-gravity. On the other hand, interstellar space is dark and cold; radiant heat flows continuously from the star to the universe. The nuclear reserves that offset this steady drain are finite and can only offer temporary respites. When they have run out, the star must die.

White dwarfs, neutron stars, and black holes

Astronomers believe that there are four possible end states for a star: (1) There may occur a violent explosion that completely overcomes self-gravity and disperses all constituent matter to interstellar space; this would leave nothing behind as the stellar remnant. (2) The free electrons in the core of the star may finally become so densely packed that quantum effects allow them to exert enough pressure (termed electron-degeneracy pressure; see below) to support the star even at zero temperature; this would leave behind a white dwarf as the stellar remnant. (3) If the mass of the core exceeds the maximum value—the Chandrasekhar limit of 1.4 solar masses—allowed for a white dwarf, the compression of the stellar matter may finally be stopped at nuclear densities; this would leave behind a neutron star. (4) If the mass of the core is so large that even nuclear forces are incapable of supporting the star against its self-gravity, the gravitational collapse of the star may continue to a highly singular state at the centre; this would leave behind a black hole.

Observations of star clusters suggest that stars initially less massive than about six solar masses are able to lose enough of their envelopes in the final stages of normal stellar evolution that their burnt-out cores fall below the Chandrasekhar limit, resulting in a white dwarf remnant. Theoretical calculations are able to reproduce this result if empirical envelope-mass loss rates are adopted for the later stages of the evolution. In the intermediate range of six to eight solar masses, the situation is more confused. Some calculations indicate that a nuclear detonation occurs when carbon ignites (under degenerate conditions when free electrons are crowded to extremely high densities, allowing little further contraction) to blow the star completely apart and leave nothing behind. Other calculations indicate a relatively undramatic evolution to a white-dwarf end state at the lower masses, while at higher masses carbon fusion can be ignited under nondegenerate conditions, and the further evolutionary stages resemble stars initially more massive than eight solar masses. In the range of eight to 25 solar masses, the star is believed to suffer an iron-core collapse, giving an implosion of the central regions to form a neutron star and an expulsion of the envelope in a supernova explosion. Above 25 solar masses or so, the central regions may collapse into a black hole.

Observations of stellar remnants are reasonably in accord with the above picture. White dwarfs slowly cooling to the same temperature as the universe (3 K) seem to account for most of the dying stars, which is consistent with the fact that most stars are born with relatively low masses. At the sites of some historical supernova explosions, astronomers have found objects called pulsars, which are thought to be rotating, magnetized neutron stars. And in some close binary systems, where a normal star is transferring matter to a compact companion, the companion can be inferred in different situations to be a white dwarf, a neutron star, or a black hole.

**The evolution of stars.** The above account of the structure of stars in the main sequence or end states bypassed two interesting issues: how stars arrive at the main-sequence state, and how stars evolve in detail from the main sequence to their final state. The first issue deals with the problem of star formation and will be covered in a subsequent section; the second primarily has to do with the origin of the chemical elements.

Whenever nuclear fuel runs out in the central regions of a star (*e.g.,* when hydrogen becomes exhausted at the end of the main-sequence stage of stellar evolution), the core must contract and heat up. This increases the flow of energy to the outside, which accelerates evolution. A shell of material outside of the contracting core may become hot enough to trigger thermonuclear fusion, and eventually the central temperature also may rise enough to ignite what was previously nuclear ash into new fuel. The entire process will then repeat. Thus, core fusion of hydrogen into helium can give way to shell hydrogen fusion. This can be followed by helium ignition in the core, with the star now possessing a shell of hydrogen fusing into helium and a core of helium fusing (with itself twice) into carbon. If the temperature rises sufficiently, the carbon also can capture a helium nucleus to become oxygen. Helium exhaustion in the core is followed by helium fusing in a shell and hydrogen fusing in another shell above that. Then, core ignition involving carbon or oxygen fusing with themselves can yield a variety of still heavier elements. The layered shell structure and the chain of possible reactions become more and more complicated, generating along the way such common elements as silicon, sulfur, and calcium, but the process cannot proceed forever. Eventually, if nothing else intervenes, iron will be created. The nucleus of the iron atom is the most bound of all atomic nuclei; it is not possible to release nuclear energy by adding nucleons to iron (or subtracting them). Hence, if iron is created, as in the cores of the more massive stars, the star must come to a catastrophic end because it will continue to lose heat to its surroundings. What happens in computer simulations of this event is that the core of the star implodes, forming a large mass of hot neutrons at temperatures and densities considerably in excess of $10^9$ K and $10^{14}$ g/cm$^3$ (grams per cubic centimetre). Under such conditions huge numbers of neutrinos are released, and these elementary particles appear capable of depositing enough energy into the dense infalling envelope of the star to drive an outwardly propagating shock wave that expels the envelope in a supernova explosion. In this way a wide variety of the nuclear products of stellar evolution can be introduced into the

upernova xplosion nd the issem- ination f heavy lements

interstellar medium to enrich the general elemental mix. From this point of view, it is encouraging that, apart from hydrogen and helium, elements that are bountiful in the natural environment (and in living species)—carbon, nitrogen, oxygen, silicon, sulfur, calcium, iron, etc.—also lie on the main line of stellar nucleosynthesis.

An interesting by-product of the supernova mechanism described above is that large numbers of free neutrons can be liberated in the envelope. Seed nuclei are able to capture these free neutrons to become heavier and eventually create many of the elements beyond iron in the periodic table, including uranium and other radioactive species. Different isotopes of uranium decay at different rates, and knowing the primitive ratios in which supernovas produce these isotopes enables radiochemists to compute, from the corresponding measured values in uranium ore, the elapsed time since these isotopes were created and introduced into the solar system. Depending on the rate of supernova explosions in the history of the Galaxy, these calculations indicate that uranium synthesis began between 6,000,000,000 and 15,000,000,000 years ago. This, then, is another method for independently estimating the age of the Galaxy. Again, within the uncertainties of the determination, the value is consistent with the Hubble expansion age—$10^{10}$ years.

The fundamental difference in evolutionary outcomes between high-mass and low-mass stars can be traced to the theory of white dwarfs. Basically, every star eventually tries to generate a white dwarf at its core as it evolves and undergoes core contraction. During the 1920s, with the dawn of modern quantum mechanics, the British physicist Ralph H. Fowler showed that a white dwarf has the peculiar property that the more massive it is, the smaller is its radius. The reason is relatively simple: a more massive white dwarf has more self-gravity, and so more internal pressure is required to counter the stronger gravity. Pressure increases when the degenerate electron gas constituting a white dwarf is compressed; it becomes strong enough to balance gravitational force only at very great densities. Consequently, equilibrium between the internal degeneracy pressure and the force of gravity is reached at a smaller size for a more massive white dwarf.

The Indian-born U.S. astrophysicist Subrahmanyan Chandrasekhar made a crucial modification to this hypothesis in order to accommodate Einstein's special theory of relativity. Chandrasekhar showed that relativistic effects imposed an upper limit on the mass of possible white dwarfs. This limit arises because electrons cannot move faster than the speed of light; therefore, as the mass of a white dwarf is increased, there comes a point at which the increase in internal degeneracy pressure is no longer able to keep the self-gravity from literally trying to crush the star to zero size. For likely white-dwarf compositions, this limit corresponds to 1.4 solar masses as noted above.

'he Chan- rasekhar mit

Consider a star that attempts to exceed the Chandrasekhar limit, assuming that it has enough material—even after envelope-mass loss—to try to build a massive white dwarf by depositing layer after layer of nuclear ash into its core. As the limit is approached, the core's outer boundary shrinks almost to arbitrarily small dimensions, generating above it enormous gravitational fields. To counteract the gravity, the pressures in the shell above the core must rise correspondingly, yielding densities and temperatures that are as high as needed to drive all thermonuclear reactions to completion. If nothing else intervenes, this situation must end in the iron catastrophe described above.

In contrast, in a low-mass star the final mass of the core may end up well below the Chandrasekhar limit. The shells outside the core may still become dense and hot enough to yield copious amounts of hydrogen and helium fusion, and this heat input into the envelope will greatly distend the envelope of the star, bringing the star to the red giant and red supergiant evolutionary phases that characterize the later stages. The outer atmospheres of such stars are often cool enough to allow the condensation of some of the heavy elements into solid particles. Dust grains of a rocky silicate are probably the most common outcome, but grains of graphite or silicon carbide are possibilities in carbon-rich stars. In any case, because the envelope of the star is so extended, the surface gravity is too weak to hold the atmospheric mix of gas and dust, and this mixture blows out of the star as a prodigious stellar wind. The observed loss of matter occurs at a rate rapid enough to strip off the entire envelope, revealing eventually a white-hot core that is now a bare white dwarf. Since the mass loss reduced the stellar mass below the Chandrasekhar limit, the core never progressed to very advanced stages of nuclear fusion, giving the most common white dwarfs in the Galaxy a likely composition of carbon and oxygen.

ted giant nd red upergiant hases

**Interstellar clouds.** Observations conducted at radio, infrared, and optical wavelengths show that the majority of stars are formed from giant clouds of gas and dust that exist in interstellar space. There are three basic varieties of clouds that astronomers distinguish on the basis of the dominant physical state in which the hydrogen gas is found: atomic, molecular, or ionized. Hydrogen is singled out in the classification scheme because of its preeminent abundance in the Cosmos.

Atomic, molecular, and ionized hydrogen clouds

Atomic hydrogen clouds are the most widely distributed in interstellar space and, together with molecular hydrogen clouds, contain most of the gaseous and particulate matter of interstellar space. Molecular hydrogen clouds contain a wide range of molecules besides the hydrogen molecule $H_2$, and for that reason are simply called molecular clouds. Ionized hydrogen clouds, called H II regions by astronomers, are fluorescent masses of gas, such as the famous Orion Nebula, which have been lit up by hot blue stars recently born from the neutral gas, the hydrogen becoming dissociated and ionized because of the outpouring of ultraviolet photons from such massive stars.

Dust particles are suspended in all three types of clouds, and their effects can be seen in the absorption and scattering of optical light or in the thermal emission of infrared radiation. The refractory cores of the dust grains were probably expelled from the atmospheres of countless red giant stars, although icy mantles may be acquired in molecular clouds by the adhesion of molecules to the cold grain surfaces when they collide. It has been estimated that dust grains typically account for 1 percent of the mass of an interstellar cloud. Because the internal constitution of dust is primarily elements heavier than hydrogen and helium, and because the cosmic mass fraction of all such elements is only a few percent of the total, dust grains must contain a significant fraction of the total cosmic abundance of heavy elements. This deduction is in accord with the observational finding that many heavy elements are severely underrepresented in the gas phase of interstellar clouds. They presumably have condensed out as solid particles.

Of greatest interest to the present discussion are the molecular clouds, because it is from giant complexes of such clouds that most stars are formed. Radiative cooling by the molecules and dust in them keeps the matter at very low average temperatures, about 10 K, and at relatively high densities as compared with atomic hydrogen clouds. These two circumstances, combined with the large mass ($10^5$ or $10^6$ solar masses) of a typical giant molecular cloud complex, make molecular clouds ideal sites for star formation because, even with dimensions spanning hundreds of light-years, they are held together by their self-gravitation. Once a gaseous astronomical body becomes self-gravitating, the formation of still more condensed states—in this case, stars—is almost inevitable.

Primary sites of star formation

**Star formation.** Detailed radio maps of nearby molecular clouds reveal that they are clumpy, with regions containing a wide range of densities—from a few tens of molecules (mostly hydrogen) per cubic centimetre to more than 1,000,000. Stars form only from the densest regions, called cloud cores, though they need not lie at the geometric centre of the cloud. Large cores (which probably contain subcondensations) up to a few light-years in size seem to give rise to unbound associations of very massive stars (called OB associations after the spectral type of their most prominent members, O and B stars) or to bound clusters of less massive stars. Whether a stellar group materializes as an association or a cluster seems to depend on the efficiency of star formation. If only a small fraction of the matter goes into making stars with the rest being

blown away in winds or expanding H II regions, then the remaining stars end up in a gravitationally unbound association, dispersed in a single crossing time (diameter divided by velocity) by the random motions of the formed stars. On the other hand, if 30 percent or more of the mass of the cloud core goes into making stars, then the formed stars remain bound to one another, and the ejection of stars by random gravitational encounters between cluster members takes many crossing times.

Low-mass stars also are formed in associations called T associations after the prototypical stars found in such groups, T Tauri stars. The stars of a T association form from loose aggregates of small molecular cloud cores, each a few tenths of a light-year in size, that are randomly distributed through a larger region of lower average density. The formation of stars in associations is the most common outcome; bound clusters account for only about 10 percent of all stars born. The overall efficiency of star formation in associations is quite small. Typically less than 1 percent of the mass of a molecular cloud becomes stars in one crossing time of the molecular cloud (about 5,000,000 years). Low efficiency of star formation presumably explains why any interstellar gas remains in the Galaxy after $10^{10}$ years of evolution. Star formation at the present time must be a mere trickle of the torrent that occurred when the Galaxy was young.

A typical cloud core rotates fairly slowly, and its distribution of mass is strongly concentrated toward the centre. The slow rotation rate is probably attributable to the braking action of magnetic fields that thread through the core and its envelope. This magnetic braking forces the core to rotate at nearly the same angular speed as the envelope as long as the core does not go into dynamical collapse. Such braking is an important process because it assures a source of matter of relatively low angular momentum (by the standards of the interstellar medium) for the formation of stars and planetary systems. It also has been proposed that magnetic fields play an important role in the very separation of the cores from their envelopes. The proposal involves the slippage of the neutral component of a lightly ionized gas under the action of the self-gravity of the matter past the charged particles suspended in a background magnetic field. This slow slippage would provide the theoretical explanation for the observed low overall efficiency of star formation in molecular clouds.

The effects of magnetic fields

At some point in the evolution of a molecular cloud, one or more of its cores become unstable and subject to gravitational collapse. Good arguments exist that the central regions should collapse first, producing a condensed protostar whose contraction is halted by the large buildup of thermal pressure when radiation can no longer escape from the interior to keep the (now opaque) body relatively cool. The protostar, which initially has a mass not much larger than Jupiter, continues to grow by accretion as more and more overlying material falls on top of it. The infall shock, at the surfaces of the protostar and the swirling nebular disk surrounding it, arrests the inflow, creating an intense radiation field that tries to work its way out of the infalling envelope of gas and dust. The photons of optical wavelengths are degraded into longer wavelengths by dust absorption and reemission, so that the protostar is apparent to a distant observer only as an infrared object. Provided that proper account is taken of the effects of rotation, this theoretical picture correlates with the radiative spectra emitted by many candidate protostars discovered near the centres of molecular cloud cores.

An interesting speculation concerning the mechanism that ends the infall phase begins by noting that the inflow process cannot run to completion. Since molecular clouds as a whole contain much more mass than what goes into each generation of stars, the depletion of the available raw material is not what stops the accretion flow. A rather different picture is revealed by observations at radio, optical, and X-ray wavelengths. All newly born stars are highly active, blowing powerful winds that clear the surrounding regions of the infalling gas and dust. It is apparently this wind that reverses the accretion flow.

The geometric form taken by the outflow is intriguing. Jets of matter seem to squirt in opposite directions along the rotational poles of the star (or disk) that sweep up the ambient matter in two lobes of outwardly moving molecular gas—the so-called bipolar flows. The exact mechanism that produces these collimated flows remains somewhat obscure, but there seems little doubt that it forms an intrinsic part of the star-formation process. Such jets and bipolar flows are doubly interesting because their counterparts were discovered some time earlier on a fantastically larger scale in the double-lobed forms of extragalactic radio sources (see below *Quasars and related objects*).

Bipolar flows

The underlying energy source that drives the outflow is unknown. Promising mechanisms invoke tapping the rotational energy stored in either the newly formed star or the inner parts of its nebular disk. In a low-mass protostar, a scaled-up version of the mechanism that drives the solar wind may work. Because newly born stars of low mass have strong convective motions throughout their interior, dynamo-generation of strong magnetic and coronal activity seems likely. This suggestion is consistent with the observational evidence that frequent flaring and strong chromospheres are present in the T Tauri stars, which are known to be pre-main-sequence stars of low mass.

An incompletely investigated area concerns the ultimate fate of the nebular disks that are found, both observationally and theoretically, around forming stars. It appears likely that companions would form out of such disks, but it remains uncertain under what conditions the companions would be stars and under what conditions they would be planets.

Knowledge of the evolution of the central stars after they disperse their circumstellar placental material is much securer. Pre-main-sequence stars of low mass first appear as visible objects, T Tauri stars, with sizes that are several times their ultimate main-sequence sizes. They subsequently contract on a time scale of tens of millions of years, the main source of radiant energy in this phase being the release of gravitational energy. When their central temperatures reach values comparable to $10^7$ K, hydrogen fusion ignites in their cores, and they settle down to long stable lives on the main sequence. The early evolution of high-mass stars is similar; the only difference is that their faster overall evolution may allow them to reach the main sequence while they are still enshrouded in the cocoon of gas and dust from which they emerged.

## GALAXIES

Astronomers have found that most of the matter in the universe is concentrated in galaxies. Paradoxically, they also have discovered from studying galaxies that the universe may contain large quantities of mass in a form not observable in galactic systems. There are some hints that this hidden mass may not even be in the form of ordinary matter, and the discrepancy between the mass that can be seen in galaxies and the mass needed to account for their gravitational binding has become one of the foremost unsolved problems in modern astrophysics.

**The Milky Way Galaxy.** Any discussion of galaxies should begin with the local system, where the wealth of information is greatest. The Galaxy contains three main structural components: (1) a thin flat disk of stars, gas, and dust; (2) a spheroidal central bulge containing only stars; and (3) a quasi-spherical halo of old stars. The Sun is found in the first component, while globular clusters are found in the third. The nucleus of the Galaxy lies at the centre of all three components, but it cannot be seen optically from the solar system because of the thick tracts of dust that lie in the disk between it and the galactic centre, obscuring the view. The nucleus can be probed at radio, infrared, X-ray, and gamma-ray wavelengths; a description of these findings is provided below in a more general discussion of the activity witnessed in galactic nuclei.

A hint of the formation and evolution processes of the Galaxy is contained in the general correlation between the spatial location of a star in the galactic system and its heavy-element abundance. Most of the stars found in the disk of the Galaxy are Population I stars; those in the halo are of the Population II type; and those in the bulge are a mixture of the two. This correlation was first noticed in the 1940s by the German-born U.S. astronomer

Distribution of stellar populations in the Galaxy

Walter Baade from his investigation of the Andromeda galaxy. Since the theory of nucleosynthesis states that the abundance of heavy elements in successive generations of stars should increase with age, it can be deduced that star formation terminated in the halo long ago while having continued to the present day in the disk.

The shapes acquired by the different stellar components can be understood in terms of the orbital characteristics of the different stellar populations. For Population I stars, the motion corresponds nearly to circular orbits in a single plane; the random velocities above the circular component are small, accounting for the flattened shape of the galactic disk. For Population II stars, the noncircular velocities are much larger; the stars orbit randomly about the Galaxy like a swarm of bees around a hive, accounting for the spheroidal shapes of the galactic bulge and halo.

Although the velocities of the stars within a few thousand light-years of the Sun in the direction perpendicular to the galactic plane are generally small, they are not zero. By investigating the statistics of these motions and the vertical structure of the disk, it is possible to deduce the vertical component of the gravitational field of the Galaxy and thereby the total mass of material required locally to supply the observed gravity. The quantity of required material is called Oort's limit (after the aforementioned J.H. Oort), and it exceeds by a factor of about two the quantity

Oort's limit

of available material, as observed in the form of known stars and gas clouds. This result constitutes the closest example of a general discrepancy arising on galactic scales whenever dynamically derived masses are compared with direct counts of observationally accessible objects. The missing matter in Oort's limit, however, pertains to a disk population and may differ in ultimate resolution from the more general hidden-mass problem (see below), which is associated with the halos of galaxies and beyond.

From star counts, one can derive another quantity of astronomical interest, the mean brightness (per unit area) in the solar neighbourhood. If one divides this quantity into the mass (per unit area) corresponding to Oort's limit, one obtains the local mass-to-light ratio, which has been determined to be about five in solar units. In other words, the gravitating mass in the Galaxy has a mean efficiency for producing light that is five times less than the Sun's. This implies, first, that the average star must be less massive than the Sun, and, second, that the amount of helium—in contrast with the heavier elements—presently inside stars cannot have been produced by stellar processes. The reason is simple. The Sun with a mass-to-light ratio of one will manage to convert about 10 percent of its mass (in the core) into helium in $10^{10}$ years (after which it leaves the main sequence); matter with a mean mass-to-light ratio of five, therefore, would convert only 2 percent of its mass to helium in $10^{10}$ years, roughly the age both of the Galaxy and the universe. The cosmic abundance of helium, however, is approximately 26 percent of the total mass; thus, unless the Galaxy was much brighter in the past than it is today (for which there is no observational evidence), the bulk of the helium in the universe must have been created by nonstellar processes. Astronomers now believe that a primordial abundance of helium of about 24 percent by mass emerged from the big bang. Among other arguments, this is the value derived from the analyses of the chemical compositions of H II regions in external galaxies where the heavy-element abundance is very low and where nuclear processing by stars has, therefore, presumably been small.

It is possible, of course, to examine the statistics of the random velocities of stars in the two directions parallel to the galactic plane as well as in the vertical direction. The Swedish astronomer Bertil Lindblad was the first to carry out such an analysis. His work, combined with Oort's study in 1927 of the constants of the differential rotation of the Galaxy, gave the period of revolution of stars such as the Sun about the galactic centre. Modern values for this period lie between 200,000,000 and 250,000,000 years. With Shapley's measurement of the distance to the centre of the Galaxy and with the assumption that such stars as the Sun circle the system because they are gravitationally bound to it, it is possible to estimate the total mass interior to the solar distance from the galactic centre. Modern estimates yield roughly $10^{11}$ solar masses. Because the Sun is somewhat more massive than a typical star, the Galaxy must contain more than $10^{11}$ stars.

Detailed information can be gleaned about the distribution of mass in the Galaxy if one possesses a knowledge of the rotational speeds of disk matter at other radial locations in the Galaxy. The most common measurements are of atomic hydrogen in its spin-flip transition at 21-centimetre wavelength and of the carbon monoxide molecule in one or another of its rotational transitions at millimetre wavelengths. These observations also provide data concerning the total amount of atomic and molecular hydrogen gas contained in the Galaxy. To convert the carbon monoxide abundance to a molecular hydrogen abundance (which cannot be measured directly except at ultraviolet wavelengths that suffer tremendous dust extinction) requires a complicated series of calibrations of nearby sources. The mass of gas in the Galaxy is a few thousand million solar masses nearly evenly divided between atomic and molecular hydrogen clouds. Most of the observed mass of the Galaxy thus exists in the form of stars; gas and dust make up only a few percent of the total.

By a combination of such measurements, astronomers can obtain the rotation curve of the Galaxy from its innermost regions to a radial distance of almost 60,000 light-years from the galactic centre. This rotation curve implies that the mass of the Galaxy measured out to a certain distance $r$ does not converge to a fixed value as $r$ increases but continues to rise roughly in linear proportion to $r$. Observations indicate, however, that the integrated light from a galaxy such as the Milky Way system does not increase with increasing $r$ but approaches asymptotically a finite value. Thus, the local mass-to-light ratio of the Galaxy, like those of other spiral galaxies, must increase dramatically toward its outer parts where the halo dominates. Another way to state the problem is that the observed rotational velocities of gas clouds in the outer parts of spiral galaxies are so large that they would not be bound to the galaxies unless the galaxies were more massive than inferred from direct measurements of their stellar and gas contents. Most astronomers now accept the likelihood of dark halos that contain as much mass as is present in the visible disks and bulges. More controversial are the claims that these halos may increase known galactic masses by factors of 10 or 100.

Existence of dark halos

**Classification of galaxies.** Astronomers judge galaxies in accordance with three criteria: morphological appearance, stellar content, and overall luminosity. Although the number of galaxies found in the universe is enormous, Edwin P. Hubble discovered that a few basic categories specify their observed shapes. Galaxies that have irregular shapes are called irregulars, denoted "Irr." Irregulars are subdivided into two categories: Irr I and Irr II. Irr I galaxies have OB stars and H II regions; examples of such systems are the Large and Small Magellanic Clouds. Irr II galaxies are amorphous in texture and show no resolution into bright stars or associations. Galaxies that have regular forms are divided into two broad groups: ellipticals and disks. Elliptical galaxies, denoted "E," have roundish shapes. Disk galaxies, on the other hand, have flattened shapes. They can be further divided into two subcategories: ordinary spirals, denoted "S," and barred spirals, denoted "SB." In addition, there exists a transition type between ellipticals and spirals, which are often called lenticulars. The lenticular galaxies are designated either "S0" or "SB0," depending on the absence or presence of a bar of stars, gas, and dust through the nucleus.

The irregulars constitute only a few percent of all galaxies; the two kinds of spirals, about 70 percent; and the ellipticals and lenticulars, the rest. These fractions refer to the brighter systems that contain most of the known matter of the universe. In terms of the total number of galaxies, bright and faint, dwarf ellipticals appear to be the most common form. An interesting unresolved question concerning such systems is whether they have the same massive dark halos that seem to reside in the giant galaxies.

Ellipticals and spirals constitute the two largest reservoirs of the stars in the universe, and the placement of indi-

vidual galaxies into these two major categories is refined by adding a numeral 1 through 7, or a letter "a" through "c" to their designation. The sequence E0 to E7 denotes increasing flattening (as seen in projection in the sky). The sequence Sa to Sc or SBa to SBc represents decreasing tightness of winding of the spiral arms and decreasing size of the central bulge relative to the disk.

Several modifications of Hubble's classification scheme for spiral galaxies have been proposed by Gerard de Vaucouleurs: (1) the more symmetric notations SA and SB for ordinary and barred spirals; (2) the continuation of the sequence a, b, c, to d and m (for Magellanic); and (3) the appendage of the symbols (r), (s), or (rs) to indicate the relative structural prominence of rings versus spiral arms.

Another useful extension of Hubble's purely morphological scheme is the inclusion by William W. Morgan, Nicholas U. Mayall, and Donald E. Osterbrock of the dominant stellar makeup in a spiral galaxy. Starlight from the central bulge can be expected to have a later composite spectral type in the sequence OBAFGKM compared to starlight from the disk and spiral arms. The relative contribution of light from the bulge to that from the disk (and spiral arm) can therefore be labeled (in increasing importance of bulge contribution) by a concentration class: a, af, f, fg, g, gk, and k.

Luminosity class

A further useful analogy with stars is the introduction by Sydney van den Bergh of the concept of luminosity class. The scheme appends to the Hubble type a luminosity-class label, from Roman numeral I for the intrinsically brightest (and most massive) spiral galaxies to Roman numeral V for the intrinsically faintest (and least massive) spirals. The utility of this scheme, as applied to spirals, rests with the fact that it is possible to assign such galaxies a luminosity class without actually measuring their distance (to obtain an absolute brightness from an observed apparent brightness). The luminosity class of a spiral correlates well with the regularity (or "prettiness") of the spiral structure: in class I galaxies, the arms are long and well developed and have a high surface brightness; in class III, they are patchy and fuzzy; and in class V, there may be barely a hint of a spiral structure. Elliptical galaxies, lacking spiral arms, cannot have their absolute brightnesses estimated by the same morphological considerations; hence, the concept of a luminosity class for them is less empirically useful. When the masses of elliptical galaxies at known distances are deduced from measured velocities or apparent luminosities, they range from a few million solar masses (dwarf ellipticals) to more than $10^{12}$ solar masses (giant ellipticals). Thus, giant ellipticals and giant spirals have comparable masses. Yet, it should be noted that the very largest elliptical galaxies in the universe, the supergiant cD systems, are unique and perhaps have masses approaching $10^{14}$ solar masses in some extreme cases.

**Dynamics of ellipticals and spirals.** The motions of stars in an external galaxy can be studied in a statistical sense by examining the Doppler shifts of the optical absorption lines in the integrated light along the line of sight through different parts of the object. Radio spectroscopic observations can yield similar information concerning the gaseous components of the system. Some important results from these studies are as follows.

The dominant motion in the disks of normal spiral galaxies is differential galactic rotation, with the random motions of stars being relatively small and that of the atomic and molecular gas smaller still. A surprising result, the first hints of which came in the 1960s but were not fully confirmed until the 1970s, is that the rotation curves of almost all well-studied spiral galaxies become flat at large radial distances. As one goes out from the centre, the rotational velocity rises to a constant value $V$ and then maintains it for as far as one can make the measurements. This implies, as already noted for the Milky Way Galaxy, that the mass contained within $r$ increases linearly with increasing $r$, and provides the firmest piece of evidence in support of the hypothesis that large amounts of hidden mass may be present in the halos of spiral galaxies.

The qualitative fact of disk galaxies rotating differentially, with the inner parts having shorter rotational periods than the outer parts, has been known since Lindblad's and Oort's investigations of the problem for the Milky Way system in the 1920s. This fact, combined with age estimates for all galaxies of about $10^{10}$ years, presents a dilemma for the origin of spiral structure. If spiral arms are viewed as consisting always of the same material (*e.g.*, the same gas clouds that give birth to the brilliant OB stars and H II regions that best define the optical spiral structure), then the arms should wind up. In particular, with a flat rotation curve, material at half the solar distance from the centre of the Galaxy should go around twice for each revolution of the material at the solar distance, and an extra turn should then be added to each spiral arm between these two radii every 200,000,000 to 250,000,000 years. This would give the spiral arms of the Galaxy (and other spiral galaxies like it) several dozens of turns over the lifetime of the system, whereas spiral galaxies have in fact never been observed with more than one or two turns.

A way out of the winding dilemma is the proposal that spiral structure is a wave phenomenon, the spiral arms being a local "piling-up" of stars and gas clouds that individually flow through the spiral pattern, much as a traffic jam is a local piling-up of cars and trucks that individually flow through the jam. The piling-up arises because the self-gravity of the excess matter in the arms causes deflections of what would otherwise be circular orbits (on average), the deflections producing self-consistently the original pileup.

The concept of density waves

Most astronomers are agreed that density waves underlie the phenomenon of spiral structure in the so-called grand design galaxies. More controversial is whether some other mechanism (*e.g.*, "stochastic star formation") might play a role in galaxies where the spiral structure is "flocculent."

In modern density-wave theory, as developed by the Chinese-born U.S. mathematician Chia-chiao Lin and his associates, spiral structure represents an unstable mode of collective oscillation. The instability provides a way by which a differentially rotating disk galaxy may release free energy of differential rotation and spontaneously generate spiral waves. The balance of the growth of these waves against their dissipation (through the response of the interstellar gas clouds) may yield a quasi-stationary state whereby gaseous matter slowly drifts to the interior and angular momentum is steadily transported to the exterior. Although the details of the entire picture remain incomplete, many of the basic predictions—as, for example, that the perturbations in density and velocity should be strongest in the component with the smallest random velocities; *i.e.*, gas and dust clouds—have already been confirmed both qualitatively and quantitatively in several well-observed spiral galaxies. Furthermore, the Hubble correlation between the tightness of spiral windings and the size of the central bulge relative to the disk, as well as the van den Bergh correlation between luminosity class and the degree of organization of the spiral structure, are simple direct consequences of density-wave theory.

A similar explanation probably underlies the barred spiral galaxies, with the basic underlying disturbance being an oval distortion. The predicted departures from circular motions are larger in barred spirals than in ordinary spirals, and this seems to be consistent with the observational evidence that currently exists for this problem. The enhanced rates at which matter is brought to the centres of such galaxies may have implications for various energetic events that take place in some galactic nuclei.

In elliptical galaxies, the constituent stars have random velocities that are generally much larger than the rotational motions. This explains why ellipticals possess neither thin disks nor spiral arms. Moreover, giant ellipticals are flatter than would be inferred from the amount of rotation that they do possess, and increasing rotation does not necessarily lead to increasing flattening, as appears to happen, for example, to ellipticals of lower luminosity and the bulges of spiral galaxies. Also, most ellipticals do not appear to have young stars, probably because there is little gas and dust from which stars can form.

Mathematical analysis and computer simulations made since the early 1970s have suggested a possible stellar-dynamical basis for understanding the basic shapes of giant elliptical galaxies. Unlike the bulges and disks of S0 galaxies, the bodies of giant ellipticals may not be figures

of revolution (*e.g.*, oblate spheroids), but may possess three axes of unequal lengths. In the models, the triaxial shape arises because the random velocities of the stars are anisotropic (not equal in all directions). Such a state of affairs seems consistent with the existing observational data. If confirmed, it could bear importantly on the issue of the origin of giant ellipticals.

An interesting empirical property shared by both ellipticals and spirals is that their luminosities $L$ seem to be proportional to the fourth power of their random or circular velocities $V$. The proportionality constant can be calibrated with the help of nearby (giant) galaxies, and the resulting relation may then be used for cosmological investigations. In particular, the determination of distances is a recurring astronomical problem, and the relation, $L$ proportional to $V^4$, provides a method for obtaining distances. In brief, a measurement of $V$ allows the determination of $L$, which, combined with the observed apparent brightness, gives the distance of the object.

**Interacting galaxies.** Strongly interacting pairs of galaxies make up less than 1 percent of all galaxies, but the more spectacular examples produce intriguing structures (bridges, tails, rings, and shells) and involve processes (stripping, merging, and sinking) that are not present in individual isolated galaxies. Computer simulations of the gravitational encounter between a large disk galaxy and a small one show that the latter can pull material from the near side of the former into a bridge that temporarily spans the gulf between the two. Encounters between two more nearly equal participants can yield one long tail from each disk galaxy, which extends away from the main bodies (Figure 1). Rings emerge in the disk of a galaxy if another massive galaxy passes through its body; the brief inward pull and subsequent rebound cause the orbits of the rings to pile together like ripples on the surface of a pond into which a stone is dropped. Shells form across the face of a large elliptical galaxy if it devours a small companion; the stars of the smaller system are strewn like wine out of a rolling barrel with the stopper removed.

Gravitational encounters between pairs of galaxies

(Top) H.C. Arp, Max-Planck-Institut fur Physik und Astrophysik, Munich, W.Ger., (bottom) Alar Toomre, Massachusetts Institute of Technology

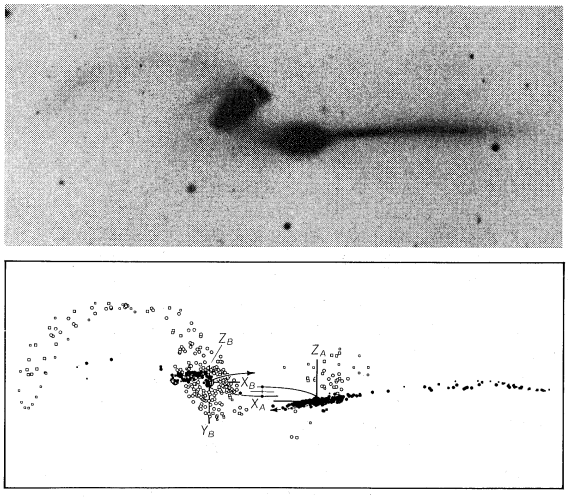



Figure 1: *The interacting galaxies NGC 4676A and B.* (Top) A negative print of these two nearly equal-sized disk galaxies shows that the close encounter between the pair has resulted in the formation of a long "tail" from each. (Bottom) A computer simulation of the same strongly interacting galaxies reveals their structures more clearly.

Stripping (of matter), merging (of the main bodies of the galaxies), and sinking (of the satellite galaxy toward the centre of the parent body) are all represented in the above example, and these processes, individually and collectively, have been invoked increasingly by theorists in a wide variety of contexts and by a wide variety of names to explain different observed galactic phenomena. The most interesting and speculative application is perhaps to the origin of elliptical galaxies.

It has been proposed by Alar Toomre of the United States that elliptical galaxies result from the merger of spiral galaxies, jumbled piles of stars from the wreckage of collisions of bound pairs of galaxies with arbitrarily oriented spins and orbits. A conservative variant of this theory suggests that only supergiant ellipticals, the cD galaxies that tend to lie at or near the centres (or density maxima) of rich clusters of galaxies, form in this manner (by "cannibalism" of many smaller neighbours).

**Galaxy formation.** Unlike star formation, which is an ongoing physical process, the existing observational evidence is consistent with the view that all galaxies formed about $10^{10}$ years ago. Thus, the problem of galaxy formation is more akin to the problem of the origin of the solar system (without the benefit of space exploration) than to star formation: astronomers have only clues from the already finished product with which to work in trying to deduce the sequence of events. The difficulty is compounded by two uncertainties: (1) the recognition that star formation is an integral aspect of galaxy formation and that the process of star formation may occur very differently under primordial conditions than under present ones where dust, molecules, and magnetic fields play primary roles; and (2) the realization that a major fraction of the mass of a galaxy may exist in a form whose properties are completely unknown. It should not be surprising therefore that knowledge in this field is in a comparatively primitive state and is subject to continual revision. The circumstance that allows even a broadbrush approach is the almost universal faith that a secure backdrop does exist against which theoretical strokes of paintwork can be attempted. This backdrop is of course the big bang scenario for the origin of the universe.

In the big bang picture, the matter in the universe during the epoch before galaxy formation was composed almost exclusively of an intense radiation field and a hydrogen and helium plasma, with everything having nearly the same density and temperature. As the universe expanded, the density and temperature decreased, and the helium and hydrogen ions recombined with free electrons, after which the matter and the radiation became decoupled—*i.e.*, the cosmic material became transparent to the propagation of photons. The significance of decoupling is that the matter of the universe at this point would be free to aggregate under the influence of its self-gravity (and that of the hidden mass) without resistance from radiation drag.

Decoupling of matter and radiation

If there existed at this juncture a spectrum of density fluctuations about the mean state, then different regions of enhanced density would have begun to slow the expansion and eventually contract (provided they were massive enough for self-gravity to overwhelm the pressure forces that would try to maintain the gas in an extended state). The character of the fluctuations that could be expected to be present and their subsequent evolution depend on a variety of complex issues: the nature of the hidden mass, the ability of the early universe to generate "adiabatic" versus "isothermal" fluctuations, the erasure of fluctuations of too small a scale by frictional coupling of matter and radiation in the era before decoupling, and the ability of the gas to cool in the time it takes the gas to cross the dimensions of the body.

Taking all of these factors into account, astronomers deduce that the most likely epoch for bodies of galactic dimensions to achieve separate identities from the rest of the Cosmos occurred when the universe expanded by a factor of about 100 beyond decoupling. The age of the universe at this time is estimated at a few hundred million years, and this is also the time matter takes to cross a typical galaxy at free-fall speeds. In the process of so contracting (perhaps in many smaller gaseous pieces, each containing roughly $10^6$ solar masses), neighbouring protogalaxies would exert gravitational torques on each other, imparting amounts of angular momenta comparable to that possessed by galaxies today. The bodies would thus flatten in the subsequent collapse.

Material that reached a completely flattened state while still in a gaseous state would have its vertical component of motion arrested in a strong shock wave and form the disk of a galaxy. Material that formed dense stars or protostars on the way down would be able to pass through the disk virtually unimpeded and, after several bounces,

settle to form the bulge and halo of the galaxy. This description forms a plausible scenario for the formation of a disk galaxy such as the Milky Way system. It was once thought that a modification would also produce elliptical galaxies—namely, if the efficiency of star formation were so high during the collapse phase that virtually all of the matter turned into stars before flattening into a disk, then a single quasi-spherical stellar component might result. Given the developments of the 1970s and early 1980s, however, there are two difficulties with this scenario for the formation of elliptical galaxies. First, why should an anisotropic distribution of random velocities, as seems necessary to account for the flattened shapes of giant ellipticals, arise naturally in such a picture? Second, why should the efficiency of primordial star formation be so high when it is so low in the present epoch even in dense molecular clouds that have access to cooling mechanisms by dust and molecules not available to primordial clouds?

The second criticism may not be as damaging as it once appeared because the low efficiency of present-day star formation may well owe its origin to (dynamo-generated) magnetic fields that may not have been present in young galaxies. The first criticism is more difficult to dismiss because whatever mechanism is invoked to explain the anisotropy of velocity dispersions produced in the collapse of elliptical galaxies would have to deal with the observation indicating that rotation, rather than anisotropy, is involved in producing the flattening of the less luminous ellipticals and the bulges of spiral galaxies. This argument may provide support for the view that (giant) elliptical galaxies are merger products rather than the direct results of galaxy formation, but many astronomers are reluctant to accept this conclusion because then a host of questions is raised concerning how random mergers can manage to produce the systematics known for these galaxies.

Search for a primeval galaxy

The observational discovery of a genuine primeval galaxy would remove many uncertainties. In a collapse environment involving only hydrogen and helium gas, the primary diagnostic would be the copious emission of Lyman-alpha radiation (corresponding to the transition between the first excited state and the ground state of atomic hydrogen). The rest wavelength of this transition lies in the ultraviolet, but in primeval galaxies the cosmological redshift would make the observed wavelength longer (*i.e.,* toward the red end of the spectrum). From this point of view, it is interesting that searches near known quasars have uncovered Lyman-alpha emitting galaxies with redshifts exceeding three (see below *Quasars and related objects*).

There also exists a body of opinion that the stars of a primeval galaxy generate so much dust at such a rapid rate that all intrinsic Lyman-alpha production by the galaxy will be degraded to thermal infrared-continuum radiation. In this case, primeval galaxies may resemble a class of infrared-bright galaxies that were discovered by the Infrared Astronomy Satellite (IRAS) in 1984. In contrast to normal galaxies such as the Milky Way system, where the ratio of infrared to visible luminosities is about unity, these sources can emit up to 100 times more infrared radiation than visible light. The only viable explanation for the infrared excess is that these galaxies are somehow undergoing an enormous burst of star formation that has enhanced the usual rates by as much as 100. Ground-based follow-up observations of the IRAS discovery showed that the activity in "starburst galaxies" is generally confined to the central portions of such a system and that many of the candidate sources correspond either to interacting galaxies or to barred spirals. This suggests that starbursts may be triggered by the gravitational perturbations that have brought large amounts of molecular gas to the central regions of the galaxy. A similar burst of star formation might be expected to occur in an era when the matter of a galaxy was nearly all gas rather than all stars.

## QUASARS AND RELATED OBJECTS

Galaxies are where astronomers find stars, the major transformers of matter into energy in the universe. Paradoxically, it is also from the study of galaxies that astronomers first learned that there exist in the universe sources of energy that are individually much more powerful than stars. These sources are radio galaxies and quasars, and their discovery in the 1950s and 1960s led to the establishment of a new branch of astronomy, high-energy astrophysics.

**Extragalactic radio sources.** Sources that emit a continuum of radio wavelengths and that lie beyond the confines of the Galaxy were divided in the 1950s into two classes on the basis of whether they present spatially extended or essentially "starlike" images. Radio galaxies belong to the former class, and quasars (short for "quasi-stellar radio sources") to the latter. The distinction is somewhat arbitrary because the ability to distinguish spatial features in cosmic radio sources has improved steadily and dramatically over the years due to Sir Martin Ryle's introduction of arrays of telescopes, which use aperture synthesis techniques to enhance the angular resolution of a single telescope. Apart from the smaller angular extent that arises from being at a greater distance, many objects originally classified as quasars are now known to have radio structures that make them indistinguishable from radio galaxies. Not every quasar, however, is a radio galaxy. For every radio-loud quasar, there exist 20 objects having the same optical appearance but not the radio emission. These radio-quiet objects are called QSOs for quasi-stellar objects. Henceforth, the term quasars will be used to refer to both quasars and QSOs when the issue of radio emission is not under discussion.

Double-lobed radio sources

The most powerful extragalactic sources of radio waves are double-lobed sources (or "dumbbells") in which two large regions of radio emission are situated in a line on diametrically opposite sides of an optical galaxy. The parent galaxy is usually a giant elliptical, sometimes with evidence of recent interaction. The classic example is Cygnus A, the strongest radio source in the direction of the constellation Cygnus. Cygnus A was once thought to be two galaxies of comparable size in collision, but more recent ideas suggest that it is a giant elliptical whose body is bifurcated by a dust lane from a spiral galaxy that it recently swallowed. The collisional hypothesis in its original form was abandoned because of the enormous energies found to be needed to explain the radio emission.

The radio waves coming from double-lobed sources are undoubtedly synchrotron radiation, produced when relativistic electrons (those traveling at nearly the speed of light) emit a quasi-continuous spectrum as they gyrate wildly in magnetic fields. The typical spectrum of the observed radio waves decreases as a power of increasing frequency, which is conventionally interpreted, by analogy with the situation known to hold for the Galaxy in terms of radiation by cosmic-ray electrons, with a decreasing power-law distribution of energies. The radio waves typically also show high degrees of linear polarization, another characteristic of synchrotron radiation in well-ordered magnetic fields.

A given amount of received synchrotron radiation can be explained in principle by a variety of assumed conditions. For example, a high energy content in particles (relativistic electrons) combined with a low content in magnetic fields will give the same radio luminosity as a low energy content in particles combined with a high content in magnetic fields. A minimum value for the sum results if the energy contents of particles and fields are assumed to be comparable. The minimum total energy computed in this way for Cygnus A (whose distance could be estimated from the optical properties of the parent galaxy) proved to be between $10^{60}$ and $10^{61}$ ergs.

Even a supernova explosion releases only about $10^{51}$ ergs, and so it is clear that $10^9$ supernovas would not suffice to explain the energetics of radio galaxies. Indeed, since at best only 1 percent of the total energy of a supernova emerges in the form of fast particles and since only massive stars explode, even the accumulated explosive power of a whole galaxy of stars would not explain Cygnus A. The most powerful of stellar phenomena pales in comparison with the most powerful of galactic phenomena.

Underlying source of power

A clue to the nature of the underlying source of power came from aperture synthesis studies of the fine structure of double-lobed radio galaxies. It was found that many such sources possess radio jets that point from the nuclei of the parent galaxies to the radio lobes. It is now believed

that the nucleus of an active galaxy supplies the basic energy that powers the radio emission, the energy being transported to the two lobes by twin beams of relativistic particles. Support for this theoretical picture exists, for example, in VLA maps (those made by the Very Large Array system of radio telescopes near Socorro, N.M.) of Cygnus A that show two jets emerging from the nucleus of the central galaxy and impacting the lobes at "hot spots" of enhanced emission (Figure 2). Other examples of this type are known, as are "head-tail" sources such as NGC 1265 where the motion of an active galaxy through the hot gas in a cluster of galaxies has apparently swept back the jets and lobes in a characteristic U shape.

Many jets are one-sided; *i.e.,* only one of the postulated twins is actually observed. This is usually interpreted to mean that the material moves at speeds approaching that of light. Relativistic effects—*e.g.,* the Doppler shift of the emitted photons—then boost the intrinsic luminosity of the jet pointing toward the observer and lower that of the counterjet, allowing measurements of limited dynamic range to detect only the former.

Support for the interpretation of relativistic jets exists in the phenomenon of "superluminal expansion." In very long baseline interferometry (VLBI) experiments performed by combining the simultaneous observations of several telescopes spaced by thousands of kilometres, radio astronomers have discovered that some of the compact radio sources located in the nuclei of active galaxies break into several components at high angular resolution. Moreover, in the course of a few years, the components move with respect to each other along a line projected against the sky that points toward more extended structures known from other observations (*e.g.,* large jets or lobes). If the source is placed at a (cosmological) distance appropriate for the redshift of the optical object, the projected motion across the line of sight has an apparent velocity that exceeds the speed of light. For example, in 3C 273, which possesses an optical jet in addition to the radio features discussed here, the apparent velocity measured over a time span from mid-1977 to mid-1980 amounted to about 10 times the speed of light.

Clearly, if Einstein's theory of special relativity is correct and if the assumed distance of the object is justified, then the computed "velocity" cannot represent the actual velocity of ejected collections of particles. The explanation now accepted by most astronomers is the model of a relativistic beam directed at a small angle to the observer along the line of sight. In this model a particle moving close to the speed of light would, according to a distant observer, almost catch up with the photons it emits, so that the duration of time that elapses between an earlier emission event and a later one is systematically underestimated by the observer (compared to one moving with the beam). Thus, under the appropriate circumstances, the apparent velocity (distance across the line of sight divided by apparent elapsed time) can exceed the actual velocity by a large factor. A beam moving at an actual velocity 99.5 percent the speed of light along an angle that lies 6° from the line of sight, for example, will appear to move across the line of sight at an apparent velocity of 10 times the speed of light.

The National Radio Astronomy Observatory, operated by Associated Universities, Inc., under contract with the National Science Foundation; observers, Richard A. Perley, John W. Dreher, and John J. Cowan

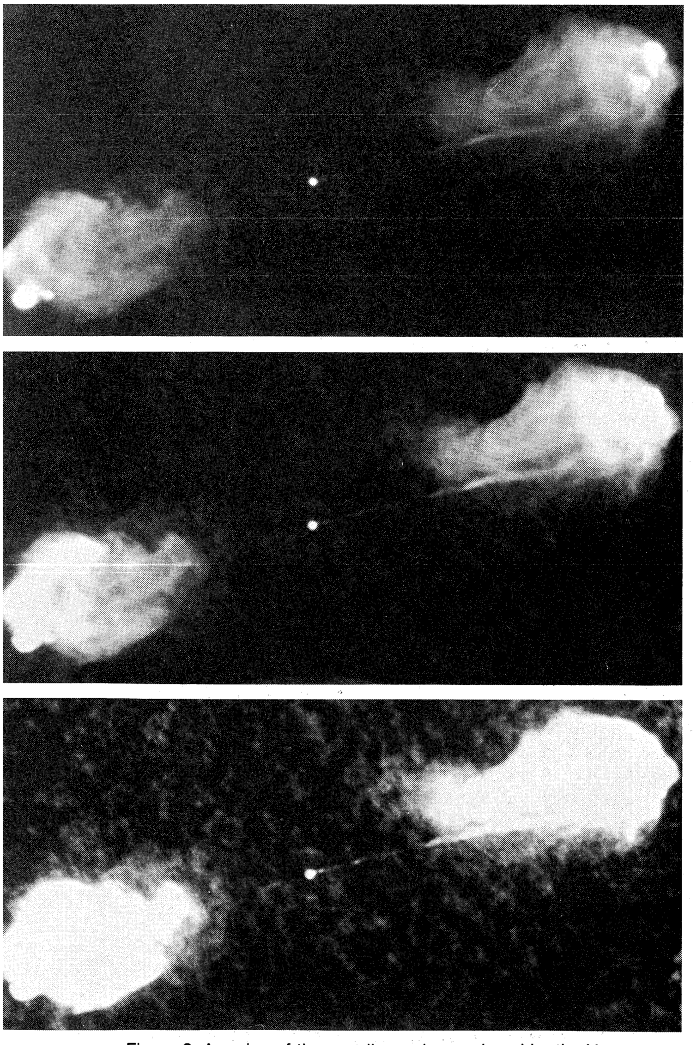


Figure 2: A series of three radiographs produced by the Very Large Array (VLA) of the double-lobed radio galaxy Cygnus A at six-centimetre wavelength. The radio jet extending from the core to the northwest lobe (right) is readily seen, but evidence for a counterjet to the southeast lobe (left) is marginal.

**Quasars.** The source 3C 273 mentioned above is officially classified by astronomers as a quasar. Quasars were first detected as unresolved sources in surveys conducted during the 1950s by radio astronomers in Cambridge, Eng. Optical photographs subsequently taken of their spectra showed locations for emission lines at wavelengths that were at odds with all celestial sources then familiar to astronomers. The puzzle was solved by the Dutch-born U.S. astronomer Maarten Schmidt, who announced in 1963 that the pattern of emission lines in 3C 273 could be understood as emanating from hydrogen atoms that had a redshift of 0.158. In other words, the wavelength of each line was 1.158 times longer than the wavelength measured in the laboratory where the source is at rest with respect to the observer. (The general formula is that if the factor is $1 + z$, astronomers say the astronomical source has a redshift of $z$. If $z$ turns out to be negative [*i.e.,* if $1 + z$ is less than 1], the source is said to be "blueshifted.")

Enormous redshifts of quasars

Schmidt's discovery raised immediate excitement, since 3C 273 had a redshift seen heretofore only among distant galaxies. Yet it had a starlike appearance, with an apparent brightness (but not a spectrum) in visible light not very different from that of a star at a distance of a few thousand light-years. If the quasar lay at a distance appropriate to remote galaxies a few thousand million light-years away, then it must be $10^{12}$ times brighter than an ordinary star. Similar conclusions were reached for other examples. Quasars seemed to be intrinsically brighter than even the largest known galaxies, yet they presented the pointlike image of a star.

Dimensions of quasars

A hint of the actual physical dimensions of quasars came when sizable variations of light output were seen from some quasars over a year or two. These variations implied that the dimensions of the regions emitting optical light must not exceed a light-year or two, since coherent fluctuations cannot be established in less time than it takes photons, which move at the fastest possible speed, to travel across the object. These conclusions were reinforced by later satellite measurements that showed that many quasars had even more X-ray emission than optical emission and that the total X-ray intensity could vary in a period of hours. In other words, quasars released energy at a rate exceeding $10^{12}$ suns, yet the central machine occupied a region only the size of the solar system.

Understandably, the implications were too fantastic for

many people to accept and a number of alternative interpretations were attempted. An idea common to several of the alternatives involved the proposal that the redshift of quasars arose from a different (*i.e.,* non-cosmological) origin than that accepted for galaxies. In that case, the distance to the quasars could be much less than assumed to estimate the energy outputs, and the requirements might be drastically relaxed. None of the alternative proposals, however, withstood close examination. For example, the proposal that quasars are objects ejected at relativistic speeds from galaxies meets the immediate objection that no known quasar has a blueshift. The counterproposal that perhaps all quasars were ejected from the Milky Way Galaxy and now all move away from the Earth is philosophically unpalatable to those who accept the full implications of the Copernican revolution. The hypothesis that quasars might be explained in terms of gravitational redshifts (the mechanism by which the wavelengths of photons increase as they propagate out of a deep gravitational well) meets with the objection that the width and intensity of the emission lines do not allow such an interpretation unless quasars have masses comparable to the largest galaxies and are located at distances that would make their redshifts likely to be of cosmological origin in the first place. The proposal that perhaps quasars have redshifts of a completely unknown origin cannot be logically assailed, but the notion is scientifically sterile.

In any case, there now exists ample evidence for the validity of attributing cosmological distances to quasars. The strongest arguments are the following. When the strong nonstellar light from the central quasar is eliminated by mechanical or electronic means, a fuzzy haze can sometimes be detected surrounding the quasar. When this light is examined carefully, it turns out to have the colour and spectral characteristics appropriate to a normal giant galaxy. This suggests that the quasar phenomenon is related to nuclear activity in an otherwise normal galaxy. In support of this view is the observation that quasars do not really form a unique class of objects. For example, not only are there elliptical galaxies that have radio-emission characteristics similar to those of quasars, but there are weaker radio sources among spiral galaxies (called Seyferts after their discoverer, the American astronomer Carl K. Seyfert), which have bright nuclei that exhibit qualitatively the same kinds of optical emission lines and nonstellar continuum light seen in quasars. There also are elliptical galaxies, N galaxies and the so-called BL Lacertae objects, which have nuclei that are exceptionally bright in optical light. Finally, a number of quasars—including the closest example, the famous source 3C 273—have been found to lie among clusters of galaxies. When the redshifts of the cluster galaxies are measured, they have redshifts that bracket the quasar's, suggesting that the quasar is located in a galaxy that is itself a cluster member.

To be sure, there also are a few quasars that exist in the same directions of the sky as known galaxies, but the quasars and the galaxies have very different redshifts. The majority opinion is that these represent chance superpositions, but surprises may yet be in order. Since the discovery in 1979 of a "double quasar" (two sources having nearly identical appearance and redshift), there has been a reassessment of the possible role that gravitational lensing (the bending of background rays by a massive foreground object) may play in focusing the light received from astronomical objects, and it is possible that chance superpositions may have an effect in producing some quasars, say, by intensifying the light from Seyfert galaxies undergoing starbursts in their nuclear regions. Nevertheless, the simplest hypothesis at the present time is that quasars merely lie at an extreme end of nuclear activity in a continuum of galaxy types that encompass Seyferts, N galaxies, BL Lacertae objects, and classical radio doubles.

**Black-hole model for active galactic nuclei.** The fact that the total output from the nucleus of an active galaxy can vary by large factors strongly suggests that the central machine is a single coherent body. The most promising theoretical candidate is a supermassive black hole that releases energy by the accretion of matter through a viscous disk. The idea is that the rubbing of gas in the shearing layers of a differentially rotating disk would frictionally generate heat, liberating photons as the mass moves inward and the angular momentum is transported outward. Scaled-down versions of the process have been invoked to model the primitive solar nebula and the disks that develop in interacting binary stars.

Mechanisms for energy release

The black hole has to be supermassive for its gravitational attraction to overwhelm the strong radiation forces that attempt to push the accreting matter back out. For a luminosity of $10^{46}$ erg/sec, which is a typical inferred X-ray value for quasars, the black hole must exceed $10^8$ solar masses. The event horizon of a $10^8$ solar-mass black hole has a circumference of about two light-hours. (The event horizon is the sphere around a black hole from inside of which not even photons can escape.) Matter orbiting in a circle somewhat outside of the event horizon would be hot enough to emit X rays and have an orbital period of several hours. If this material is lumpy or has a nonaxisymmetric distribution as it disappears into the event horizon, variations of the X-ray output on a time scale of a few hours might naturally be expected.

To produce $10^{46}$ erg/sec, the black hole has to swallow about two solar masses per year if the process is assumed to have an efficiency of about 10 percent for producing energy from accreted mass. The rough estimate that 10 percent of the rest energy of the matter in an accretion disk would be eventually liberated as photons, in accordance with Einstein's formula $E = mc^2$, should be contrasted with a total efficiency of about 1 percent in nuclear reactions if a mass of hydrogen were to be converted entirely into iron. If large-scale annihilation of matter and antimatter is excluded from consideration, the release of gravitational binding energy when matter settles onto compact objects is the most powerful mechanism for generating energy in the known universe. (Even supernovas use this mechanism, for most of the energy released in the explosion comes from the gravitational binding energy or mass deficit of the remnant neutron star.)

The matter swirling into the black hole may come externally from the sinking of a companion gas-rich galaxy into the central regions of the parent galaxy or from the settling of gas out of an intracluster medium; or it may come internally through the tidal breakup of stars that wander too close to the black hole, through mass loss from stars in the galaxy, or through bar forcing the inward spiraling of a larger gaseous disk. Gathering matter at a rate of two solar masses per year (90 percent of which ends up as the gravitating mass of the black hole) will build up a black hole of $10^8$ solar masses in several tens of millions of years. This estimate for the lifetime of an active galactic nucleus is in approximate accord with the statistics of such objects. This is not to imply that supermassive black holes at the centres of galaxies necessarily accumulate from a seed of very small mass by steady accretion. There remain many viable routes for their formation, the study of such processes being in a state of infancy.

**Observational tests.** If there are supermassive objects at the centres of elliptical galaxies, gravitational perturbations of the spatial distribution or velocity field of nearby stars may be discernible. For a spherical distribution of stars surrounding a black hole, theoretical calculations indicate that the number of stars per unit volume should rise as the negative 7/4 power of the radial distance and the dispersion of random velocities should rise as the negative 1/2 power of the radial distance from the black hole. In other words, rather than gently rounded or flat profiles as the centre is approached, cusps of stellar light and random velocities should be seen, the upturn beginning at a radial distance where the escape velocity from the black hole is comparable to the natural dispersion of random velocities in the central regions of an elliptical galaxy.

Except for the largest black holes or the nearest galaxies, the region interior to the turnup point is not resolvable by ground-based optical telescopes because of the blurring effects produced by turbulence in the Earth's atmosphere. Excess central starlight and velocity dispersions have been seen in M87—a giant elliptical with a well-known optical jet emerging from its nucleus, which is located in the Virgo cluster, the nearest large cluster of galaxies. The

excesses are consistent with a central black hole of several thousand million solar masses. Atmospheric effects, however, prevent astronomers from determining whether the upturns represent true cusps or merely shoulders that taper to constant values. Mere shoulders could be explained, without invoking a black hole, by the stars in the central regions of this galaxy having a nonstandard distribution of random velocities.

The closest galactic nucleus of all is of course located at the centre of the Milky Way Galaxy. Unfortunately the nucleus of the system is not observable at the wavelengths of visible light, ultraviolet light, or soft X rays (those of lower energy) because of the heavy absorption by intervening dust. It can be probed by radio, infrared, hard (higher energy) X-ray, and gamma-ray techniques; such studies have revealed many intriguing features.

Among the more suggestive findings is the detection of gamma-ray emission from within 4° of the galactic centre (the experiment did not have high directional sensitivity) at a line energy of 510,000 electron volts, corresponding to the annihilation of electrons with positrons (the antiparticle of an electron). Evidently microscopic events energetic enough to produce electron–positron pairs occur at the centre of the Galaxy, and their subsequent annihilation of each other (or positrons with ambient electrons) produce the observed gamma-ray line emission. The process of pair production has been theoretically studied as a possibility in the environment of an accreting black hole.

The most likely candidate for the nucleus of the Galaxy has long been regarded to be a compact radio-continuum source denoted Sagittarius A*. This synchrotron-radiation source is unique in the Galaxy: it is variable on a time scale of one day, implying that the radio emission arises from a region with dimensions not much larger than the solar system; and it shows evidence for synchrotron self-absorption, a condition consistent with a region being compactly filled with relativistic particles and fields. Hard X-ray observations of the central region of the Galaxy, however, reveal only low-level emission from a diffuse component and several discrete sources with characteristics similar to coronal emission from luminous young stars. In particular, a discrete X-ray source is found at the position of Sagittarius A*, but its luminosity is only about $10^{35}$ erg/sec, which hardly compares with a full-blown quasar. Broadband, near-infrared measurements at a wavelength centred near two micrometres (0.002 millimetres) show the presence of a dense star cluster. Surprisingly the maximum concentration of light of the star cluster does not seem to centre on Sagittarius A*, as it should if the radio source were a very massive object (both because the object would attract stars toward itself and because the "dynamical friction" of the stars would cause any heavy object to sink eventually to the centre of the star cluster).

Sagittarius A*

Spectroscopic investigations at radio and infrared wavelengths allow a more promising interpretation. Spectroscopic observations at infrared wavelengths (*e.g.*, "vibrational transition" of shock-excited molecular hydrogen, "fine-structure" line of atomic oxygen) demonstrate that the molecular gas at distances of several light-years from the galactic centre is distributed in a (tilted) ring having circular rotation. The inferred rotational velocities are consistent with motion under the central force field of an object with a mass of about several million solar masses. Unfortunately the molecular gas disappears before the centre can be approached very closely; fortunately its disappearance is compensated by the appearance of ionized gas within the central few light-years. The densest parts of this ionized gas appear arranged in the form of a "mini-spiral" with three "arms," whose Doppler shifts can be measured by infrared techniques (*e.g.*, via the fine-structure line of singly ionized neon) or by radio techniques (*e.g.*, via the "radio-recombination" lines of hydrogen). The interpretation, however, is made difficult by this ionized gas having a complex velocity field and by one of the arms of the mini-spiral apparently interacting with a supernova remnant near the centre of the Galaxy. Different models of the many competing processes lead to different interpretations. One simple attempt identifies a special streamer of ionized gas as an infalling parabolic trajectory, yielding a value of 4,000,000 solar masses in a compact object at the galactic nucleus. If the Galaxy has a central black hole, this is probably the best existing estimate of its mass.

Radio-continuum studies on a scale of hundreds of light-years from the Galaxy's centre show the nucleus to be embedded in an extraordinary arc of filamentary streamers that passes perpendicularly through the galactic plane. Magnetic fields may play a role in defining the streamers, perhaps in a fashion analogous to the eruption of solar prominences. Details such as these can be seen only because the nucleus of the Galaxy is so close (a mere 30,000 light-years away). This complexity should serve as a sobering reminder that most theoretical models of the active nuclei of external galaxies must vastly oversimplify the actual state of affairs.

## OTHER COMPONENTS

Every second of every day, the Earth is bombarded by high-speed particles, electromagnetic radiation, and perhaps gravitational waves of cosmic origin. As has already been discussed, a part of this steady rain is, directly or indirectly, of planetary, stellar, or galactic origin, but another part may be remnants from a time in the universe before there were any planets, stars, or galaxies.

**Cosmic rays and magnetic fields.** In the years following the discovery of natural radioactivity by the French physicist Henri Becquerel in 1896, investigators used ionization chambers to detect the presence of the fast charged particles that are produced in the phenomenon. These workers found that low-level ionization events still occurred even when the source of radioactivity was removed. The events persisted with heavy shielding, and in 1912 Victor F. Hess of the United States found that they increased drastically in intensity if the detecting instruments were carried to high altitudes by balloons. There was little difference in intensity between day and night; thus, the Sun could not be the primary source. The penetrating radiation had to have a cosmic component, and the earliest suggestion was that it was composed of high-energy photons, gamma rays—hence, the name cosmic rays. In 1927 it was shown that the cosmic-ray intensity was higher at the magnetic poles than at the magnetic equator. For the incoming trajectories to be affected by the geometry of the Earth's magnetic field, cosmic rays had to be charged particles.

It is now known that cosmic rays come with both signs of electric charge and with a wide distribution of energies. About 83 percent of the positively charged component of cosmic rays consists of protons, the nuclei of hydrogen atoms, and about 16 percent of alpha particles, the nuclei of helium nuclei (Figure 3). The nuclei of heavier atoms occur roughly in their cosmic abundances except that the light elements, lithium, beryllium, and boron—which are quite rare elsewhere in the universe—are vastly overrepresented in the cosmic rays. The negatively charged component consists of mostly electrons at a level of 1 percent of the protons. Positrons also can be found, about 10 percent as frequently as electrons. A very slight contribution from antiprotons is also known. Cosmic-ray positrons and antiprotons are believed to be by-products of collisions between cosmic-ray nuclei and the ambient atomic nuclei that exist in interstellar gas clouds. Cosmic gamma rays, which have been detected emanating from the Milky Way and show a strong correlation with the distribution of interstellar gas, are another manifestation of such collisions.

Cosmic-ray composition

The cosmic-ray protons that freely enter the solar system, in spite of the outward sweep of the solar wind and the magnetic fields it carries, have energies ranging from a few times their rest energies to 1,000,000 times and more. Thus, these particles must move at speeds approaching the speed of light. In this range, the number of particles at energy $E$ varies with $E$ to the negative 2.7 power. A similar decreasing power law seems to hold for cosmic-ray electrons with energies from a few thousand to tens of thousands times their rest energies. Within uncertainties, this energy distribution is consistent with the synchrotron-radiation interpretation of the nonthermal radio emission from the Galaxy. At higher energies, there are fewer cosmic-ray electrons than predicted by extrapolation of the power law found at lower energies, and this depletion

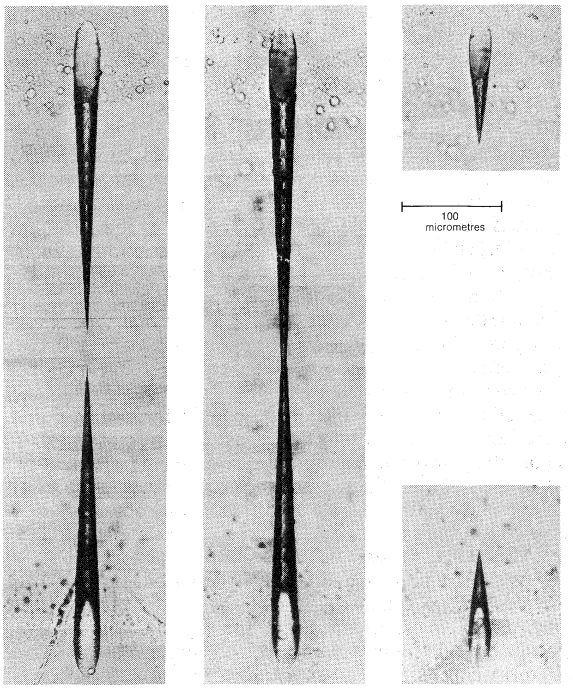



Figure 3: Cosmic-ray particles traversing a sheet of Lexan plastic leave ionization tracks on it. These tracks are made visible by chemical etching. The properties of the tracks and the subsequent etch patterns depend on the energy loss of the particles.
By courtesy of P. Buford Price

can be understood on the basis of the large synchrotron-radiation losses suffered by the most energetic electrons.

Above $10^7$ times the rest energy of the proton, there also are fewer positively charged particles than predicted by the extrapolation of the power law $E^{-2.7}$; however, synchrotron losses cannot account for this deficiency. A more likely interpretation is that the cosmic-ray nuclei of lower energies are commonly produced and confined to the Galaxy, whereas those with very high energies may have an origin in exotic or even extragalactic objects. This is consistent with the fact that protons with energies less than $10^7$ times their rest energies would be bent by the interstellar magnetic field to follow spiraling trajectories that would be confined to the thickness of the galactic disk. Nevertheless, these particles can eventually escape from the disk if the magnetic fields buckle out of the galactic plane (as they sometimes do because of certain instabilities).

An estimate of the total residence time of cosmic-ray nuclei within the disk of the Galaxy can be obtained by examining the anomalous abundances of lithium, beryllium, and boron. These elements are only somewhat less abundant in cosmic rays than carbon, nitrogen, and oxygen, and this has been conventionally interpreted to mean that the former group were mostly produced by spallation reactions (breakup of heavier nuclei) of the latter group as the cosmic-ray particles traversed interstellar space and interacted with the matter there. From the amount of spallation that has occurred, it can be estimated that the cosmic rays reside, on average, roughly $10^7$ years among the gas clouds in the galactic disk before escaping.

Origin of cosmic rays

The origin of cosmic rays is an incompletely resolved problem. At one time, astronomers believed that all cosmic rays, except those at the highest energies, originated with supernova explosions. The total energetics is right, and the presence in cosmic rays of nuclei as heavy as iron, etc., could receive a natural explanation under the supernova hypothesis. Unfortunately, doubt was cast on the hypothesis by later work that questioned first, whether particles could really be accelerated to cosmic-ray energies in a single supernova shock, and second, whether these particles, even if accelerated, could propagate through the interstellar medium very far from the site of the original explosion. The second objection also applies to other possible point sources, such as pulsars.

A more promising possibility seems to be the proposal that cosmic rays are accelerated to their high energies by repeated reflections in magnetic shock waves in the interstellar medium (whose ultimate energy may be derived from the total of all supernova explosions). The idea is that gas and the magnetic field threading it move at very different speeds on the two sides of the front of a shock wave. Cosmic-ray particles rattling through magnetic inhomogeneities may be shuttled back and forth between these two regions, gaining statistically an extra boost in energy every time they "bounce" off the moving set of magnetic field lines. The process is akin to the increasing energy that would be gained by a tennis ball in the absence of air drag if it were banged back and forth between a vigorously swinging player and a stationary wall. The great attractiveness of the strong shock-wave picture for accelerating cosmic rays is that it automatically gives, in the simplest models, a decreasing power-law distribution of particle energies. The exponent is 2 instead of the measured 2.7, and the discrepancy is believed to be related to an energy-dependent escape rate from the region of acceleration. The enhancement of the escape rate with increasing energy is not completely understood, but no fundamental obstacle appears likely in this direction to rule out the shock-acceleration model.

A more serious failing of this model is that it cannot easily explain the origin of ultra-high-energy cosmic rays, particles with energies that lie between $10^8$ and $10^{11}$ times the rest energy of the proton. There is some indication from measurements of ultra-high-energy gamma rays from certain binary X-ray sources that these objects may copiously produce ultra-high-energy cosmic rays, but the exact acceleration mechanism remains obscure. At the highest observed cosmic-ray energies, the particles arrive preferentially from northern galactic latitudes, a fact interpreted by some to indicate a large contribution from the Virgo supercluster. In this picture, even higher energy cosmic rays from more distant parts of the universe do not reach the Earth because such particles would suffer serious losses en route as they interact with the photons of the cosmic microwave background.

**Microwave background radiation.** Beginning in 1948 the Russian-born U.S. cosmologist George Gamow and his coworkers, Ralph Alpher and Robert Herman, investigated the idea that the chemical elements might have been synthesized by thermonuclear reactions that took place in a primeval fireball. The high temperature associated with the early universe would give rise to a thermal radiation field, which has a unique distribution of intensity with wavelength (known as Planck's radiation law), that is a function only of the temperature. As the universe expanded, the temperature would have dropped, each photon being redshifted by the cosmological expansion to longer wavelength, as the American physicist Richard C. Tolman had already shown in 1934. By the present epoch, the radiation temperature would have dropped to very low values, about 5° above absolute zero (0 K, or −273° C) according to the estimates of Alpher and Herman.

Interest in these calculations waned among most astronomers when it became apparent that the lion's share of the synthesis of elements heavier than helium must have occurred inside stars rather than in a hot big bang. During the early 1960s physicists at Princeton University, as well as in the Soviet Union, took up the problem again and began to build a microwave receiver that might detect, in the words of the Belgian cleric and cosmologist Georges Lemaître, "the vanished brilliance of the origin of the worlds."

Discovery of relict radiation

The actual discovery of the relict radiation from the primeval fireball, however, occurred by accident. In experiments conducted in connection with the first Telstar communication satellite, two scientists of the Bell Telephone Laboratories, Arno Penzias and Robert Wilson, measured excess radio noise that seemed to come from the sky in a completely isotropic fashion. When they consulted Bernard Burke of the Massachusetts Institute of Technology about the problem, Burke realized that Penzias and Wilson had most likely found the cosmic background radiation that Robert H. Dicke, P.J.E. Peebles, and their

colleagues at Princeton were planning to search for. Put into touch with one another, the two groups published simultaneously in 1965 papers detailing the prediction and discovery of a universal thermal radiation field with a temperature of about 3 K. Subsequent investigations have substantiated this finding in every respect; the accepted value for the cosmic radiation temperature is 2.7 K.

In the mid-1980s efforts were directed toward the measurement of possible distortions in the wavelength distribution of the radiation field or its angular position in the sky from the ideal Planck form and the total isotropy predicted by the simplest theory. Also of interest is whether there is any net polarization in the radiation field. Such departures could possibly yield information about the slight fluctuations of density and velocity in the early universe that are believed to have given rise to structure (*e.g.*, the observed distribution of galaxies) in an otherwise homogeneous and uniformly expanding medium. So far, except for a so-called dipole anisotropy, any departures that might exist lie below quite small detection limits—on the order of one part in $10^4$.

The high degree of observed isotropy is a mixed blessing for theoretical cosmology. On the one hand, it provides a strong justification for the assumption of homogeneity and isotropy that is common to most cosmological models. On the other hand, such homogeneity and isotropy are difficult to explain because of the "event horizon" problem. In the context of the cosmic microwave background, the problem can be expressed in the following way. Consider the background radiation coming to an observer on the Earth from any two opposite sides of the sky. Clearly whatever are the ultimate sources (hot plasma) for this radiation, the photons, traveling at the speed of light since their emission by the sources, have only had time to reach Earth now. The matter on one side of the sky could not have had time to have "communicated" with the matter on the other side (they are beyond each other's event horizon), so how is it possible (with respect to an observer in the right rest frame) that they "know" to have the same temperature to a precision approaching one part in $10^4$? What accounts for the high degree of angular isotropy of the cosmic microwave background? Or, for that matter, for the large-scale distribution of galaxies? As will be seen below (*Cosmological models*), a mechanism called inflation may offer an attractive way out of this dilemma.

The dipole anisotropy refers to the cosmic background radiation being 3.2 millidegrees warmer than average toward a direction a little south of the constellation Leo, and 3.2 millidegrees cooler toward the opposite direction of the sky, a little north of the constellation Aquarius. This result has a simple interpretation—namely, that the solar system is moving with a velocity of about 360 km/sec in the general direction of Leo, and that this velocity, which is 0.12 percent of the speed of light, Doppler-shifts the energies of the photons up or down by this fraction in the directions parallel or antiparallel to the motion. If the motion of the solar system around the centre of the Galaxy and the motion of the Galaxy relative to the centre of mass of the Local Group are taken into account, this implies that the Local Group of galaxies is moving at a velocity of about 600 km/sec in a direction that is some 45° from the direction of the Virgo cluster of galaxies. Such motion is not measured relative to the galaxies themselves (the Virgo galaxies have an average velocity of recession of about 1,000 km/sec with respect to the Milky Way system), but relative to a local frame of reference in which the cosmic microwave background radiation would appear completely isotropic (at least, to one part in $10^4$) and would ideally have a perfect Planck spectrum with a single radiation temperature.

Velocity of the Local Group

The origin of the "peculiar velocity" of 600 km/sec for the Local Group is not completely understood. There is probably a component of this velocity that is induced by the gravitational attraction of the excess mass above the cosmological mean represented by the Virgo cluster; however, it is now believed the Virgo component is fairly small—only 200–300 km/sec. A more important contribution may come from the mass within a distance of 100,000,000 light-years connected to the Local Supercluster.

One can compute, under certain simplifying assumptions of the three-dimensional distribution, the excess mass needed to produce the observed peculiar velocity of the Local Group. Although the uncertainties of the calculation are large, one finds that the mass typically exceeds by a healthy margin that which can be inferred to exist in the observable parts of the galaxies in the Local Supercluster. One way to state the result is that the dynamical models require an average mass-to-light ratio of several hundred, whereas observed stellar systems rarely have mass-to-light ratios much in excess of 20 or 30. This is another manifestation of the hidden-mass problem, one occurring on the scale of hundreds of millions of light-years.

**Intergalactic gas.** At one time it was thought that large amounts of mass might exist in the form of primordial gas clouds in the spaces between galaxies. One by one, however, the forms that this intergalactic gas might take were eliminated until the only possible form that might have escaped early detection was a very hot plasma. Thus, there was considerable excitement and speculation when astronomers found evidence in the early 1970s for a seemingly uniform and isotropic background of hard X-radiation (photons with energies greater than 1,000,000 electron volts). There also was a diffuse background of soft X rays, but this had a patchy distribution and was definitely of galactic origin—hot gas produced by many supernova explosions inside the Galaxy. The hard X-ray background, in contrast, seemed to be extragalactic, and a uniform plasma at a temperature of roughly $10^8$ K was a possible source. The launch in 1978 of an imaging X-ray telescope aboard the "Einstein Observatory" (the HEAO 2 satellite), however, showed that a large fraction of the seemingly diffuse background of hard X rays, perhaps all of it, could be accounted for by a superposition of previously unresolved point sources—*i.e.*, quasars and QSOs. Unfortunately subsequent research demonstrated that the shape of the X-ray spectrum of these objects does not match that of the diffuse background.

X-ray emission from hot intergalactic gas

Very hot gas that emits X rays at tens to hundreds of millions of kelvins does indeed reside in the spaces between galaxies in rich clusters, and the amount of this gas seems comparable to that contained in the visible stars of the galaxies; however, because rich clusters are fairly rare in the universe, the total amount of such gas is small compared to the total mass contained in the stars of all galaxies. Moreover, an emission line of iron can frequently be detected in the X-ray spectrum, indicating that the intracluster gas has undergone nuclear processing inside stars and is not of primordial origin. Except inside the bodies of some of the supergiant galaxies, the cooling time for the intracluster gas exceeds the age of the universe, $10^{10}$ years, so no difficulty arises from explaining the observed temperatures of $10^8$ K. Presumably this gas was expelled from the galaxies in the cluster during an earlier epoch of high activity. The existence of substantial amounts of primordial intergalactic gas in the vast regions of the Cosmos not occupied by rich clusters thus remains speculation rather than a confirmed fact. The same holds true for the hypothesis of primordial intergalactic magnetic fields.

**Low-energy neutrinos.** Another hypothesized component of the Cosmos is a universal sea of very low-energy neutrinos. Although nearly impossible to detect by direct means, the existence of this sea has a strong theoretical basis. This basis rests with the notion that a hot big bang will produce not only a primeval fireball of electromagnetic radiation but also enormous numbers of neutrinos and antineutrinos (both referred to in cosmological discussions as neutrinos for brevity's sake). Estimates suggest that every cubic metre of space in the universe contains about 100,000,000 low-energy neutrinos. This number considerably exceeds the cosmological density of atomic nuclei (mostly hydrogen) obtained by averaging the known matter in the universe over scales of hundreds of millions of light-years. The latter density amounts to less than one particle per cubic metre of space. Nevertheless, because neutrinos interact with matter only weakly (they do not, for example, emit electromagnetic radiation), they can be detected experimentally by very sophisticated instruments only if they have relatively high energies. The very low-

energy neutrinos of cosmological origin cannot be observed by any conventional means known at present.

Such low-energy neutrinos, nonetheless, attracted considerable astronomical interest during the late 1970s, because experiments conducted in the Soviet Union and the United States suggested, contrary to the prevailing belief in particle physics, that neutrinos may possess a nonzero rest mass. Even if the rest mass were very small—say, 10,000 times smaller than the rest mass of the electron, the lightest known particle of matter—the result could be of great potential importance because neutrinos, if relatively abundant cosmologically, could then be the dominant source of mass in the universe. Unfortunately, later experiments cast doubts on the conclusions of the earlier findings, and theoretical investigations of "massive neutrinos" as the hidden mass in the universe turned up as many new difficulties to be explained as possible solutions to old problems.

**Gravitational waves.** Superficially, there are many similarities between gravity and electricity; for example, Newton's law for the gravitational force between two point masses and Coulomb's law for the electric force between two point charges both vary as the inverse square of the separation distance. Yet, in James Clerk Maxwell's theory for electromagnetism accelerated charges emit signals (electromagnetic radiation) that travel at the speed of light, whereas in Newton's theory of gravitation accelerated masses transmit information (action at a distance) that travels at infinite speed. This dichotomy is repaired by Einstein's theory of gravitation, wherein accelerated masses also produce signals (gravitational waves) that travel only at the speed of light. And just as electromagnetic waves can make their presence known by the pushing to and fro of electrically charged bodies, so can gravitational waves be detected, in principle, by the tugging to and fro of massive bodies. However, because the coupling of gravitational forces to masses is intrinsically much weaker than the coupling of electromagnetic forces to charges, the detection of gravitational radiation is much more difficult than that of electromagnetic radiation. Indeed, since the time of Einstein's formulation of general relativity in the early 1900s, there has yet to be a single instance of the detection of gravitational waves that is direct and undisputed.

Binary pulsar as a possible source of gravitational radiation

There are, however, some indirect pieces of evidence that accelerated astronomical masses do emit gravitational radiation. The most convincing concerns radio-timing observations of a pulsar located in a binary star system with an orbital period of 7.75 hours. This object, discovered in 1974, has a pulse period of about 59 milliseconds that varies by about one part in 1,000 every 7.75 hours. Interpreted as Doppler shifts, these variations imply orbital velocities on the order of $^1/_{1,000}$ the speed of light. The nonsinusoidal shape of the velocity curve with time allows a deduction that the orbit is quite noncircular (indeed, an ellipse of eccentricity 0.62 whose long axis precesses in space by 4.2° per year). It is now believed that the system is composed of two neutron stars, each having a mass of about 1.4 solar masses, with a semimajor axis separation of only 2.8 solar radii. According to Einstein's theory of general relativity, such a system ought to be losing orbital energy through the radiation of gravitational waves at a rate that would cause them to spiral together on a time scale of about $3 \times 10^8$ years. The observed decrease in the orbital period in the years since the discovery of the binary pulsar does certainly indicate that the two stars are spiraling toward one another at exactly the predicted rate.

The implosion of the core of a massive star to form a neutron star prior to a supernova explosion, if it occurs in a nonspherically symmetric way, ought to provide a powerful burst of gravitational radiation. Simple estimates yield the release of a fraction of the mass-energy deficit, roughly $10^{53}$ erg, with the radiation primarily coming out at wave periods between the vibrational period of the neutron star, about 0.3 millisecond, and the gravitational-radiation damping time, about 300 milliseconds.

A cosmic background of gravitational waves is a possibility that has also been discussed. Such a background might be generated if the early universe expanded in a chaotic fashion rather than in the smooth homogeneous fashion that is currently observed. The energy density of the gravitational waves produced, however, is unlikely to exceed the energy density of electromagnetic radiation, and each graviton (the gravitational analogue of the photon) would be susceptible to the same cosmological redshift by the expansion of the universe. A roughly thermal distribution of gravitons at a present temperature of about 1 K would be undetectable by foreseeable technological developments in gravitational-wave astronomy.

## Large-scale structure and expansion of the universe

Hubble inferred a uniformity in the spatial distribution of galaxies through number counts in deep photographic surveys of selected areas of the sky. This inference applies only to scales larger than several hundred million light-years. On smaller scales, galaxies tend to bunch together in clusters and superclusters, a fact already known to Hubble, who deliberately avoided the more conspicuous examples in order not to bias his results. This clustering did excite debate among both observers and theorists in the earliest discussions of cosmology, particularly over the largest dimensions where there are still appreciable departures from homogeneity and over the ultimate cause of the departures. During the 1950s and early 1960s, however, attention tended to focus on homogeneous cosmological models because of the competing ideas of the big bang and steady state scenarios. Only after the discovery of the cosmic microwave background—which, together with the successes of primordial nucleosynthesis, signaled a clear victory for the hot big bang picture—did the issue of departures from homogeneity in the universe again attract widespread interest.

From a more pragmatic point of view, clusters and groups of galaxies are important to cosmological studies because they are useful in establishing the extragalactic distance scale. A fundamental problem that recurs over and over again in astronomy is the determination of the distance to an object. Individual stars in star clusters and associations provide an indispensable tool in gauging galactic distances. The brightest stars—in particular, the brightest variable stars among the so-called Cepheid class—allow the distance ladder to be extended to the nearest galaxies; but at distances much larger than $10^7$ light-years individual stars become too difficult to resolve, at least from the ground, and astronomers have traditionally resorted to other methods. (The original confusion of H II regions with individual stars by Hubble led him to underestimate the distances to relatively nearby groups of galaxies.) Clusters and groups of galaxies play a key role in the calibration of these other methods because the individual galaxies of a cluster or group can usually be assumed to lie at nearly the same distance, and there is a good chance that the distance to one of the member galaxies can be determined if it resembles a nearby system whose properties have already been studied at closer range. The establishment of the extragalactic distance scale is therefore very much a bootstrap operation; errors at one level propagate to the next. Accordingly, confidence in the entire scheme can be gained only if each step outward is corroborated by measurements using several techniques.

Distance determinations

### CLUSTERING OF GALAXIES

Clusters of galaxies fall into two morphological categories: regular and irregular. The regular clusters show marked spherical symmetry and have a rich membership. Typically, they contain thousands of galaxies, with a high concentration toward the centre of the cluster. Rich clusters, such as the Coma cluster, are deficient in spiral galaxies and are dominated by ellipticals and S0s. The irregular clusters have less well-defined shapes, and they usually have fewer members, ranging from fairly rich systems such as the Hercules cluster to poor groups that may have only a few members. Galaxies of all types can be found in irregular clusters: spirals and irregulars, as well as ellipticals and S0s. Most galaxies are not to be found in rich clusters but rather in loose groups. The Milky Way Galaxy belongs to one such loose group—the Local Group.

Types of clusters

**The Local Group.** The Local Group contains some two dozen galaxies, of which the dominant pair is the Milky Way and Andromeda, both giant spirals of Hubble type Sb and luminosity class II. The distance to the Andromeda system was first measured by Hubble, but his estimate was too low by a factor of two because astronomers at that time did not recognize the distinction between variable stars belonging to Population II (like those studied by Shapley) and Population I (those studied by Hubble). Another spiral in the Local Group—M33, Hubble type Sc and luminosity class III—is fairly prominent, but the rest are intermediate to dwarf systems, either irregulars or ellipticals. Most of the mass of the Local Group is associated with the Milky Way and Andromeda, and with a few exceptions the smaller systems tend to congregate about one or the other of these galaxies. The size of the Local Group is therefore only slightly larger than the 2,000,000 light-years separating the Milky Way system and the Andromeda galaxy, and the centre of mass lies roughly halfway between these two giants.

The Andromeda galaxy is one of the few galaxies in the universe that actually has a velocity of approach with respect to the centre of the Galaxy. If this approach results from the reversal by the mutual gravitational attraction of a former recession, then the total mass of the Local Group probably amounts to a few million million solar masses. This is more than the mass inferred for the optically visible parts of the galaxies and is another manifestation of the problem of hidden mass.

**Neighbouring groups and clusters.** Beyond the fringes of the Local Group lie many similar small groups. The best studied of these is the M81 group, whose dominant galaxy is the spiral galaxy M81. Much like the Andromeda and Milky Way systems, M81 is of Hubble type Sb and luminosity class II. The distance to M81, as well as to the outlying galaxy NGC 2403, can be determined from various stellar calibrators to be about 10,000,000 light-years. It is not known whether NGC 2403 and its companion NGC 2366 are truly bound to M81 or whether they constitute an independent pair seen by chance to lie near the M81 group. If they are bound to M81, then a measurement of their velocity along the line of sight relative to that of M81 yields, by an argument similar to that used for the Andromeda and Milky Way, an estimate of the gravitating mass of M81. This estimate equals $2 \times 10^{12}$ solar masses and exceeds by an order of magnitude what is deduced from measurements of the rotation curve of M81 inside its optically visible disk.

The M81 group

The M81 group also has a few normal galaxies with classifications similar to those of galaxies in the Local Group, and it was noticed by some astronomers that the linear sizes of the largest H II regions (which are illuminated by many OB stars) in these galaxies had about the same intrinsic sizes as their counterparts in the Local Group. This has led Allan Sandage and Gustav Tammann to the technique of using the sizes of H II regions as a distance indicator, because a measurement of their angular sizes, coupled with knowledge of their linear sizes, allows an inference of distance.

This method can be used, for example, to obtain the distance to the M101 group, whose dominant galaxy M101 is a supergiant spiral—the closest system of Hubble type Sc and luminosity class I. Since Sc I galaxies are the most luminous spiral galaxies, with very large H II regions strung out along their spiral arms, determining the distance to M101 is a crucial step in obtaining the absolute sizes of the giant H II regions of these important systems. The sizes of the H II regions in the companion galaxies of M101 compared with the calibrated values for nearby galaxies of the same class yield a distance to the M101 group of approximately 20,000,000 light-years.

Having calibrated the sizes of the giant H II regions in M101, Sandage and Tammann could then determine the distances to 50 field Sc I galaxies. Once this had been done, it became possible to measure the absolute brightnesses of Sc I galaxies, and it has been ascertained that all such systems have nearly the same luminosity. Since Sc I galaxies like M101 or M51 can be recognized on purely morphological grounds (well-organized spiral structure with massive arms dominated by giant H II regions), they serve as "standard candles" for determining the distances to irregular clusters that contain them (*e.g.*, the Virgo cluster containing the Sc I galaxy M100).

Virgo and Coma clusters

The Virgo cluster is the closest large cluster and is located at a distance of about 50,000,000 light-years in the direction of the constellation Virgo. About 200 bright galaxies occupy the Virgo cluster, scattered in various subclusters whose largest concentration (near the famous system M87) is about 5,000,000 light-years in diameter. Of the galaxies in the Virgo cluster, 68 percent are spirals, 19 percent are ellipticals, and the rest are irregulars or unclassified. Although spirals are more numerous, the four brightest galaxies are giant ellipticals, among them M87. Calibration of the absolute brightnesses of these giant ellipticals allow a leap to the distant regular clusters.

The nearest rich cluster containing thousands of systems, the Coma cluster, lies about seven times farther than the Virgo cluster in the direction of the constellation Coma Berenices. The main body of the Coma cluster has a diameter of about 25,000,000 light-years, but enhancements above the background can be traced out to a supercluster having a diameter of about 200,000,000 light-years. Ellipticals or S0s constitute 85 percent of the bright galaxies in the Coma cluster; the two brightest ellipticals in Coma are located near the centre of the system and are individually more than 10 times as luminous as the Andromeda galaxy. These galaxies have a swarm of smaller companions orbiting them and may have grown to their bloated sizes by a process of "galactic cannibalism" like that hypothesized to explain the supergiant elliptical cD systems (see above).

The spatial distribution of galaxies in rich clusters such as the Coma cluster closely resembles what one would expect theoretically for a bound set of bodies moving in the collective gravitational field of the system. Yet, if one measures the dispersion of random velocities of the Coma galaxies about the mean, one finds that it amounts to almost 900 km/sec. For a galaxy possessing this random velocity along a typical line of sight to be gravitationally bound within the known dimensions of the cluster requires Coma to have a total mass of about $5 \times 10^{15}$ solar masses. The total luminosity of the Coma cluster is measured to be about $3 \times 10^{13}$ solar luminosities; therefore, the mass-to-light ratio in solar units required to explain Coma as a bound system exceeds by an order of magnitude what can be reasonably ascribed to the known stellar populations. A similar situation exists for every rich cluster that has been examined in detail. This hidden mass problem for rich clusters was known to the Swiss astronomer Fritz Zwicky as early as 1933. The discovery of X-ray emitting gas in rich clusters has alleviated the dynamical problem by a factor of about two, but a substantial discrepancy remains.

**Superclusters.** In 1932 Harlow Shapley and Adelaide Ames introduced a catalog that showed the distributions of galaxies brighter than 13th magnitude to be quite different north and south of the plane of the Galaxy. Their study was the first to indicate that the universe might contain substantial regions that departed from the assumption of homogeneity and isotropy. The most prominent feature in the maps they produced in 1938 was the Virgo cluster, but already apparent at that time were elongated appendages that stretched on both sides of Virgo to a total length exceeding 50,000,000 light-years. This configuration is the kernel of what came to be known later through the work of Erik Holmberg, Gerard de Vaucouleurs, and George O. Abell as the Local Supercluster, a flattened collection of about 100 groups and clusters of galaxies including the Local Group. The Local Supercluster is centred approximately on the Virgo cluster and has a total extent of roughly 200,000,000 light-years, though precise boundaries are difficult to define inasmuch as the local enhancement in numbers of galaxies above the cosmological average probably just blends smoothly into the background.

The Local Supercluster

Also apparent in the Shapley–Ames maps were three independent concentrations of galaxies, separate superclusters viewed from a distance. Astronomers now believe superclusters fill perhaps 10 percent of the volume of the universe. Most galaxies, groups, and clusters belong to superclusters, the space between superclusters being rela-

tively empty. The dimensions of superclusters range up to a few hundred million light-years. For larger scales, the distribution of galaxies is essentially homogeneous and isotropic—that is, there is no evidence for the clustering of superclusters. This fact can be understood by recognizing that the time it takes a randomly moving galaxy to traverse the long axis of a supercluster is comparable to the age of the universe. Thus, if the universe started out homogeneous and isotropic on small scales, there simply has not been enough time for it to become inhomogeneous on scales much larger than superclusters. This interpretation accords with the observation that superclusters themselves look dynamically unrelaxed—that is, they lack the regular equilibrium shapes and central concentrations that typify systems well mixed by several crossings.

**Statistics of clustering.** The description of galaxy clustering given above is qualitative and therefore open to a charge of faulty subjective reasoning. To remove human biases, it is possible to take a statistical approach, a path pioneered by Jerzy Neyman and Elizabeth L. Scott and extended by P.J.E. Peebles and his coworkers. Their line of attack begins by considering the correlation of the angular positions of galaxies in the northern sky surveyed by C.D. Shane and C.A. Wirtanen of Lick Observatory. If the intrinsic distribution in the direction along the line of sight is assumed to be similar to that across it, then it is possible to derive from the analysis the two-point correlation function that expresses the joint probability for finding two galaxies in certain positions separated by a distance $r$. Of special interest is the enhancement in the probability above a random distribution of locations, which Peebles and his colleagues found to be well represented, up to scales of about 50,000,000 light-years, as a simple power law, $(r/r_0)^{-1.8}$, with $r_0$ equal to about 20,000,000 light-years. Beyond 50,000,000 light-years, the enhancement drops more quickly with distance than $r^{-1.8}$, but the exact way it does this is somewhat controversial.

To summarize, then, when one knows a galaxy to be present, there is a considerable statistical enhancement in the likelihood that other galaxies will be near it for distances of 50,000,000 light-years or less, whereas at much larger distances the probability drops off to the expectation for a purely random distribution in space. This result provides a quantitative expression for the phenomenon of galaxy clustering. A similar power-law representation seems to be appropriate for the correlation of galaxy clusters, which provides empirical evidence for the phenomenon of superclustering.

In addition to angular positions, it is possible to derive empirical information about the large-scale distribution of galaxies in the direction along the line of sight by examining the redshifts of galaxies under the assumption that a larger redshift implies a greater distance in accordance with Hubble's law. A number of investigators have carried out such a program, some in fairly restricted areas of the sky and others over larger regions but to shallower depths. A primary finding of such surveys is the existence of huge holes and voids, regions of space measuring hundreds of millions of light-years across where galaxies seem notably deficient or even totally absent. The presence of holes and voids forms in some sense, of course, a natural complement to the idea of superclusters, but the surprising finding is the degree of the density contrast between the large-scale regions where galaxies are found and those where they are not. There was even a report in 1986 that clusters of galaxies may occur only on the interfaces between voids (Figure 4). This picture may present considerable difficulties for theories that envisage the clustering of galaxies in the universe to arise solely from the action of gravitational forces.

Immense holes and voids in space

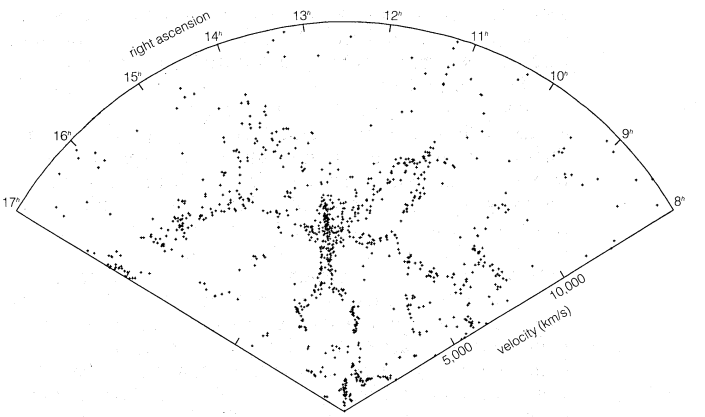



Figure 4: A two-dimensional representation of a three-dimensional map of galaxies distributed on the surfaces of what are thought to be enormous bubble-like voids measuring approximately 60,000,000 to 150,000,000 light-years across. The pie-shaped segment of the sky shown here (6° wide and 117° long) contains about 1,000 galaxies, all located within 300,000,000 light-years from Earth. The Coma cluster of galaxies, lying near the middle of the map and resembling a human torso, seems to occur where several voids intersect.

Margaret Geller, John Huchra, and Valerie de Lapparent, Smithsonian Astrophysical Observatory

## GRAVITATIONAL THEORIES OF CLUSTERING

The fact that gravitation universally affects all masses suggests that its attraction may explain why the astronomical universe is not uniform but contains structure. This natural idea, which is the basis of much of the modern theoretical work on the problem, had already occurred to Newton in 1692. Newton wrote to the noted scholar and clergyman Richard Bentley:

> It seems to me that, if the matter of our sun and planets, and all the matter of the universe, were evenly scattered throughout all the heavens, and every particle had an innate gravity towards all the rest, and the whole space throughout which this matter was scattered, was finite, the matter on the outside of this space would by its gravity tend towards all the matter on the inside, and by consequence fall down into the middle of the whole space, and there compose one great spherical mass. But if the matter were evenly disposed throughout infinite space, it could never convene into one mass; but some of it would convene into one mass and some into another, so as to make an infinite number of great masses, scattered at great distances from one to another throughout all that infinite space. And thus might the sun and the fixed stars be formed, supposing the matter were of a lucid nature.

**Modes of gravitational instability.** It was the English physicist and mathematician Sir James Jeans who, in 1902, first provided a quantitative criterion for the gravitational instability speculated upon by Newton. Jeans considered the idealized initial state of a homogeneous static gas of infinite extent and uniform temperature and asked under what conditions the compressed portions of a small sinusoidal fluctuation would continue to contract gravitationally and become denser and denser (eventually to form galaxies and stars presumably) rather than re-expand because of the increased internal pressure. He found that for gravitational instability to occur the wavelength of the density fluctuation had to exceed a certain critical value, now called the Jeans length, which is proportional to the square root of the ratio of temperature to density.

Two new considerations enter to modify the picture in a universe that begins with a hot big bang: the expansion of the background and the coexistence with matter of a thermal radiation field. The expansion of the background causes the dense portions of unstable small fluctuations to grow much more slowly, at least at first, than the static Jeans theory—as a power of time rather than as an exponential. The thermal radiation field causes greater complications.

First, the existence of a component in the universe other than ordinary matter, radiation, means that one has to specify—particularly in the early stages of the expansion when the energy density of radiation dominates that of matter—whether the radiation field fluctuates together with matter or whether it maintains a uniform level inside which matter fluctuates. Density fluctuations of the first type are called adiabatic perturbations and those of the second type isothermal perturbations (because the temperature of the radiation field remains uniform in space and the matter temperature locally equals that of the radiation when they are well coupled).

Adiabatic and isothermal perturbations

In the early universe when the radiation temperature was high and matter existed as a highly ionized plasma, nei-

ther adiabatic nor isothermal fluctuations could grow because the intense radiation field resisted compression and, through its strong coupling to ionized matter, prevented the latter also from contracting relative to the overall expansion of the universe. Indeed, the tendency for the excess radiation in the compressed regions of adiabatic fluctuations to try to diffuse out of such regions implies that such fluctuations tend to decay. Therefore, given an arbitrary initial spectrum of adiabatic fluctuations, only those with a large enough scale can survive the decay for the age of the universe up to that point.

Decoupling between ordinary matter and radiation occurs when the temperature drops low enough for free (hydrogen) ions and electrons to recombine. When electrons become attached to atoms, they have a much smaller cross section for interaction with photons than when they were free. This occurs for reasonable cosmological models at a temperature of about 4,000 K. At this time, by coincidence (but perhaps ultimately one of great physical significance), the radiation-energy density also begins to drop below the rest-energy density of matter, and the universe turns from being radiation dominated to being matter dominated. Past the decoupling epoch, the density fluctuations of the type previously labeled isothermal can grow if they satisfy the original Jeans criterion, whereas those previously labeled adiabatic can grow only if they have survived the prior epoch of damping. Calculations indicate that the smallest unstable fragment of the former type has a mass comparable to that of a globular cluster. In the case of the latter type, the fragment has a mass comparable to that of a giant galaxy or of a large cluster of galaxies, depending on various assumptions.

Possible forms of the hypothesized hidden mass

Among these assumptions is the choice of the form for the hidden mass. If the hidden mass is not ordinary matter, as current arguments seem to indicate, but is contained in exotic forms of elementary particles whose properties have yet to be deciphered, then one needs to specify if and when this hidden mass decouples from the thermal radiation field. Two extremes are often considered: "warm" dark matter and "cold" dark matter. Warm dark matter is typified by such hypothetical particles as neutrinos that have small but nonzero rest mass, which decouple relatively early from the radiation field. Particles of this sort stream freely (nearly at the speed of light in the early universe) and erase initial fluctuations of all scale smaller than a critical coherence length (analogous to but larger than the critical scale introduced by photons for adiabatic fluctuations), above which self-gravity can finally cause growth (when the neutrinos are moving much less rapidly). Cold dark matter is typified by particles that interact only weakly with radiation and ordinary matter and that have sufficient rest mass so as always to possess random thermal motions much less than the speed of light at any stage relevant to the problem of galaxy formation. Density fluctuations of such particles can grow in a fashion similar to that described for isothermal fluctuations of ordinary matter after decoupling; hence, on the scale of galaxies and larger groups, cold dark matter possesses no coherence length. In either picture, warm or cold, the dark component of the universe supposedly forms a lumpy background into whose concentrations ordinary matter falls eventually to produce galaxies and stars.

Origin of large-scale cosmic structures

**Top-down and bottom-up theories.** The scenarios described in the previous section lead, in the extremes, to two different pictures for the origin of large-scale structure in the universe, which can be given the labels "top-down" and "bottom-up." In top-down theories the regions with the largest scale sizes, comparable to superclusters and clusters, collapse first, yielding flat gaseous "pancakes" of ordinary matter (a description coined by the primary proponent of this theory, Yakov B. Zel'dovich of the Soviet Union) from which galaxies condense. In bottom-up theories the regions with the smallest scale sizes, comparable to galaxies or smaller, form first, giving rise to freely moving entities that subsequently aggregate gravitationally (perhaps by a hierarchial process) to produce clusters and superclusters of galaxies. Adiabatic fluctuations of ordinary matter tend to yield a top-down picture, and isothermal fluctuations a bottom-up picture. When hidden mass is added to the calculations, warm dark matter tends to give the top-down picture, and cold dark matter the bottom-up picture.

To make comparisons with observational data, the amplitudes of the initial fluctuations are required as input to numerical simulations on a computer to follow the subsequent growth of structure. In principle, this data should come from a first principles calculation of the (quantum) origin of the big bang; in practice, workers use heuristic arguments, along with constraints set by the relative lack of anisotropy of the cosmic microwave background except for the dipole component. When this is done and models are computed, it is found that top-down theories tend to give a better, though not perfect, account of the observed spatial distributions of galaxies (*e.g.,* flattened superclusters and holes and voids), but cluster formation and galaxy formation take place too late (*i.e.,* too recently relative to the present epoch) to be compatible with the observational data. Moreover, neutrinos, with their large coherence length, probably cannot explain the hidden mass that is inferred to reside in the dark halos of individual galaxies.

Bottom-up theories that include cold dark matter as a component can yield acceptable accounts of the spatial distributions of galaxies but only at the expense of introducing larger random velocities than are observed. A possible escape from this conclusion lies in the suggestion that the distribution of galaxies (consisting mostly of ordinary matter) may not trace the distribution of mass (consisting mostly of cold dark matter). This scheme, called biased galaxy formation, may have a physical basis if it can be argued that galaxies only form from larger than average fluctuations. The argument is not implausible because star formation works that way; stars only form from the densest portions (the cores) of a molecular cloud and not just anywhere. Acceptance of the analogous argument for the problem of galaxy formation, however, would imply that there should be an appreciable intergalactic medium composed of gas that did not condense into galaxies, just as there is an appreciable interstellar medium composed of gas that did not condense into stars. Without invoking additional ad hoc mechanisms (*e.g.,* reheating the gas), it is not clear that such an intergalactic medium would have resisted detection for so long.

## UNORTHODOX THEORIES OF CLUSTERING AND GALAXY FORMATION

Given the somewhat unsatisfactory state of affairs with gravitational theories for the origin of large-scale structure in the universe, some cosmologists have abandoned the orthodox approach altogether and have sought alternative mechanisms. One of the first to be considered was primordial turbulence. This idea enjoys little current favour for a variety of reasons, the most severe being the following. Because it tends to decay over time, turbulence of a magnitude sufficient to cause galaxy formation after decoupling would have had to be much larger during earlier epochs. This seems both unlikely and unnatural. Too delicate a balance is required for primordial turbulence to produce galaxies rather than, say, black holes.

Another alternative suggestion is that energetic galactic explosions resulting from the formation of a first wave of massive stars may have compressed large shells of intergalactic gas that subsequently became the sites for further galaxy formation and more explosions. Such a picture is attractive because it predicts holes and voids with galaxies at the interfaces (though the holes tend to be much too small), but it does not avoid the criticism that a "seed" galaxy needs to be formed at the centre of each shell by some other process. If such a process exists, why should it not be the dominant mechanism?

The notion of cosmic strings

Finally, there is a suggestion that galaxy and cluster formation might take place by accretion around "cosmic strings." These long strands or loops of mass-energy are a consequence of certain theories of elementary particle physics. The strings are envisaged to arise from phase transitions in the very early universe in a fashion analogous to the way faults occur in a crystal that suffers dislocations because of imperfect growth from, say, a liquid medium. The dynamic properties of cosmic strings

are imperfectly understood, but some heuristic arguments suggest that they may give a clustering hierarchy similar to that observed for galaxies. Unfortunately the same particle physics that produces cosmic strings also produces magnetic monopoles (isolated magnetic charges), whose possible abundance in the universe can be shown by experiments to lie below very low limits. Particle physicists like to explain the absence of magnetic monopoles in the Cosmos by invoking for the very early universe the mechanism of inflation (see below). The same mechanism would also inflate away cosmic strings.

In summary, it can be seen that mechanisms alternative to the growth of small initial fluctuations by self-gravitation all have their own difficulties. Most astronomers hope some dramatic new observation or new idea may yet save the gravitational instability approach, whose strongest appeal has always been the intuitive notion that the force which dominates the astronomical universe, gravity, will automatically promote the growth of irregularities. But until a complete demonstration is provided, the lack of a simple convincing picture of how galaxies form and cluster will remain one of the prime failings of the otherwise spectacularly successful hot big bang theory.

### THE EXTRAGALACTIC DISTANCE SCALE AND HUBBLE'S CONSTANT

It was noted earlier that the galaxies in the Virgo cluster had an average recession velocity $v$ (as measured by their redshift) of roughly 1,000 km/sec with respect to the Local Group. If the distance $r$ to the Virgo cluster is 50,000,000 light-years and if the Virgo galaxies can be assumed to be far enough away to partake in the general Hubble flow, then the application of the Hubble law, $v = H_0 r$, yields Hubble's constant as $H_0 = 20$ km/sec per 1,000,000 light-years. The reciprocal of Hubble's constant is called the Hubble time; for the value given above, $H_0^{-1} = 15{,}000{,}000{,}000$ years.

The Hubble time

The most naive interpretation for the Hubble time is that of a free expansion of the universe, wherein a Hubble-time ago the distant galaxies started receding from one another, reaching a distance $r = vH_0^{-1}$ in time $H_0^{-1}$ if they fly away at speed $v$, the fastest receding galaxies getting the farthest away. Rearranging terms yields the Hubble law $v = H_0 r$. The interpretation is naive in two respects: (1) it ignores the role of gravitation in slowing down the expansion, so that Hubble's "constant" does not always have the value it does at the present epoch; and (2) it overlooks the part played by gravitation in regulating the global structure of space-time, so that the interpretation of the "velocity" $v$ and "distance" $r$ is modified when distances or redshifts approach values such that $v$ given by the above formula becomes comparable to or exceeds the speed of light. Nevertheless, as will be seen in the discussion of relativistic cosmologies below, the Hubble time does provide a useful rough estimate for the age of the universe.

The exact value of Hubble's constant is an issue of great controversy among astronomers. Modern estimates for $H_0$ range from 15 to 30 km/sec per 1,000,000 light-years. The source of the discrepancy lies partly in the interpretation of the amount of distortion superimposed atop a pure Hubble flow by the gravitational effects of the Local Supercluster in which the Local Group and the Virgo cluster are embedded, and partly in the different calibrators used or emphasized by different workers for the distances to various extragalactic objects.

To avoid the first complication, the interpretation of the velocity field in the Local Supercluster, it is possible to examine the redshift-distance relation implied by Sandage's and Tammann's study of 50 Sc I galaxies. There is little controversy that these distant galaxies do empirically satisfy the idealized linear relationship of the Hubble law. The faintest galaxies in the sample have recession velocities of 9,000 km/sec, and if they lie at the calibrated distance of 600,000,000 light-years, then $H_0 = 15$ km/sec per 1,000,000 light-years, the same value as Sandage and Tammann derived from their study of the Virgo cluster. Unfortunately many workers do not accept the determination of Sandage and Tammann of the distances to the nearest Sc I galaxies (in particular, M101). They regard as suspect the technique using the sizes of H II regions as a distance indicator. Some astronomers advocate, for example, using the relationship found to exist between the luminosity $L$ of a spiral galaxy and the velocity $V$ of its (flat) rotation curve, $L$ proportional to $V^4$, as a basis for measuring extragalactic distances, and they obtain values for $H_0$ that lie on the high end of the range cited above. For the sake of definiteness, this article adopts the fiducial value $H_0 = 20$ km/sec per 1,000,000 light-years, but it should be noted that uncertainties of the magnitude discussed still remain. The corollary of this warning is that the distances quoted for extragalactic objects also are uncertain by the same factor.

## Cosmological models

### EARLY COSMOLOGICAL IDEAS

Immediate issues that arise when anyone contemplates the universe at large are whether space and time are infinite or finite. After many centuries of thought by some of the best minds, humanity has still not arrived at conclusive answers to these questions. Aristotle's answer was that the material universe must be spatially finite, for if stars extended to infinity, they could not perform a complete rotation around the Earth in 24 hours. Space must then itself also be finite because it is merely a receptacle for material bodies. On the other hand, the heavens must be temporally infinite without beginning or end, since they are imperishable and cannot be created or destroyed.

Except for the infinity of time, these views came to be accepted religious teachings in Europe before the period of modern science. The most notable person to publicly express doubts about restricted space was the Italian philosopher-mathematician Giordano Bruno, who asked the obvious question that if there was a boundary or edge to space, what was on the other side? For his advocacy of an infinity of suns and earths, he was burned at the stake in 1600.

In 1610 Kepler provided a profound reason for believing that the number of stars in the universe had to be finite. If there were an infinity of stars, he argued, then the sky would be completely filled with them and night would not be dark! This point was rediscussed by the astronomers Edmond Halley and Jean-Philippe-Loys de Cheseaux in the 18th century, but it was not popularized as a paradox until Heinrich Wilhelm Olbers of Germany took up the problem in the 19th century. The difficulty became potentially very real with Hubble's measurement of the enormous extent of the universe of galaxies with its large-scale homogeneity and isotropy. His discovery of the systematic recession of the galaxies provided an escape, however. At first, people thought that the redshift effect alone would suffice to explain why the sky is dark at night—namely, that the light from the stars in distant galaxies would be redshifted to long wavelengths beyond the visible regime. The modern consensus is, however, that a finite age for the universe is a far more important effect. Even if the universe is spatially infinite, photons from very distant galaxies simply do not have the time to travel to the Earth because of the finite speed of light. There is a spherical surface, the cosmic event horizon (roughly $10^{10}$ light-years in radial distance from the Earth at the current epoch), beyond which nothing can be seen even in principle; and the number (roughly $10^{10}$) of galaxies within this cosmic horizon, the observable universe, are too few to make the night sky bright.

When one looks to great distances, one is of course seeing things as they were a long time ago, again because light takes a finite time to travel to Earth. Over such great spans, do the classical notions of Euclid concerning the properties of space necessarily continue to hold? The answer given by Einstein was: No, the gravitation of the mass contained in cosmologically large regions may warp one's usual perceptions of space and time: in particular, the Euclidean postulate that parallel lines never cross need not be a correct description of the geometry of the actual universe. And in 1917 Einstein presented a mathematical model of the universe in which the total volume of space was finite yet had no boundary or edge. The

model was based on his theory of general relativity that utilized a more generalized approach to geometry devised in the 19th century by the German mathematician Bernhard Riemann.

## GRAVITATION AND THE GEOMETRY OF SPACE-TIME

The physical foundation of Einstein's view of gravitation, general relativity, lies on two empirical findings that he elevated to the status of basic postulates. The first postulate is the relativity principle: local physics is governed by the theory of special relativity. The second postulate is the equivalence principle: there is no way for an observer to distinguish locally between gravity and acceleration. The motivation for the second postulate comes from Galileo's observation that all objects—independent of mass, shape, colour, or any other property—accelerate at the same rate in a (uniform) gravitational field.

The principles of relativity and equivalence

Einstein's theory of special relativity had as its basic premises (1) the notion (also dating back to Galileo) that the laws of physics are the same for all inertial observers and (2) the constancy of the speed of light in a vacuum—namely, that the speed of light has the same value ($3 \times 10^{10}$ cm/sec) for all inertial observers independent of their motion relative to the source of the light. Clearly, this second premise is incompatible with Euclidean and Newtonian precepts of absolute space and absolute time, resulting in a program that merged space and time into a single structure, with well-known consequences. The space-time structure of special relativity is often called "flat" because, among other things, the propagation of photons is easily represented on a flat sheet of graph paper with equal-sized squares. Let each tick on the vertical axis represent one light-year ($9.46 \times 10^{17}$ cm) of distance in the direction of the flight of the photon, and each tick on the horizontal axis represent the passage of one year ($3.16 \times 10^{7}$ sec) of time. The propagation path of the photon is then a 45° line because it flies one light-year in one year (with respect to the space and time measurements of all inertial observers no matter how fast they move relative to the photon).

Warping of space-time by gravitation

The principle of equivalence in general relativity allows the locally flat space-time structure of special relativity to be warped by gravitation, so that (in the cosmological case) the propagation of the photon over thousands of millions of light-years can no longer be plotted on a globally flat sheet of paper. To be sure, the curvature of the paper may not be apparent when only a small piece is examined, thereby giving the local impression that space-time is flat (*i.e.*, satisfies special relativity). It is only when the graph paper is examined globally that one realizes it is curved (*i.e.*, satisfies general relativity).

Fourth spatial dimension

In Einstein's 1917 model of the universe, the curvature occurs only in space, with the graph paper being rolled up into a cylinder on its side, a loop around the cylinder at constant time having a circumference of $2\pi R$—the total spatial extent of the universe. Notice that the "radius of the universe" is measured in a "direction" perpendicular to the space-time surface of the graph paper. Since the ringed space axis corresponds to one of three dimensions of the actual world (any will do since all directions are equivalent in an isotropic model), the radius of the universe exists in a fourth spatial dimension (not time) which is not part of the real world. This fourth spatial dimension is a mathematical artifice introduced to represent diagrammatically the solution (in this case) of equations for curved three-dimensional space that need not refer to any dimensions other than the three physical ones. Photons traveling in a straight line in any physical direction have trajectories that go diagonally (at 45° angles to the space and time axes) from corner to corner of each little square cell of the space-time grid; thus, they describe helical paths on the cylindrical surface of the graph paper, making one turn after traveling a spatial distance $2\pi R$. In other words, always flying dead ahead, photons would return to where they started from after going a finite distance without ever coming to an edge or boundary. The distance to the "other side" of the universe is therefore $\pi R$, and it would lie in any and every direction; space would be closed on itself.

Now, except by analogy with the closed two-dimensional surface of a sphere that is uniformly curved toward a centre in a third dimension lying nowhere on the two-dimensional surface, no three-dimensional creature can visualize a closed three-dimensional volume that is uniformly curved toward a centre in a fourth dimension lying nowhere in the three-dimensional volume. Nevertheless, three-dimensional creatures could discover the curvature of their three-dimensional world by carrying out surveying experiments of sufficient spatial scope. They could draw circles, for example, by tacking down one end of a string and tracing along a single plane the locus described by the other end when the string is always kept taut in between (a straight line) and walked around by a surveyor. In Einstein's universe, if the string were short compared to the quantity $R$, the circumference of the circle divided by the length of the string (the radius of the circle) would nearly equal $2\pi = 6.2837853\ldots$, thereby fooling the three-dimensional creatures into thinking that Euclidean geometry gives a correct description of their world. Yet, the ratio of circumference to length of string would become less than $2\pi$ when the length of string became comparable to $R$. Indeed, if a string of length $\pi R$ could be pulled taut to the antipode of a positively curved universe, the ratio would go to zero. In short, at the tacked-down end the string could be seen to sweep out a great arc in the sky from horizon to horizon and back again; but, to make the string do this, the surveyor at the other end need only walk around a circle of vanishingly small circumference.

To understand why gravitation can curve space (or more generally, space-time) in such startling ways, consider the following thought experiment that was originally conceived by Einstein. Imagine an elevator in free space accelerating upward, from the viewpoint of a woman in inertial space, at a rate numerically equal to $g$, the gravitational field at the surface of the Earth. Let this elevator have parallel windows on two sides, and let the woman shine a brief pulse of light toward the windows. She will see the photons enter close to the top of the near window and exit near the bottom of the far window because the elevator has accelerated upward in the interval it takes light to travel across the elevator. For her, photons travel in a straight line, and it is merely the acceleration of the elevator that has caused the windows and floor of the elevator to curve up to the flight path of the photons.

Let there now be a man standing inside the elevator. Because the floor of the elevator accelerates him upward at a rate $g$, he may—if he chooses to regard himself as stationary—think that he is standing still on the surface of the Earth and is being pulled to the ground by its gravitational field $g$. Indeed, in accordance with the equivalence principle, without looking out the windows (the outside is not part of his local environment), he cannot perform any local experiment that would inform him otherwise. Let the woman shine her pulse of light. The man sees, just like the woman, that the photons enter near the top edge of one window and exit near the bottom of the other. And just like the woman, he knows that photons propagate in straight lines in free space. (By the relativity principle, they must agree on the laws of physics if they are both inertial observers.) However, since he actually sees the photons follow a curved path relative to himself, he concludes that they must be bent by the force of gravity. The woman tries to tell him there is no such force at work; he is not an inertial observer. Nonetheless, he has the solidity of the Earth beneath him, so he insists on attributing his acceleration to the force of gravity. According to Einstein, they are both right. There is no need to distinguish locally between acceleration and gravity—the two are in some sense equivalent. But if that is the case, then it must be true that gravity—"real" gravity—can actually bend light. And indeed it can, as many experiments have shown since Einstein's first discussion of the phenomenon.

It was the genius of Einstein to go even further. Rather than speak of the force of gravitation having bent the photons into a curved path, might it not be more fruitful to think of photons as always flying in straight lines—in the sense that a straight line is the shortest distance between two points—and that what really happens is that gravitation bends space-time? In other words, perhaps gravitation is curved space-time, and photons fly along the shortest

paths possible in this curved space-time so as to give the appearance of being bent by a "force" when one insists on thinking that space-time is flat. The utility of taking this approach is that it becomes automatic that all test bodies fall at the same rate under the "force" of gravitation, for they are merely producing their natural trajectories in a background space-time that is curved in a certain fashion independent of the test bodies. What was a minor miracle for Galileo and Newton becomes the most natural thing in the world for Einstein.

To complete the program and to conform with Newton's theory of gravitation in the limit of weak curvature (weak field), the source of space-time curvature would have to be ascribed to mass (and energy). The mathematical expression of these ideas constitutes Einstein's theory of general relativity, one of the most beautiful artifacts of pure thought ever produced. The American physicist John Archibald Wheeler and his colleagues summarized Einstein's view of the universe in these terms:

> Curved spacetime tells mass-energy how to move;
> mass-energy tells spacetime how to curve.

Contrast this with Newton's view of the mechanics of the heavens:

> Force tells mass how to accelerate;
> mass tells gravity how to exert force.

Notice therefore that Einstein's worldview is not merely a quantitative modification of Newton's picture (which is also possible via an equivalent route using the methods of quantum field theory) but represents a qualitative change of perspective. Modern experiments have amply justified the fruitfulness of Einstein's alternative interpretation of gravitation as geometry rather than as force. His theory would have undoubtedly delighted the Greeks.

## RELATIVISTIC COSMOLOGIES

**Einstein's model.** To derive his 1917 cosmological model, Einstein made three assumptions that lay outside the scope of his equations. The first was to suppose that the universe is homogeneous and isotropic in the large (*i.e.*, the same everywhere on average at any instant in time), an assumption that the English astrophysicist Edward A. Milne later elevated to an entire philosophical outlook by naming it the cosmological principle. Given the success of the Copernican revolution, this outlook is a natural one. Newton himself had it implicitly in mind in his letter to Bentley (see above) when he took the initial state of the Cosmos to be everywhere the same before it developed "the sun and the fixed stars."

The second assumption was to suppose that this homogeneous and isotropic universe had a closed spatial geometry. As described in the previous section, the total volume of a three-dimensional space with uniform positive curvature would be finite but possess no edges or boundaries (to be consistent with the first assumption).

The third assumption made by Einstein was that the universe as a whole is static—*i.e.*, its large-scale properties do not vary with time. This assumption, made before Hubble's observational discovery of the expansion of the universe, was also natural; it was the simplest approach, as Aristotle had discovered, if one wishes to avoid a discussion of a creation event. Indeed, the philosophical attraction of the notion that the universe on average is not only homogeneous and isotropic in space but also constant in time was so appealing that a school of British cosmologists—Hermann Bondi, Fred Hoyle, and Thomas Gold—would call it the perfect cosmological principle and carry its implications in the 1950s to ultimate refinement in the so-called steady state model.

The cosmological constant

To his great chagrin Einstein found, in 1917, that with his three adopted assumptions, his equations of general relativity—as originally written down—had no meaningful solutions. To obtain a solution, Einstein realized that he had to add to his equations an extra term, which came to be called the cosmological constant. If one speaks in Newtonian terms, the cosmological constant could be interpreted as a repulsive force of unknown origin that could exactly balance the attraction of gravitation of all the matter in Einstein's closed universe and keep it from moving. The inclusion of such a term in a more general context, however, meant that the universe in the absence of any mass-energy (*i.e.*, consisting of a vacuum) would not have a space-time structure that was flat (*i.e.*, would not have satisfied the dictates of special relativity exactly). Einstein was prepared to make such a sacrifice only very reluctantly, and when he later learned of Hubble's discovery of the expansion of the universe and realized that he could have predicted it had he only had more faith in the original form of his equations, he regretted the introduction of the cosmological constant as the "biggest blunder of his life." Ironically, recent theoretical developments in particle physics suggest that in the early universe there may very well have been a nonzero value to the cosmological constant and that this value may be intimately connected with precisely the nature of the vacuum state (see below).

**De Sitter's model.** It was also in 1917 that the Dutch astronomer Willem de Sitter recognized that he could obtain a static cosmological model differing from Einstein's simply by removing all matter. The solution remains stationary essentially because there is no matter to move about in his postulated universe. If some test particles are reintroduced into the model, the cosmological term would propel them away from each other. Astronomers now began to wonder if this effect might not underlie the recession of the spirals.

**Friedmann–Lemaître models.** In 1922 Aleksandr A. Friedmann, a Russian meteorologist and mathematician, and in 1927 Georges Lemaître, the aforementioned Belgian cleric, independently discovered solutions related to Einstein's equations that included realistic amounts of matter. These evolutionary models correspond to big bang cosmologies. Friedmann and Lemaître adopted Einstein's assumption of spatial homogeneity and isotropy (the cosmological principle). They rejected, however, his assumption of time independence and considered both positively curved spaces ("closed" universes) as well as negatively curved spaces ("open" universes). The difference between the approaches of Friedmann and Lemaître is that the former set the cosmological constant equal to zero, whereas the latter retained the possibility that it might have a nonzero value. To simplify the discussion, only the Friedmann models are considered here.

Divergence from Einstein

The decision to abandon a static model meant that the Friedmann models evolve with time. As such, neighbouring pieces of matter have recessional (or contractional) phases when they separate from (or approach) one another with an apparent velocity that increases linearly with increasing distance. Friedmann's models thus anticipated the Hubble law before it had been formulated on an observational basis. It was Lemaître, however, who had the good fortune of deriving the results at the time when the recession of the galaxies was being recognized as a fundamental cosmological observation, and it was he who clarified the theoretical basis for the phenomenon.

The geometry of space in Friedmann's closed models is similar to that of Einstein's original model, but there is a curvature to time as well as one to space. Unlike Einstein's model, where time runs eternally at each spatial point on an uninterrupted horizontal line that extends infinitely into the past and future, there is a beginning and end to time in Friedmann's version of a closed universe when material expands from or is recompressed to infinite densities. These instants are called the instants of the "big bang" and the "big squeeze," respectively. The global space-time diagram for the middle half of the expansion-compression phases can be depicted as a barrel lying on its side (Figure 5). The space axis corresponds again to any one direction in the universe and it wraps around the barrel. Through each spatial point runs a time axis that extends along the length of the barrel on its (space-time) surface. Because the barrel is curved in both space and time, the little squares in the grid of the curved sheet of graph paper marking the space-time surface are of nonuniform size, stretching to become bigger when the barrel broadens (universe expands) and shrinking to become smaller when the barrel narrows (universe contracts).

Expansion-compression phases of the universe

It should be remembered that only the surface of the barrel has physical significance. The dimension off the

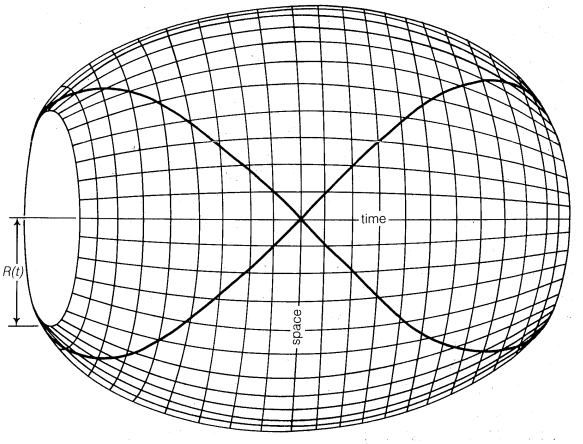



Figure 5: Curved space-time in a matter dominated, closed universe during the middle half of its expansion-compression phases. At each instant of time $t$, the space axis forms a closed loop with radius $R(t)$, the so-called radius of the universe, in an unobservable fourth dimension (see text).
From F.H. Shu, *The Physical Universe* (1982); University Science Books

surface toward the axle of the barrel represents the fourth spatial dimension, which is not part of the real three-dimensional world. The space axis circles the barrel and closes upon itself after traversing a circumference equal to $2\pi R$, where $R$, the radius of the universe (in the fourth dimension), is now a function of the time $t$. In a closed Friedmann model $R$ starts equal to zero at time $t = 0$ (not shown in barrel diagram), expands to a maximum value at time $t = t_m$ (the middle of the barrel), and recontracts to zero (not shown) at time $t = 2t_m$, with the value of $t_m$ dependent on the total amount of mass that exists in the universe.

Imagine now that galaxies reside on equally spaced tick marks along the space axis. Each galaxy on average does not move spatially with respect to its tick mark in the spatial (ringed) direction but is carried forward horizontally by the march of time. The total number of galaxies on the spatial ring is conserved as time changes, and therefore their average spacing increases or decreases as the total circumference $2\pi R$ on the ring increases or decreases (during the expansion or contraction phases). Thus, without in a sense actually moving in the spatial direction, galaxies can be carried apart by the expansion of space itself. From this point of view, the recession of galaxies is not a "velocity" in the usual sense of the word. For example, in a closed Friedmann model, there could be galaxies that started, when $R$ was small, very close to the Milky Way system on the opposite side of the universe. Now, $10^{10}$ years later, they are still on the opposite side of the universe but at a distance much greater than $10^{10}$ light-years away. They reached those distances without ever having had to move (relative to any local observer) at speeds faster than light—indeed, in a sense, without having had to move at all. The separation rate of nearby galaxies can be thought of as a velocity without confusion in the sense of Hubble's law, if one wants, but only if the inferred velocity is much less than the speed of light.

On the other hand, if the recession of the galaxies is not viewed in terms of a velocity, then the cosmological redshift cannot be viewed as a Doppler shift. How, then, does it arise? The answer is contained in the barrel diagram by noticing that as the universe expands, each small cell in the space-time grid also expands. Consider the propagation of a wave train of electromagnetic radiation consisting at a given instant of time of $n$ wavelengths with its head at one vertex and its tail at another vertex (for simplicity of discussion) one spatial cell length back (Figure 6). Suppose an elliptical galaxy emits such a wave at some time $t_1$. The head of the wave propagates from corner to corner on the little square grids that look locally flat, and the tail propagates from corner to corner one vertex back. At a later time $t_2$, a spiral galaxy begins to intercept the head of the wave. At time $t_2$, the tail is still one vertex back, and therefore the wave train, still containing $n$ wavelengths, now spans one current spatial grid spacing. In other words, independent of $n$, the wavelength has grown in direct proportion to the linear expansion factor of the universe. Consequently, all electromagnetic radiation from a given object will show the same cosmological redshift if the universe (or, equivalently, the average spacing between galaxies) was smaller at the epoch of transmission than at the epoch of reception. The wavelength will have been stretched in direct proportion to the expansion of the universe in between.

A nonzero peculiar velocity for an emitting galaxy with respect to its local cosmological frame can be taken into account by Doppler-shifting the emitted photons before applying the cosmological redshift factor; *i.e.*, the observed redshift would be a product of two factors. When the observed redshift is large, one usually assumes that the dominant contribution is of cosmological origin. When this assumption is valid, the redshift is a monotonic function of both distance and time during the expansional phase of any cosmological model. Thus, astronomers often use the redshift $z$ as a shorthand indicator of both distance and elapsed time. Following from this, the statement "object $X$ lies at $z = a$" means that "object $X$ lies at a distance associated with redshift $a$;" the statement "event $Y$ occurred at redshift $z = b$" means that "event $Y$ occurred a time ago associated with redshift $b$."

Redshift as a monotonic function of distance and time

The open Friedmann models differ from the closed models in both spatial and temporal behaviour. In an open universe the total volume of space and the number of galaxies contained in it are infinite. The three-dimensional spatial geometry is one of uniform negative curvature in the sense that if circles are drawn with very large lengths of string, the ratio of circumferences to lengths of string are greater than $2\pi$. The temporal history begins again with expansion from a big bang of infinite density, but now the expansion continues indefinitely, and the average density of matter and radiation in the universe would eventually become vanishingly small. Time in such a model has a beginning but no end.

From F.H. Shu, *The Physical Universe* (1982); University Science Books

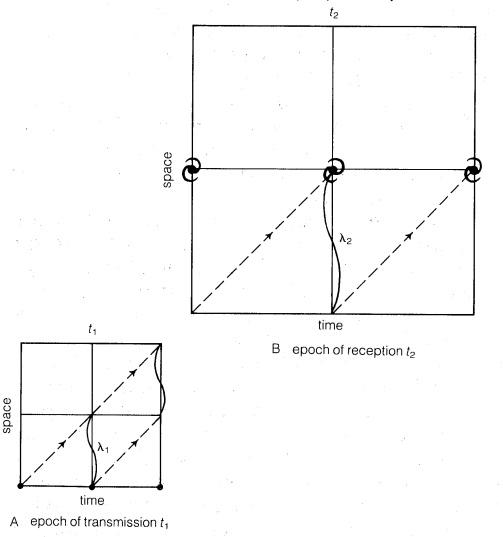



Figure 6: The cosmological redshift as a stretching of the wavelengths of photons. The cells in A and B correspond to the squares in the space-time grid of Figure 5. Each individual cell becomes larger, paralleling the linear expansion of the universe from the epoch of transmission $t_1$ to the epoch of reception $t_2$ (see text).

**The Einstein–de Sitter universe.** In 1932 Einstein and de Sitter proposed that the cosmological constant should be set equal to zero and derived a homogeneous and isotropic model that provides the separating case between the closed and open Friedmann models; *i.e.*, Einstein and de Sitter assumed that the spatial curvature of the universe is neither positive nor negative but rather zero. The spatial geometry of the Einstein–de Sitter universe is Euclidean (infinite total volume), but space-time is not globally flat (*i.e.*, not exactly the space-time of special relativity). Time

again begins with a big bang and the galaxies recede forever, but the recession rate (Hubble's "constant") asymptotically coasts to zero as time advances to infinity.

Because the geometry of space and the gross evolutionary properties are uniquely defined in the Einstein–de Sitter model, many people with a philosophical bent have long considered it the most fitting candidate to describe the actual universe. During the late 1970s strong theoretical support for this viewpoint came from considerations of particle physics (the model of inflation to be discussed below), and mounting, but as yet undefinitive, support also seems to be gathering from astronomical observations.

**Bound and unbound universes and the closure density.** The different separation behaviours of galaxies at large times in the Friedmann closed and open models and the Einstein–de Sitter model allow a different classification scheme than one based on the global structure of space-time. The alternative way of looking at things is in terms of gravitationally bound and unbound systems: closed models where galaxies initially separate but later come back together again represent bound universes; open models where galaxies continue to separate forever represent unbound universes; the Einstein–de Sitter model where galaxies separate forever but slow to a halt at infinite time represents the critical case.

The advantage of this alternative view is that it focuses attention on local quantities where it is possible to think in the simpler terms of Newtonian physics—attractive forces, etc. In this picture, it is intuitively clear that the feature that should distinguish whether or not gravity is capable of bringing a given expansion rate to a halt depends on the amount of mass (per unit volume) present. This is indeed the case; the Newtonian and relativistic formalisms give the same criterion for the critical, or closure, density (in mass equivalent of matter and radiation) that separate closed or bound universes from open or unbound universes. If Hubble's constant at the present epoch is denoted as $H_0$, then the closure density (corresponding to an Einstein–de Sitter model) equals $3H_0^2/8\pi G$, where $G$ is the universal gravitational constant that appears in both Newton's and Einstein's theories of gravity. If the numerical value of Hubble's constant $H_0$ is 20 km/sec per 1,000,000 light-years, then the closure density equals $8 \times 10^{-30}$ g/cm$^3$, the equivalent of about five hydrogen atoms on average per cubic metre of cosmic space. If the actual cosmic average is greater than this value, then the universe is bound (closed) and, though currently expanding, will end in a crush of unimaginable proportion. If it is less, then the universe is unbound (open) and will expand forever. The result is intuitively plausible since the smaller the mass density, the smaller the role for gravitation, so the more the universe will approach free expansion (assuming that the cosmological constant is zero).

Criterion for closure density

The mass in galaxies observed directly, when averaged over cosmological distances, is estimated to be only a few percent of the amount required to close the universe. The radiation field (most of which is in the cosmic microwave background) contributes negligibly to the total at present. If this were all, the universe would be open and unbound. However, the hidden mass that has been deduced from various dynamical arguments multiplies the known amount by factors of a few to 10 or more as one considers phenomena of ever-increasing scale—from galaxies to superclusters. Thus, the total average mass density has been estimated to be 20–40 percent or more of the closure density, and many investigators would like to believe that new observations and refined estimates will eventually bring this number up to 100 percent of closure.

**The age of the universe.** An indirect method of inferring whether the universe is bound or unbound involves estimates of the age of the universe. The basic idea is as follows. For a given present rate of expansion (*i.e.,* Hubble's constant), it is clear that the deceleration produced by gravitation must act to make the expansion faster in the past and slower in the future. Thus, the age of the universe (in the absence of a cosmological constant) must always be less than the free expansion age, $H_0^{-1}$, which equals 15,000,000,000 years. The bigger the role for gravity, the smaller is the true age compared to the Hubble time $H_0^{-1}$. It can be demonstrated that a matter-dominated Einstein–de Sitter universe has a present age two-thirds of the Hubble time, or 10,000,000,000 years. Consequently, the actual universe (which has been matter-dominated for a long time) is closed if it can be shown to be younger than 10,000,000,000 years, and open if older than that critical value.

As noted previously, estimates of the ages of globular cluster stars and of the radioactive elements, which must be at least as old as the universe itself, give ranges of values that are not incompatible with the critical value of 10,000,000,000 years. The formal estimates for globular cluster ages, however, seem somewhat too old to be compatible with the critical value, and some people have inferred that either the universe is unbound or the cosmological constant is not zero. These sentiments may be premature, since the errors in the determinations are not small. Moreover, until astronomers arrive at a better understanding of the source of the discrepancy concerning predicted and observed solar neutrino emission, it cannot be claimed that the knowledge of all physics relevant to the theory of stellar structure and evolution is completely secure.

**Global observational tests.** Since neither local test of average mass density nor age of the oldest accessible objects in the universe has proved decisive in showing whether the universe is bound or unbound, one might investigate large-scale diagnostics of the global structure of space-time to discriminate between closed and open universes. Conducting surveying experiments by means of space exploration of the scope described earlier is of course out of the question. Fortunately there exist in the universe accessible natural probes with which to explore the deepest reaches of space and time—namely, photons from distant galaxies. To be able to use these probes effectively as diagnostics—say, in the apparent-brightness redshift or the angular-size redshift tests of classical observational cosmology—it is important to know the intrinsic properties of the emitting sources and to examine the objects with the largest possible redshifts (so one is going farthest out into space and farthest back in time). Unfortunately these two goals yield incompatible requirements.

Photons from remote galaxies as scientific probes

The problem is that astronomers know the properties of nearby galaxies best; *i.e.,* galaxies as they appear today. The assumption that more distant (and therefore younger) galaxies look the same as they do now becomes more and more suspect as one probes deeper and deeper sources because of the increasing possibility of evolutionary effects (*e.g.,* stellar populations being younger and galaxies not yet having suffered mergers). The difficulty of disentangling the evolutionary effects from the purely cosmological ones remains the biggest obstacle to this line of research. The use of quasars or QSOs fares even worse because, though they are observable at great distances, they have a very large spread in intrinsic luminosities, and they may also suffer from evolutionary effects.

**The ultimate fate of the universe.** In the absence of definitive observational conclusions, one can only speculate on the possible fate of the actual universe. If the universe is unbound, the cosmological expansion will not halt and eventually the galaxies and stars will all die, leaving the Cosmos a cold, dark, and virtually empty place. If the universe is bound, the mass-energy content in the distant but finite future will come together again; the cosmic background radiation will be blueshifted, raising the temperature of matter and radiation to incredible levels, perhaps to reforge everything in the fiery crucible of the big squeeze. Because of the development of structure in previous epochs, the big squeeze may not occur simultaneously everywhere at the end of time as its explosive counterpart, the big bang, seems to have done at the beginning of time. Discussions of recurring cycles of expansions and contractions thus remain highly speculative.

## THE HOT BIG BANG

Given the measured radiation temperature of 2.7 K, it is easy to calculate that the energy density of the cosmic microwave background is about $10^3$ times smaller than the average rest-energy density of ordinary matter in the present universe as matter dominated

universe. Thus, the current universe is matter dominated. If one goes back in time to redshift $z$, the average number densities of particles and photons are both bigger by the same factor $(1 + z)^3$ because the universe was more compressed, and the ratio of number densities would have maintained their current value of about $10^{-9}$—*i.e.*, about one hydrogen nucleus, or proton, for every 1,000,000,000 photons. The wavelength of each photon, however, was shorter by the factor $1 + z$ in the past than it is now; therefore, the energy density of radiation increases faster by one factor of $1 + z$ than the rest-energy density of matter—the dominant form except in the early universe. Thus, the energy density of radiation becomes comparable to the energy density of ordinary matter at a redshift of about $10^3$. At redshifts larger than $10^4$, radiation would have dominated even over the hidden mass of the universe. Between these two values radiation would have decoupled from matter when hydrogen recombined. It is not possible to use photons to observe redshifts larger than about 1,500, because the cosmic plasma at temperatures above 4,000 K is essentially opaque before recombination. One can think of the spherical surface as an inverted "photosphere" of the observable universe. Thus, for example, the earliest stages of the universe's history, when temperatures were $10^9$ K and higher, cannot be examined by light received through any telescope. Clues must be sought by comparing the matter content with theoretical calculations.

For this purpose, fortunately, the cosmological evolution of model universes is especially simple and amenable to computation at redshifts much larger than $10^4$ (or temperatures substantially above $3 \times 10^4$ K) because the physical properties of the dominant component, photons, are then completely known. In a radiation-dominated early universe, for example, the radiation temperature $T$ is very precisely known as a function of the age of the universe, the time $t$ after the big bang.

**Primordial nucleosynthesis.** According to the considerations outlined above, at a time $t$ less than $10^{-4}$ seconds, the creation of matter–antimatter pairs would have been in thermodynamic equilibrium with the ambient radiation field at a temperature $T$ of about $10^{12}$ K. Nevertheless, there was a slight excess of matter particles (*e.g.*, protons) compared to antimatter particles (*e.g.*, antiprotons) of roughly a few parts in $10^9$. This is known because as the universe aged and expanded, the radiation temperature would have dropped and the antiprotons and antineutrons would have annihilated with the protons and neutrons to yield two gamma rays; and later, the antielectrons would have done the same with the electrons to give two more gamma rays. After annihilation, however, the ratio of the number of remaining protons to photons would have been conserved in the subsequent expansion to the present day. Since that ratio is known to be one part in $10^9$, it is easy to work out that the original matter-antimatter asymmetry must have been a few parts per $10^9$.

In any case, after proton–antiproton and neutron–antineutron annihilation but before electron–antielectron annihilation, it is easy to calculate that for every excess neutron there were about five excess protons in thermodynamic equilibrium with one another through neutrino and antineutrino interactions at a temperature of about $10^{10}$ K. When the universe reached an age of a few seconds, the temperature would have dropped significantly below $10^{10}$ K and electron–antielectron annihilation would have occurred, liberating the neutrinos and antineutrinos to stream freely through the universe. With no neutrino–antineutrino reactions to replenish their supply, the neutrons would have started to decay with a half-life of 10.6 minutes to protons and electrons (and antineutrinos). However, before the decay was completed at an age of 1.5 minutes, the temperature would have dropped to $10^9$ K, low enough to allow neutrons to be captured by protons to form a nucleus of heavy hydrogen, or deuterium. (Before that time, the reaction could still have taken place, but the deuterium nucleus would immediately have broken up under the prevailing high temperatures.) Once deuterium had formed, a very fast chain of reactions set in, quickly assembling most of the neutrons and deuterium nuclei with protons to yield helium nuclei. If the decay of neutrons is ignored, an original mix of 10 protons and two neutrons (one neutron for every five protons) would have been assembled into one helium nucleus (two protons plus two neutrons), leaving more than eight protons (eight hydrogen nuclei). This amounts to a helium-mass fraction of $4/12 = 1/3$—*i.e.*, 33 percent. A more sophisticated calculation that takes into account the concurrent decay of neutrons and other complications yields a helium-mass fraction in the neighbourhood of $1/4 = 25$ percent and a hydrogen-mass fraction of $3/4 = 75$ percent, which are close to the deduced primordial values from astronomical observations. This agreement provides one of the primary successes of hot big bang theory.

Formation of the deuterium nucleus

**The deuterium abundance.** Not all of the deuterium formed by the capture of neutrons by protons will be further reacted to produce helium. A small residual can be expected, the exact fraction depending sensitively on the density of ordinary matter existing in the universe when the universe is a few minutes old. The problem can be turned around: given measured values of the deuterium abundance (corrected for various effects), what density of ordinary matter needs to be present at a temperature of $10^9$ K so that the nuclear reaction calculations will reproduce the measured deuterium abundance? The answer is known, and this density of ordinary matter can be expanded by simple scaling relations from a radiation temperature of $10^9$ K to one of 2.7 K. This yields a predicted present density of ordinary matter and can be compared with the density inferred to exist in galaxies when averaged over large regions. The two numbers seem to be within a factor of two or three of each other. In other words, the deuterium calculation implies that a substantial fraction of all of the ordinary matter in the universe (and perhaps all of it) has already been seen in observable galaxies. Ordinary matter cannot be the hidden mass of the universe unless a large mistake has been made somewhere.

## THE VERY EARLY UNIVERSE

**Matter-antimatter asymmetry.** A curious number that appeared in the above discussion was the few parts in $10^9$ asymmetry initially between matter and antimatter (or equivalently, the ratio $10^{-9}$ of protons to photons in the present universe). What is the origin of such a number—so close to zero yet not exactly zero?

At one time, the question posed above would have been considered beyond the ken of physics, because the net "baryon" number (for present purposes, protons and neutrons minus antiprotons and antineutrons) was thought to be a conserved quantity. Therefore, once it exists, it always exists, into the indefinite past and future. Developments in particle physics during the 1970s, however, suggested that the net baryon number may in fact undergo alteration. It is certainly very nearly maintained at the relatively low energies accessible in terrestrial experiments, but it may not have been conserved at the almost arbitrarily high energies with which particles may have been endowed in the very early universe.

An analogy can be made with the chemical elements. In the 19th century most chemists believed the elements to be strictly conserved quantities; although oxygen and hydrogen atoms can be combined to form water molecules, the original oxygen and hydrogen atoms can always be recovered by chemical or physical means. However, with the discovery and elucidation of nuclear forces in the 20th century, chemists came to realize that the elements are conserved if they are subjected only to chemical forces (basically electromagnetic in origin); they can be transmuted by the introduction of nuclear forces, which enter characteristically only when much higher energies per particle are available than in chemical reactions.

Possible transmutations of the net baryon number in the very early universe

In a similar manner, it turns out that at very high energies new forces of nature may enter to transmute the net baryon number. One hint that such a transmutation may be possible lies in the remarkable fact that a proton and an electron appear at first sight to be completely different entities, yet they have, as far as one can tell to very high experimental precision, exactly equal but opposite electric charges. Is this a fantastic coincidence, or does it represent a deep physical connection? A connection would obvi-

ously exist if it can be shown, for example, that a proton is capable of decaying into a positron (an antielectron) plus electrically neutral particles. Should this be possible, then the proton would necessarily have the same charge as the positron, for charge is exactly conserved in all reactions; and the positron in turn would necessarily have the opposite charge of the electron, as it is its antiparticle. Indeed, in some sense, the proton (a baryon) can even be said to be merely the "excited" version of an antielectron (an "antilepton"). As a consequence, evidence for proton decay is an actively pursued goal of experimental particle physics.

With respect to the cosmological problem of the matter-antimatter asymmetry, one theoretical approach is founded on the idea of a grand unified theory (GUT), which seeks to explain the electromagnetic, weak-nuclear, and strong-nuclear forces as a single grand force of nature. This approach indicates that an initial collection of very heavy particles having zero baryon and lepton number may decay into many lighter particles (baryons and leptons) with the desired average for the net baryon number (and net lepton number) of a few parts per $10^9$. This event was supposed to have occurred at a time when the universe was $10^{-39}$ second old. The idea is elegant, but the simplest form of GUT on which such calculations are based has not given the correct predictions.

**Superunification and the Planck era.** Why should a net baryon fraction initially of 0 be more appealing aesthetically than $10^{-9}$? The underlying motivation here is perhaps the most ambitious undertaking in the history of science—the attempt to explain the creation of truly everything from literally nothing. In other words, is the creation of the entire universe from a vacuum possible?

The evidence for such an event lies in another remarkable fact. It can be estimated that the total number of protons in the observable universe is an integer 80 digits long. No one of course knows all 80 digits, but for the argument about to be presented, it suffices only to know that they exist. The total number of electrons in the observable universe is also an integer 80 digits long. In all likelihood, these two integers are equal, digit by digit—if not exactly, then very nearly so. This inference comes from the fact that, as far as astronomers can tell, the total electric charge in the universe is zero. Is this another coincidence, or does it represent a deeper connection?

The apparent coincidence becomes trivial if the entire universe was created from a vacuum because a vacuum has by definition zero electric charge. It is a truism that one cannot get something for nothing. The interesting question is whether one can get everything for nothing. Clearly, this is a very speculative topic for scientific investigation, and the ultimate answer depends on a sophisticated interpretation of what "nothing" means.

The words "nothing," "void," and "vacuum" usually suggest uninteresting empty space. To modern quantum physicists, however, the vacuum has turned out to be rich with complex and unexpected behaviour. They envisage it as a state of minimum energy where quantum fluctuations, consistent with the uncertainty principle of Werner Heisenberg, can lead to the temporary formation of particle–antiparticle pairs. In flat space-time, destruction follows closely upon creation (the pairs are said to be virtual) because there is no source of energy to give the pair permanent existence and all the known forces acting between a particle and antiparticle; they pull the pair together to annihilate one another. In the expanding space-time of the very early universe, however, particles and antiparticles may separate and become part of the observable world. In other words, sharply curved space-time can give rise to the creation of real pairs with positive mass-energy, a fact first demonstrated in the context of black holes by the British astrophysicist Stephen W. Hawking.

Yet Einstein's picture of gravitation is that the curvature of space-time itself is a consequence of mass-energy. Now, if curved space-time is needed to give birth to mass-energy, and mass-energy is needed to give birth to curved space-time, which came first, space-time or mass-energy? The suggestion that they both rose from something still more fundamental raises a new question: What is more fundamental than space-time and mass-energy? No one knows the answer to this question, and perhaps some would argue that the answer is not to be sought within the boundaries of natural science. Nevertheless, attempts are being made, and one area on which intense interest has been focused is the theory of "superstrings."

The theory of superstrings

Superstrings represent one example of a variety of so-called superunification theories formulated to explain the four known forces of nature—gravitational, electromagnetic, weak, and strong—on a single unifying basis. Common to all such schemes are the postulates that quantum principles and special relativity underlie the theoretical framework. Another common feature is supersymmetry, the notion that particles with half-integer values of the spin angular momentum (fermions) can be transformed into particles with integer spins (bosons).

The distinguishing feature of superstring theory is the postulate that elementary particles are not mere points in space but have linear extension. The characteristic linear dimension is a certain combination of three fundamental constants of nature: (1) Planck's constant $h$ (named after Max Planck of Germany, the founder of quantum physics); (2) the speed of light $c$; and (3) the universal gravitational constant $G$. The combination, called the Planck length $(Gh/c^3)^{1/2}$, equals roughly $10^{-33}$ cm, far smaller than the distances to which elementary particles can be probed in particle accelerators on Earth.

The energies needed to smash particles to within a Planck length of each other were available to the universe at a time equal to the Planck length divided by the speed of light. This time, called the Planck time $(Gh/c^5)^{1/2}$, equals approximately $10^{-43}$ second. At the Planck time, the mass density of the universe is thought to approach the Planck density, $c^5/hG^2$, roughly $10^{93}$ g/cm$^3$. Contained within a Planck volume is a Planck mass $(hc/G)^{1/2}$, roughly $10^{-5}$ g. An object of such mass would be a quantum black hole, with an event horizon close to both its own Compton wavelength (distance over which a particle is quantum mechanically "fuzzy") and the event horizon of the universe at the Planck time. Under such extreme conditions, space-time cannot be treated as a classical continuum and must be given a quantum interpretation.

Quantum interpretation of space-time

The latter is the goal of the superstring theory, which has as one of its features the curious notion that the four space-time dimensions (three space dimensions plus one time dimension) of the familiar world may be an illusion. Real space-time, in accordance with this picture, has 26 or 10 dimensions, depending on whether one is dealing with bosons or fermions, but all of these dimensions except the usual four are somehow compacted to a size comparable to the Planck scale. Thus has the existence of these other dimensions escaped detection. It is presumably only during the Planck era, when the usual four space-time dimensions also acquire their natural Planck scales, that the existence of what is more fundamental than the usual ideas of mass-energy and space-time becomes fully revealed.

Some versions of superstring even allow for the existence of a "shadow world" containing bodies that do not interact with the familiar world in any way except through gravitation. There are speculations that these bodies may constitute the hidden mass of the universe.

**Inflation.** One of the more concrete results contributed by particle physicists to cosmology is the notion of inflation. The basic idea is that during the time of superunification (Planck era, $10^{-43}$ sec) or grand unification (GUT era, $10^{-39}$ sec), space may have been largely in a "false vacuum" state. This state was completely different from the true vacuum state familiar today. Notably, the false vacuum had a negative pressure capable of behaving as a cosmological constant—*i.e.*, repulsive force—of considerable magnitude when a bubble containing matter and antimatter appeared as a result of pair creation. According to the inflationary concept, this bubble, which eventually became the entire universe, expanded much more rapidly than it would have if the vacuum were real. With a large cosmological constant, the size of the bubble surface would have inflated exponentially in time through many orders of magnitude. Eventually, as the temperature dropped, the energy of the true vacuum state became lower than that of the false vacuum state. A phase transition to the

true vacuum state ensued. This brought the value of the cosmological constant to zero, and the universe entered an expansional epoch that closely conforms to a conventional relativistic model of the expanding universe.

Inflation serves a number of useful purposes. First, the drastic stretching during inflation flattens any initial space curvature, so the universe after inflation closely resembles an Einstein–de Sitter universe. Second, inflation so dilutes the concentration of any magnetic monopoles appearing as "topological knots" during the GUT era that their cosmological density drops to negligibly small and acceptable values. Finally, inflation provides a means of understanding the isotropy of the microwave background because the matter and radiation of the entire observable universe were in good thermal contact (within the cosmic event horizon) before inflation and therefore acquired the same temperature. Rapid inflation carried different portions outside of their individual event horizons. When inflation halted and the universe resumed normal expansion, these different portions, through the natural passage of time, reappeared on our event horizon. And through the observed isotropy of the cosmic microwave background, they are inferred still to have the same temperatures.

### THE STEADY STATE THEORY AND OTHER ALTERNATIVE COSMOLOGIES

Big bang cosmology is the theory of choice among almost all astronomers, but as successful and fruitful as the theory has been, it has not resolved all difficulties. For example, no consensus has been reached concerning the origin in the cosmic gas of fluctuations thought to produce the observed galaxies, clusters, and superclusters. To most astronomers these shortcomings indicate the incompleteness of the development of the theory, but it is conceivable that surprises are in order.

An early problem encountered by big bang theorists was an apparent large discrepancy between the Hubble time and other indicators of cosmic age. This discrepancy was resolved by revision of Hubble's original estimates for $H_0$, which were about 10 times too large because of confusion between Population I and II variable stars and between H II regions and bright stars. Nonetheless, the apparent difficulty motivated Bondi, Hoyle, and Gold to offer the alternative theory of steady state cosmology in 1948.

By that year, of course, the universe was known to be expanding; therefore, the only way to explain a constant (steady state) matter density was to postulate the continuous creation of matter to offset the attenuation caused by the cosmic expansion. This aspect was physically very unappealing to many people, who consciously or unconsciously preferred to have all creation completed in virtually one instant in the big bang. In the steady state theory the average age of matter in the universe is one-third the Hubble time, but any given galaxy could be older or younger than this mean value. Thus, the steady state theory had the virtue of making very specific predictions. Also, it was vulnerable to observational disproof.

Observational disproofs of the steady state theory

The first blow was delivered by Ryle's counts of extragalactic radio sources during the 1950s and '60s. These counts involved the same methods discussed above for the star counts of Kapteyn and the galaxy counts of Hubble except that radio telescopes were used. (These instruments are generally more sensitive than optical ones, permitting radio galaxies to be fairly sampled to much larger volumes than optical systems.) Ryle found more radio galaxies at large distances from the Earth than can be explained under the assumption of a uniform spatial distribution no matter which cosmological model was assumed, including that of steady state. This seemed to imply that radio galaxies must evolve over time in the sense that there were more powerful sources in the past (and therefore observable at large distances) than there are at present. Such a situation contradicts a basic tenet of the steady state theory, which holds that all large-scale properties of the universe, including the population of any subclass of objects like radio galaxies, must be constant in time.

Hoyle has argued that an alternative interpretation could still save the steady state hypothesis—namely, that there is a deficit of radio galaxies locally rather than an excess at large distances. The relative measure would be the same, but a local deficit could be explained away as a statistical fluke, whereas a distant excess would be a global property of the universe. However, the argument seemed forced, and the steady state theory could not survive the second blow dealt it in 1965 by the discovery of the cosmic microwave background radiation. Although it has few adherents today, the steady state theory is credited as having been a useful idea for the development of modern cosmological thought as it stimulated much work in the field.

At various times, other alternative theories have also been offered as challenges to the prevailing view of the origin of the universe in a hot big bang: cold big bang theory (to account for galaxy formation), symmetric matter–antimatter cosmology (to avoid an asymmetry between matter and antimatter), variable $G$ cosmology (to explain why the gravitational constant is so small), tired-light cosmology (to explain redshift), and the notion of shrinking atoms in a nonexpanding universe (to avoid the singularity of the big bang). The motivation behind these suggestions is, as indicated in the parenthetical comments, to remedy some perceived problem in the standard picture. Yet, in most cases, the cure offered is worse than the disease, and none of the mentioned alternatives has gained much following. Hot big bang theory has ascended to primacy because, unlike its many rivals, it attempts to address, not isolated individual facts, but a whole panoply of cosmological issues. And, though some sought-after results remain elusive, no glaring weakness has yet been uncovered.

Ascendancy of the hot big bang theory

## Summary

The history of human thought on the nature of the Cosmos offers a number of remarkable lessons, the most striking of which is that the architecture of the universe is open to reason. The plan is intricate and subtle, and each glimpse of another layer has led philosophers and scientists to a deeper mental image of the physical world. These images have surprising clarity and coherence—from the view of the Cosmos as geometry by the Greeks, to the mechanistic clockwork of the Newtonian universe, to the quirky subatomic "dance" of quantum particles and fields, to a geometric worldview with a relativistic and quantum twist. Each generation has had members who thought that they had found the path that would penetrate to the centre of innermost truth. The present generation is no different, but is there any real reason to believe that the process has stopped with its conclusions?

Yet, incomplete though it may be, the scope of modern scientific understanding of the Cosmos is truly dazzling. It envisages that four fundamental forces, along with matter-energy and space itself, emerged in a big bang. Forged in the heat of the primeval fireball were the two simplest elements, hydrogen and helium. As the fireball expanded and cooled, the dominance of gravity over matter led to the birth of galaxies and stars. As the stars evolved, hydrogen and helium were molded in their fiery interiors, and molded again into the heavy elements, which were subsequently spewed into interstellar space by titanic explosions that occurred with the death of massive stars. The enriched debris mixed with the gas of interstellar clouds, which collected into cool dense pockets and formed new generations of stars. At the outskirts of a spiral galaxy, the gravitational collapse of a rotating molecular cloud core resulted in the formation of the Sun, surrounded by a spinning disk of gas and dust. The dust, composed of the heavy elements produced inside stars, accumulated to form planetary cores of rock and ice. One such planet was fortunate enough to have water in all three phases; and carbon chemistry in the liquid oceans of that planet gave rise to living organisms that evolved and eventually conquered the land. The most intelligent of these land animals looked up at the sky and saw the planets and the stars, and in wonderment pondered the underlying plan of the Cosmos.

### BIBLIOGRAPHY

*General references:* A beautiful compendium of modern astronomy with many full-colour photographs and illustrations

is SIMON MITTON (ed.), *The Cambridge Encyclopaedia of Astronomy* (1977). Topical surveys of more limited scope are covered in the *Harvard Books on Astronomy* series; especially recommended are ROBERT W. NOYES, *The Sun, Our Star* (1982); LAWRENCE H. ALLER, *Atoms, Stars, and Nebulae,* rev. ed. (1971); CECELIA PAYNE-GAPOSCHKIN, *Stars and Clusters* (1979); BART J. BOK and PRISCILLA F. BOK, *The Milky Way,* 5th ed. (1981); HARLOW SHAPLEY, *Galaxies,* 3rd ed. rev. by PAUL W. HODGE (1972); and WALLACE TUCKER and RICCARDO GIACCONI, *The X-Ray Universe* (1985).

There are many introductory astronomy textbooks available that suppose little mathematical sophistication on the part of the reader; one of the most comprehensive is GEORGE O. ABELL, *Exploration of the Universe,* 4th ed. (1982). An introduction that begins with the big bang and works forward in time is DONALD GOLDSMITH, *The Evolving Universe,* 2nd ed. (1985). At a somewhat more advanced level is FRANK H. SHU, *The Physical Universe: An Introduction to Astronomy* (1982).

*History of astronomy:* The standard reference is A. PANNEKOEK, *A History of Astronomy* (1961, reissued 1969; originally published in Dutch, 1951). Excellent accounts of early ideas can be found in J.L.E. DREYER, *A History of Astronomy from Thales to Kepler,* 2nd ed. (1953); and GIORGIO DE SANTILLANA, *The Origins of Scientific Thought: From Anaximander to Proclus, 600 B.C.–500 A.D.* (1961, reissued 1970). Developments in the first half of the 20th century are covered authoritatively by OTTO STRUVE and VELTA ZEBERGS, *Astronomy of the 20th Century* (1962). The history of galactic astronomy is discussed in CHARLES A. WHITNEY, *The Discovery of Our Galaxy* (1971), while an engaging personal account of the early years of research in extragalactic astronomy is provided in EDWIN HUBBLE, *The Realm of the Nebulae* (1936, reprinted 1982).

*Planets:* Useful summaries are two collections of *Scientific American* articles: *The Solar System* (1975); and BRUCE MURRAY (ed.), *The Planets* (1983). Also recommended is J. KELLY BEATTY, BRIAN O'LEARY, and ANDREW CHAIKIN (eds.), *The New Solar System,* 2nd ed. (1982). The relationship of the origin of the solar system to theories of star formation is discussed at a technical level in DAVID C. BLACK and MILDRED SHAPLEY MATTHEWS (eds.), *Protostars and Planets II* (1985).

*Stars:* Very readable semipopular accounts of the theory of stellar evolution are ROBERT JASTROW, *Red Giants and White Dwarfs,* rev. ed. (1984); and IOSIF S. SHKLOVSKII, *Stars, Their Birth, Life, and Death* (1978; originally published in Russian, 1977). A classic text is MARTIN SCHWARZSCHILD, *Structure and Evolution of the Stars* (1958, reissued 1965). Stellar nucleosynthesis is the emphasis of DONALD D. CLAYTON, *Principles of Stellar Evolution and Nucleosynthesis* (1968, reprinted 1983). The properties of gravitationally compact stellar remnants are discussed by STUART L. SHAPIRO and SAUL A. TEUKOLSKY, *Black Holes, White Dwarfs, and Neutron Stars: The Physics of Compact Objects* (1983); a more elementary treatment is contained in HARRY L. SHIPMAN, *Black Holes, Quasars, and the Universe,* 2nd ed. (1980).

*Galaxies:* Beautiful photographs of galaxies together with nontechnical commentary are contained in TIMOTHY FERRIS, *Galaxies* (1980, reissued 1982). Equally enjoyable for the amateur and professional alike are ALLAN SANDAGE, *The Hubble Atlas of Galaxies* (1961); HALTON ARP, *Atlas of Peculiar Galaxies* (1966, reprinted 1978); and ALLAN SANDAGE and G.A. TAMMANN, *A Revised Shapley-Ames Catalog of Bright Galaxies* (1981). Extragalactic astronomy is discussed at a level appropriate for professionals in ALLAN SANDAGE, MARY SANDAGE, and JEROME KRISTIAN (eds.), *Galaxies and the Universe* (1975, reprinted 1982); S.M. FALL and D. LYNDEN-BELL (eds.), *The Structure and Evolution of Normal Galaxies* (1981); and C. HAZARD and SIMON MITTON (eds.), *Active Galactic Nuclei* (1979). The problems of galaxy formation or galaxy clustering are described by JOSEPH SILK, *The Big Bang: The Creation and Evolution of the Universe* (1980); and in mathematical detail by P.J.E. PEEBLES, *The Large-Scale Structure of the Universe* (1980).

*Cosmology:* Several excellent semipopular accounts are available: TIMOTHY FERRIS, *The Red Limit: The Search for the Edge of the Universe,* 2nd rev. ed. (1983); STEVEN WEINBERG, *The First Three Minutes: A Modern View of the Origin of the Universe* (1977, reissued 1984); NIGEL CALDER, *Einstein's Universe* (1979, reissued 1982); EDWARD R. HARRISON, *Cosmology, the Science of the Universe* (1981); ROBERT V. WAGONER and DONALD W. GOLDSMITH, *Cosmic Horizons: Understanding the Universe* (1982, reprinted 1983); PAUL DAVIES, *The Edge of Infinity: Where the Universe Came From and How It Will End* (1981, reissued 1982; U.K. title, *The Edge of Infinity: Naked Singularities and the Destruction of Spacetime*); and JOHN BARROW and JOSEPH SILK, *The Left Hand of Creation: The Origin and Evolution of the Expanding Universe* (1983, reissued 1984). Standard textbooks on general relativity and cosmology include P.J.E. PEEBLES, *Physical Cosmology* (1971); STEVEN WEINBERG, *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity* (1972); and CHARLES W. MISNER, KIP S. THORNE, and JOHN ARCHIBALD WHEELER, *Gravitation* (1973). The interface between particle physics and cosmology is the concern of G.W. GIBBONS, S.W. HAWKING, and S.T.C. SIKLOS, *The Very Early Universe* (1983).

(F.H.Sh.)

# Crime and Punishment

Within a broad spectrum of cultural and historical variations, crime constitutes the intentional commission of an act usually deemed socially harmful or dangerous and specifically defined, prohibited, and punishable under the criminal law. Most countries have enacted a criminal code in which all of the criminal law can be found, although English law—the source of many other criminal law systems—remains uncodified. The definitions of particular crimes contained in a code must be interpreted in the light of many principles, some of which are not expressed in the code itself. The most important of these are related to the mental state of the accused person at the time of the act that is alleged to constitute a crime. Crimes are classified by most legal systems for purposes such as determining which court has authority to deal with the case. Social changes often result in the adoption of new criminal laws and the obsolescence of older ones.

The purpose of punishing offenders has been debated for centuries. A variety of often conflicting theories are held, and in practice each is followed to some extent. Prison is not the most common penalty for crime—a wide variety of punishments that do not involve incarceration have developed, including financial sanctions, such as fines, and schemes for service to the community in general or to the victim in particular. Juveniles are usually dealt with by courts set aside exclusively for the prosecution of young offenders.

The prison systems of most countries are subject to many problems, especially overcrowding, but the recognition by some legal systems that prisoners have rights that the courts can enforce has led to some improvements. The death penalty is now rare in Western countries, although it has been reinstated in some parts of the United States after a period of disuse.

The present article treats the definition, incidence, and prevailing theories of criminal activity, the conduct of all stages of criminal proceedings, and various theories and practices of punishment. The material draws principally from common, or Anglo-American, law, with supplementary treatment of other systems. For full treatment of particular legal aspects of crime and punishment, see the articles CRIMINAL LAW; JUDICIAL AND ARBITRATIONAL SYSTEMS; LEGAL SYSTEMS, THE EVOLUTION OF MODERN WESTERN; POLICE; and PROCEDURAL LAW.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 521, 522, and 543.

The article is divided into the following sections:


The concept of crime 796
- Criminal codes and other legal formulations 796
- General principles of criminal law 797

Classification of crimes 798
- General considerations 798
- Some particular crimes 798
- Other modes of criminal activity 800

Measurement of crime 800
Analysis of crime 801
- Characteristics of offenders 801
- Theories of causation 802

Detection of crime 803
- The role of forensic science 803
- Modus operandi and suspect identification 804
- Gathering evidence 804
- Interrogation and confession 804

Prosecution 805
The court system 805
- Criminal procedure in English-speaking countries 805
- Other systems of criminal procedure 805

Sentencing 806
Parole 806
Theories and objectives of punishment 807
- Retribution 807
- Deterrence 808
- Incapacitation 808
- Rehabilitation 808
- Theories in conflict 809
- Punishment in other systems 809

Treatment of juvenile offenders 809
Prisons 810
- Development of the penitentiary 810
- The prison population 811

The death penalty 812
- In English law 812
- In the United States 813
- In other countries 813
- The continuing controversy 813

Alternatives to prison 814
Crime and social policy 815
Bibliography 815


## The concept of crime

### CRIMINAL CODES AND OTHER LEGAL FORMULATIONS

Crime is whatever conduct the laws of a particular jurisdiction designate as criminal, and there are many differences from one country to another as to what behaviour is prohibited. Conduct that is lawful in one country may be criminal in another, and activity that amounts to a trivial infraction in one country may constitute a serious crime elsewhere. Changing times and social attitudes may lead to changes in the criminal law, so that behaviour that was once criminal becomes lawful. Abortion, once prohibited except in the most unusual circumstances, has become lawful in many countries, as has homosexual behaviour in private between consenting adults, which was once a serious offense. Suicide and attempted suicide, once criminal, have also been removed from the scope of the criminal law in many countries. Nonetheless, the trend generally is to increase the scope of the criminal law rather than to reduce it. It is more common to find that statutes create new criminal offenses than that they abolish old ones. New technologies give rise to new opportunities for their abuse, which in turn give rise to legal restrictions; just as the invention of the motor vehicle led to the development of a whole body of criminal laws designed to regulate its use, so the widening use of computers has created the need to legislate against a variety of new abuses and frauds—or old frauds committed in new ways.

**The English-speaking world.** In most countries the criminal law is contained in a single statute, known as the criminal code or penal code. Although the criminal codes of most English-speaking countries are derived from English criminal law (in many cases with substantial modifications), England itself has never had a criminal code. English criminal law still consists of a collection of statutes of varying age (the oldest still in force being the Treason Act, 1351) and a set of general principles that are chiefly expressed in the decisions of the courts (case law). The absence of a criminal code in England is not for want of effort; since the early 19th century there has been a series of attempts to reduce the English criminal law to a code. The first effort, which occurred between 1833 and 1853, was by two panels of criminal law commissioners, who systematically surveyed the prevailing state of the

Attempts to formulate an English criminal code

criminal law. They were confronted by a vast number of often overlapping and inconsistent statutes. Determining precisely what the law provided on any particular topic was enormously difficult; the existence of different statutes covering the same conduct, often with widely varying penalty provisions, permitted wide judicial discretion and inconsistency of punishment. The English criminal law commissioners drew up a number of draft codes that were presented to Parliament but not enacted. Eventually the resistance of the judiciary led to the abandonment of the movement toward codification, and instead there was a consolidation of most of the statutory criminal law in 1861 into a number of statutes—the Larceny Act, 1861, the Malicious Damage Act, 1861, and the Offences against the Person Act, 1861, were the most important of these. As these statutes were consolidations rather than codifications, they preserved many of the difficulties of the earlier legislation, which was in effect reproduced in the form of a single statute without substantial change. The Offences against the Person Act is still largely in force; the others have been replaced by more modern provisions.

Interest in codification was not limited to England; a similar process ensued in India, then under British rule, and a criminal code in the true sense was written during the 1830s and eventually enacted in 1860. It remains substantially in force in India and Pakistan and in certain parts of Africa that were once British colonies. The effort to produce a criminal code in England was resumed in 1877, and a further Criminal Code Bill was presented to Parliament in 1879–80. This draft code, largely the work of the celebrated legal author and judge James Fitzjames Stephen, was not enacted in England largely because Parliament was preoccupied with other matters at the time, but it was subsequently enacted in Canada as the Canadian Criminal code in 1892 and in several Australian states and other British colonies.

Efforts in the United States

Reform of the criminal law became one of the interests of the U.S. states in the period following the Revolution, and in the early 1820s a comprehensive draft code was prepared for Louisiana, although it was never enacted. Other states moved to codify their criminal laws, and New York enacted a criminal code in 1881, setting an example that was eventually followed by the majority of states. Because criminal law is primarily a matter for the individual states (in contrast to Canada, for example, where the national parliament enacts the criminal code for the country as a whole), there has been considerable variation in the content of the criminal code from one state to another. In 1950 the American Law Institute began more than a decade of effort that was to lead eventually to the publication of the Model Penal Code in 1962. This was an attempt to rationalize the criminal law in relation to modern society and to establish a logical framework for defining offenses and a consistent body of general principles on such matters as criminal intent and the liability of accomplices. The Model Penal Code had a profound influence on the revision of many individual state codes over the following 20 years; the code itself was never enacted completely, but it inspired and influenced a long period of criminal code reform. (D.A.T.)

**Other systems.** Criminal offenses in the legislation of modern African countries are, with the exception of Sierra Leone and some southern African states, now defined in criminal or penal codes. As far as the common-law countries are concerned, this is a radical departure from the originally uncodified criminal law of England, on which these codes are largely based. Because of their origins, these codes reflect the penal assumptions of the original colonial metropole. The only concessions to local African values or problems are, first, the inclusion of legislation against various customary practices, notably witchcraft; second, the extension of the criminal law in states with a planned economy to cover economic crimes against the state; and third, as a consequence of the soaring rate of some kinds of crime, special provision for offenses such as armed robbery. Special tribunals, not subject to the ordinary rules of procedure, have been established in many countries to deal with such offenses; similarly, special tribunals and commissions with punitive powers have been set up to investigate the assets or the misdeeds of former rulers displaced by coups d'état or revolutions.

The states of the world that have large Muslim communities fall into four main categories in matters of criminal law. First, there are those that have an English colonial past and that have in the main adopted English criminal law and procedure, such as Pakistan, Bangladesh, Jordan, and some of the Persian Gulf states. The second group comprises those states that came under French colonial influence and adopted French laws; these are the states of the Maghrib and North Africa, including Egypt, and also Syria and Iraq. The third group comprises those states that were relatively little influenced by a colonial presence and that retain Islāmic law (called Sharī'ah) with few or no reforms; these include Saudi Arabia and Iran (the last shah of Iran had reformed a large amount of the law, building on previous colonial laws, but this was almost totally eradicated following the revolution of 1979). The fourth group comprises those states, such as India and some East African countries, where Muslims are a minority community.

In many respects the pattern of crime and punishment found in the Soviet Union resembles that in western Europe and North America, but it has distinctive features as well. Some of them derive from the Soviet commitment to socialist or Marxist-Leninist ideas; others stem from long-standing Russian traditions. The theory and practice of criminology and penology in China have been influenced both by China's indigenous legal tradition and by Marxist-Leninist theory, especially as practiced in the Soviet Union. (A.N.A./I.D.E./P.H.S./D.C.C.)

## GENERAL PRINCIPLES OF CRIMINAL LAW

Despite differences of form and detail, the general principles of criminal law have much in common throughout the English-speaking world. One widely accepted principle is the rule against retroactivity—an individual may not be punished for an action that was not designated a crime at the time it was carried out. This rule, which restricts the authority of the judges to declare new offenses (although not necessarily to expand the scope of old ones by interpretation) has not always been accepted in England; as late as 1960 the House of Lords in its judicial capacity claimed that the courts retained the authority to recognize new offenses as social needs changed and declared that it was criminal to publish a directory of prostitutes. In Scotland (the legal system and criminal courts of which are totally separate from those of England) the claim is still maintained. A shopkeeper who supplied equipment for glue sniffing to children was convicted and imprisoned, notwithstanding that the law contained no express provision prohibiting such conduct.

The rule against retroactivity

Determining which particular conduct constitutes a crime usually requires an examination of the terms of the relevant provisions of the code or statutory provisions (a few offenses in English law have not been defined in statute), but these must be interpreted in the light of a number of general principles. The most important of these is that an individual is not normally to be held guilty of a crime unless he intended or foresaw the consequences of his action or was aware of the circumstances that make it criminal. In advanced Western societies legal codes frequently recognize mental abnormality such as schizophrenia, mental retardation, or paranoia as at least mitigating, if not absolving, factors, though the claim of insanity may be contested. Criminal law as a general rule does not punish accidental or negligent behaviour. This principle, known as the mens rea ("guilty mind") principle, is subject to many exceptions and qualifications. For many offenses (offenses of strict liability) it is abandoned completely, and in other cases it is allowed only a limited scope. If the offense is one that requires proof of intention or knowledge on the part of the accused, the court or jury must be satisfied that the accused himself had the necessary intention or knowledge at the time when he committed the act that constitutes the crime—it is not normally sufficient to prove that any ordinary person in his place would have realized what was likely to happen. This is a difficult matter to prove, and in practice the members of the jury can be guided

Proof of intention

only by what they would have intended themselves, if they can imagine themselves in the same situation as the accused, and by the accused person's explanations of his behaviour. The fact that an individual had been drinking before committing a crime is not in itself a defense, but it may in some cases be evidence that the accused person did not have the intention that the law requires for the offense with which he is charged. It is no defense for an accused person to say that because he had been drinking he acted out of character and did things that he would not have done if sober, but it may help him to persuade the court that he did not realize what the consequences of his actions would be if he shows that he was affected by drink. Provocation is not generally a defense to a criminal charge, except in the case of murder; in a murder case evidence of a high degree of provocation (in English law, sufficient to provoke a reasonable person to act in the same way as the accused did) results in a verdict of manslaughter, even if the killing was intentional. One very rare condition that gives a complete exemption from criminal liability is a form of involuntary conduct known as automatism. This is a state (such as sleepwalking or certain effects of concussion) in which the conscious mind does not control the bodily movements, and thus the individual cannot be held accountable for potential consequences, however serious they may be.

Criminal responsibility is not limited only to those who perform the criminal acts themselves. As a general principle, anyone who "aids and abets" the perpetrator by encouraging or in any way knowingly helping him (for instance, by providing information, implements, or practical help) is an accomplice and is considered equally guilty. Those who actually perform the criminal act (*e.g.,* wielding the weapon that strikes the fatal blow) are called principals in the first degree; those who assist at the time of the commission of the offense (*e.g.,* holding the victim down while the principal in the first degree strikes the blow) are principals in the second degree; and those who assist before the crime takes place (*e.g.,* by lending the weapon or by providing information) are accessories before the fact. As a general rule, all are equally responsible in the eyes of the law and liable to the same punishment. In many cases the accessory before the fact will be considered more culpable—if, for instance, he has instigated the offense and arranged for it to be committed by an associate. In some cases the person who actually performs the act that causes the crime is completely innocent of all intent—for instance, the nurse who administers to a patient, on the doctor's instructions, what she believes to be medicine but what is in fact poison. In this situation the person who carries out the act is an innocent agent and is not criminally responsible; the person who causes the innocent agent to act is the principal in the first degree. The accessory after the fact is one who helps a felon to evade arrest or conviction, by, for example, hiding him or destroying evidence. In England and some other jurisdictions the expression is no longer in use, as specific offenses have been enacted to deal with this kind of behaviour.

## Classification of crimes

### GENERAL CONSIDERATIONS

Most legal systems find it necessary to divide crimes into categories for various purposes connected with the procedure of the courts—determining, for instance, which kind of court may deal with which kind of offense. The common law originally divided crimes into two categories—felonies (the graver crimes, generally punishable with death, which resulted in forfeiture of the perpetrator's land and goods to the crown) and misdemeanours (for which the common law provided fines or imprisonment). There were many differences in the procedure of the courts according to whether the charge was felony or misdemeanour, and other matters that depended on the distinction included the power of the police to arrest a suspect on suspicion that he had committed an offense (which was generally permissible in felony, but not in misdemeanour). By the early 19th century it had become clear that the growth of the law had rendered this classification obsolete and in many cases inconsistent with the gravity of the offenses concerned (theft was a felony, irrespective of the amount stolen; obtaining by fraud was always a misdemeanour). Efforts to abolish the distinction in English law did not succeed until 1967, when the distinction was replaced by that between arrestable offenses and other offenses (an arrestable offense is one punishable with five years' imprisonment or more; offenders may be arrested for other offenses subject to certain conditions). In later legislation it has proved necessary to devise further classifications—for the purposes of powers of investigation, a category of serious arrestable offenses has been created, and, for the purpose of deciding in which courts the case should be tried, a different classification of offenses into three categories of indictable, "either way," and summary has been adopted. The traditional classification between felony and misdemeanour has been retained in many U.S. jurisdictions (although there has been rationalization of the allocation of offenses to one category or the other) and is used as the basis of determining the court that will hear the case. In some jurisdictions a further class of offense (violations) has been added, to include minor offenses (corresponding broadly to the English category of summary offenses).

Felony and misdemeanour

### SOME PARTICULAR CRIMES

**Murder.** In English tradition, murder was defined as the willful killing with malice aforethought of a human creature in being within the king's peace, the death occurring within a year and a day of the injury. Most of these elements remain in modern definitions of the crime—the requirement that the victim is "in being," for instance, distinguishes abortion from murder—although in some respects the definition has become more complex. Many of the problems of defining murder have centred on the mental element—the "malice aforethought." The old English rule extended this concept to include not only intentional or deliberate killings but also accidental killings in the course of some other serious crime (such as robbery or rape). This rule, the felony murder rule, was adopted in many other jurisdictions, although it has often produced harsh results when death has been caused accidentally in the course of what was intended to be a minor crime. The rule was abolished in England in 1957, but since then English law has been in a state of confusion over the precise definition of murder. It is now settled that an intention to kill is not necessary and that an intention to cause serious bodily injury is sufficient, but the precise interpretation of intention in this context remains controversial. Similar problems have arisen in many U.S. jurisdictions, some of which distinguish between different degrees of murder—first-degree murder may require proof of premeditation over and above the normal requirement of intention. Surprisingly, murder and manslaughter are not mentioned in the Qur'ān and are subject in Islāmic countries to customary law as amended by Sharī'ah. Virtually all systems treat murder as a crime of the utmost gravity, providing in some cases the death penalty or a special form of sentence, such as a mandatory life sentence, often with restrictions on parole. A high proportion of murders in all societies are committed spontaneously by persons acquainted with the deceased, often a member of the same family, as a result of quarrels or provocation. The convicted murderer is often a person with no other criminal conviction.

"Malice aforethought"

**Rape.** The traditional legal definition of rape is the performance of sexual intercourse by a man other than her husband with a woman against her will, by force or fraud. This definition has been adapted in the statutes of some jurisdictions; in Canada the crime of rape has been abolished as a separate offense and merged into a wider general category of sexual assault. Most jurisdictions do not treat as rape an act of sexual intercourse by a husband with his wife without her consent, unless the marriage has effectively been terminated by a legally recognized separation. Although many rapes involve the application or threat of violence, it is possible to commit rape by fraud—either by persuading the victim that what is to take place is not sexual intercourse (by representing it as medical treatment, for instance) or by impersonating some other person, such as the victim's husband. Under the provi-

sions of most criminal codes, rape requires penetration of the female organ by the male organ (but does not require ejaculation); other forms of sexual abuse (such as oral penetration or anal penetration) are dealt with, if at all, under different provisions. In the United States, however, the crime of rape may also include forcible sodomy, and victims of rape may be women or men. In many rape trials the issue is whether the victim consented to the sexual intercourse, and this may lead to distressing cross-examination, in some cases about the woman's previous sexual behaviour, whether with the accused or with other persons. In many jurisdictions cross-examination of the complainant on such matters is now restricted, and the embarrassment of the complainant is further mitigated by provisions restricting publication of the woman's identity. Proof is made more difficult by the common need to prove not only that the victim did not consent but also that the accused knew this—or at least was aware of that possibility.

When guilt is established, rape in most systems of criminal law is treated as a grave crime; 95 percent of those convicted of rape in England, for instance, are sentenced to imprisonment. A high proportion of rapists escape conviction for a variety of reasons. The victim may be reluctant to report the incident, possibly because of fears of hostile treatment by investigating authorities or by defense lawyers in court; there is a higher than average acquittal rate of those indicted for rape, as a result of the difficulty of proving a crime of which there are rarely any witnesses other than the complainant and the accused. The motivation of rapists is now acknowledged to be a more complex matter than was formerly believed; it has come to be widely accepted that rape is not necessarily the result of sexual desire but is more likely to be motivated by aggression and the desire to humiliate or exercise domination over the victim.

**Incest.** The crime of incest consists of sexual intercourse between near relatives. Incest was not a crime under the common law but was punishable historically in the ecclesiastical courts. Legislation prohibiting incest was enacted in England in 1908, and most English-speaking jurisdictions now prohibit intercourse between close relatives, but there are differences among the systems in the relationships within which intercourse is forbidden. Most systems forbid intercourse between immediate relatives—father and daughter, brother and sister, mother and son. There are some anomalies—English law prohibits intercourse between grandfather and granddaughter but not between grandmother and grandson. Consent to intercourse is irrelevant to the charge of incest, but if there is no consent, the crime of rape may also be committed. Generally, sexual intercourse between family members who are not related by blood—for instance, stepfather and stepdaughter—is not considered incest, but this is prohibited in some jurisdictions. Both parties are considered guilty if incest occurs, but in many systems there is an exemption from liability for women below a certain age (16 in England, 18 in some U.S. jurisdictions). Most cases of incest that come before criminal courts concern sexual intercourse between fathers and relatively young daughters, and it is believed that incest in this form is far more common than the statistics of court cases suggest. Treatment of the offenders in such cases presents acute difficulties to the courts—on the one hand, the offense is widely regarded as serious, involving sexual abuse of children and a breach of the parent's responsibility for the child's welfare; on the other hand, to impose a severe penalty, such as imprisonment, on the father may result in the destruction of the family unit and the infliction of other deprivations on the child victim, in particular feelings of guilt for being responsible for the imprisonment of the father.

Sexual abuse of children

**Perjury.** Perjury originally consisted in the giving of false evidence on oath to a court of law; it has now been expanded in most jurisdictions to include evidence given otherwise than on oath (under affirmation, for instance, by a person who objects to swearing) and to other tribunals that have the authority of the law. Perjury may be committed by witnesses from either the prosecution or the defense (or by witnesses on either side in civil litigation) and in proceedings before the jury or after the verdict in proceedings leading to sentence. As a general rule, the accused person must make a false statement that he either knows to be false or does not believe to be true, and the false statement must normally be material to the matters in issue in the proceedings. In many jurisdictions the law imposes special requirements for the proof of perjury—it is not normally sufficient to rely on the evidence of one witness as to the falsity of the alleged perjured statement. Crimes associated with perjury include subornation of perjury (persuading other persons to commit perjury) and a wide variety of statutory offenses involving making false statements in official documents (such as applications for drivers' licenses) that are not normally treated so seriously.

**Prostitution.** The laws regulating prostitution vary greatly from one jurisdiction to another. In some jurisdictions prostitution—generally defined as the provision of sexual services for money—is itself illegal; in others the act of prostitution is not illegal in itself, but many associated activities are unlawful. English law, for instance, does not prohibit prostitution itself but does prohibit soliciting for prostitution in a public place, living on the earnings of prostitution, exercising control over prostitutes, or keeping a brothel (any premises where two or more prostitutes are employed). In some jurisdictions, notably in the state of Nevada, prostitution is lawful and practiced openly subject only to health and related controls.

**Arson.** In common law, arson consisted of setting fire to the dwelling of another person. Subsequent statutes have expanded the scope of the offense to include setting fire to other types of buildings, and in English law any kind of damage deliberately caused by fire—even setting fire to rubbish—is now arson, but generally setting fire to a building is necessary. The gravity of the crime may depend on the extent to which life is endangered—the law may distinguish between arson endangering life, or arson of occupied buildings, and other forms of arson, but most systems consider the crime a serious one. The motivation of those who commit arson differs—arson may be committed as an act of revenge against an employer or by a jealous lover, for example, or by persons who find excitement in fires or have pathological impulses to set fires. Schools are sometimes set on fire by pupils out of resentment or simple vandalism. Some arson is more rationally motivated—a burglar may set fire to a house to conceal the evidence of his crime, as may an employee who is anxious to conceal accounts from an auditor. Another phenomenon is setting fire to premises belonging to the fire setter in order to make a fraudulent insurance claim.

**Theft and burglary.** Theft, sometimes still known by the traditional name of larceny, is probably the most common crime involving a criminal intent. The crime of grand larceny in some U.S. jurisdictions consists of stealing more than a specified sum of money or property worth more than a specified amount. The traditional definition of theft specified the physical removal of an object that was capable of being stolen, without the consent of the owner and with the intention of depriving the owner of it permanently. This intention, which has always been an essential feature of theft, does not necessarily mean that the thief must intend to keep the property—an intention to destroy it, or to abandon it in circumstances where it will not be found, is sufficient. In many legal systems the old definition has been found to be inadequate to deal with modern forms of property that may not be physical or tangible (a bank balance, for instance, or data stored on a computer), and more sophisticated definitions of theft have been adopted in modern legislation. The distinction that the common law made between theft (taking without consent) and fraud (obtaining with consent, as a result of deception) has been preserved in many modern statutes, but the two crimes are rarely regarded as mutually exclusive, as they were in the past. It is now accepted that an act may constitute both theft and fraud, as in the theft and subsequent sale of an automobile.

Fraud

Burglary consisted originally of breaking into a dwelling by night with intent to commit a felony, but as in the case of many other crimes the definition has been expanded in many legal systems. In English law, any entry by an

individual into a building as a trespasser with intent to commit theft or certain other offenses is burglary, and some jurisdictions recognize an offense of burglary of an automobile—breaking into it to steal the contents. The essence of burglary is normally the entry into a building with a criminal intent. Entry without the intent to commit a crime of the kind specified in the burglary statute is not burglary—it is merely a trespass, which is not criminal in many jurisdictions. Although the motivation of most burglars is theft, an intention to commit various other offenses converts a trespass into a burglary—it is possible, for instance, to commit burglary with intent to rape.

Robbery is the commission of theft in circumstances of violence. It involves the application or the threat of force in order to commit the theft or to secure escape. Robbery takes many forms—from the mugging of a stranger in the street, in the hope of stealing whatever he may happen to have in his possession, to much more sophisticated robberies of banks or similar premises, involving numerous participants and careful planning. (D.A.T.)

OTHER MODES OF CRIMINAL ACTIVITY

**Organized crime.** In addition to that segment of the population made up of individual criminals acting independently or in small groups, there exists a so-called underworld of criminal organizations engaged in offenses such as cargo theft, fraud, robbery, kidnapping for ransom, and the demanding of "protection" payments. In the United States and Canada, the principal source of income for organized crime is the supply of goods and services that are illegal but for which there is continued public demand. Examples include drugs, prostitution, loan-sharking (lending money at extremely high rates of interest), and gambling. To ensure freedom from the law, the organizations must subvert both the police and the courts.

Organized crime in the United States is best viewed as a set of shifting coalitions, normally local or regional in scope, between groups of gangsters, business people, politicians, and union leaders. Many of these people have legitimate jobs and sources of income. So-called street-level criminals are normally independent of major crime syndicates. Among other advanced industrial nations, the closest similarities to this organizational model occur in Australia, where extensive narcotics, cargo theft, and labour racketeering rings have been discovered, and in Japan, where there are gangs that specialize in vice and extortion. In Britain groups of organized criminals have not developed in this way, principally because the supply and consumption of alcohol and opiates, gambling, and prostitution remain legal but partly regulated, owing to a more liberal and pragmatic attitude of successive governments aware of the impossibility of total enforcement. This apparent laxity reduces the profitability of supplying such demands criminally. Far Eastern groups such as the Chinese Triads are important in the supply of drugs to and by way of Britain. Except for cargo thieves who work at airports and local vice, protection, and pornography syndicates, British crime organizations tend to be relatively short-term groups drawn together for specific projects, such as fraud and armed robbery, from a pool of long-term professional criminals.

In many Third World countries, apart from the drug trade, the principal form of organized crime is black-marketeering, including smuggling and corruption in the granting of licenses to import goods and to export foreign exchange. Armed robbery, cattle theft, and maritime piracy and fraud are organized crime activities in which politicians have less complicity. Robbery is particularly popular and easy because of the widespread availability of arms supplied to nationalist movements by those who seek political destabilization of their own or other states, and who may therefore exploit the dissatisfaction of ethnic and tribal groups.

**"White-collar" crime.** Crimes committed by business people, professionals, and politicians in the course of their occupation are known as "white-collar" crimes, after the typical attire of their perpetrators. Contrary to popular usage, criminologists tend to restrict the term to those illegal actions intended by the perpetrators principally to further the aims of their organizations rather than to make money for themselves personally. Examples include conspiring with other corporations to fix prices of goods or services in order to make artificially high profits or to drive a particular competitor out of the market; bribing officials or falsifying reports of tests on pharmaceutical products to obtain manufacturing licenses; and constructing buildings or roads with cheap, defective materials while charging for components meeting full specifications. Often such activities are attributable to over-enthusiastic employees or executives acting on their own initiative, but sometimes they represent a form of "upperworld" organized crime.

Corporate crime

The cost of corporate crime in the United States has been estimated at $200,000,000,000 a year—three times the cost of organized crime. Such crimes have a huge impact upon the safety of workers, consumers, and the environment, but they are seldom detected. Compared with crimes committed by juveniles or the poor, corporate crimes are very rarely prosecuted in the criminal courts, and executives seldom go to jail, though companies may pay large fines.

The term white-collar crime is used in another sense, by the public and academics, to describe fraud and embezzlement. Rather than being crime "by the firm, for the firm," this constitutes crime for profit by the individual against the organization, the public, or the government. (Tax fraud costs at least 5 percent of the gross national product in most developed countries.) Because of the concealed nature of many frauds and the fact that few are reported even when discovered, the cost is impossible to estimate precisely, but in the United States it is thought to be at least 10 times the combined cost of theft, burglary, and robbery. (Mi.L.)

**Terrorism.** From the 1960s, international terrorist crimes, such as the hijacking of passenger aircraft, political assassinations and kidnappings, and urban bombings, constituted a growing phenomenon of increasing concern, especially to Western governments. Most terrorist groups are associated either with millenarian revolutionary movements on an international scale (such as some Marxist organizations) or with nationalist movements of particular ethnic, religious, or other cultural focus.

Three broad categories of terrorist crime may be distinguished, not in legal terms, but by intention. Foremost is the use of violence and the threat of violence to create public fear. This may be done by making random attacks to injure or kill anyone who happens to be in the vicinity when an attack takes place. Because such crimes deny, by virtue of their being directed at innocent bystanders, the unique worth of the individual, terrorism is said to be a form of crime that runs counter to all morality and so undermines the foundations of civilization. Another tactic generating fear is the abduction and assassination of heads of state and members of governments in order to make others afraid of taking positions of leadership and so to spread a sense of insecurity. Persons in responsible positions may be abducted or assassinated on the grounds that they are "representatives" of some institution or system to which their assailants are opposed.

A second category of terrorist crime is actual rule by terror. It is common practice for leaders of terrorist organizations to enforce obedience and discipline by terrorizing their own members. A community whose collective interests the terrorist organization claims to serve may be terrorized so that their cooperation, loyalty, and support are ensured. Groups that come to power by this means usually continue to rule by terror.

Third, crimes are committed by terrorist organizations in order to gain the means for their own support. Bank robbery, kidnapping for ransom, extortion, gambling rake-offs (profit skimming), illegal arms dealing, and drug trafficking are among the principal crimes of this nature. In the Middle East, hostages are frequently sold as capital assets by one terrorist group to another. (Ji.B.)

## Measurement of crime

Estimating the amount of crime actually committed has troubled criminologists for many years; the figures for

recorded crime do not give an accurate picture because they are influenced by variable factors such as the willingness of victims to report crimes. It is widely believed that only a small fraction of the crime actually committed is reported to authorities. For this reason the criminal who is detected is not necessarily representative of all those who commit crime, and thus attempts to explain the causes of crime by reference to those who are identified as criminals must be approached with caution.

The public's view of the frequency and gravity of crime, obtained largely from the news media, may be seriously distorted, as the media tend to concentrate on serious or sensational crimes and often fail to give a full and accurate picture of what has happened. A brief report of a case in court, for instance, inevitably is selective; much of the evidence that the court has heard is omitted. A more detached view may be provided by detailed statistics of crime compiled and published by a department of government—in the United States, for instance, the Federal Bureau of Investigation publishes an annual, the *Uniform Crime Reports,* and in England the Home Office produces each year a volume entitled *Criminal Statistics, England and Wales,* which (among other things) gives an account of the trends in different types of crime. Official statistics such as these are frequently used by policymakers as the basis for new procedures in crime control—they may show, for instance, that there has been an increase in the incidence of a particular type of crime over a period of years and suggest, therefore, that some change in the methods of dealing with that type of crime is necessary. In fact, many official statistics of crime are subject to serious error and may be almost as misleading as the general impressions formed by the public through the news media, particularly if they are used without an understanding of the processes by which they are compiled and the limitations to which they are necessarily subject. The statistics are usually compiled on the basis of reports from police forces and other law enforcement agencies and are generally known as statistics of reported crime, or crimes known to the police. Because only incidents observed by the police or reported to them by victims or witnesses are included in the reports, the picture of the amount of crime actually committed may be distorted. One factor accounting for this distortion is the extent to which police resources are allocated to the investigation of one kind of crime rather than another, particularly with regard to what are known as "victimless crimes," such as possession of drugs. These crimes are not discovered unless the police set out to look for them, and they do not figure in the statistics of reported crime unless the police take the initiative; thus, a sudden increase in the reported incidence of a crime from one year to the next may merely show that the police have taken more interest in that crime and devoted more resources to its investigation. Ironically, efforts to discourage or eliminate a particular kind of crime through more vigorous law enforcement may create the impression that the crime concerned has increased rather than decreased, because more instances are detected and thus enter the statistics.

Crime statistics

A second factor that can have a striking effect on the apparent statistical incidence of a particular kind of crime is a change in the willingness of victims of the crime to report it to the police. It is believed by most criminologists, on the basis of research, that the crime reported to the police amounts to only a small proportion of the crime actually committed. Estimates of the so-called dark figure (the number of unreported crimes) vary, but it is thought that in some cases the reported crimes may constitute less than 10 percent of those actually committed. Victims of crimes have many reasons for not reporting them: they may not realize that a crime has been committed against them (children who have been sexually molested, for instance); they may believe that the police will not be able to detect the offender; they may be afraid of involvement in the processes of the law as witnesses; they may be embarrassed by their own conduct that has led them to become the victim of the crime (a man robbed by a prostitute, for instance, or a person who has been the victim of a confidence trick as a result of his own greed or credulity). A particular type of crime may not appear sufficiently serious to make it worthwhile to inform the police, or there may be ways in which the matter can be resolved without involving them—an act of violence by one schoolchild against another may be dealt with by the school authorities, or a dishonest employee may be dismissed without prosecution. All of these factors are difficult to measure with any degree of accuracy, and there is no reason to suppose that they remain constant over a period of time. Thus, a change in any one of these factors may produce the appearance of an increase or a decrease in a particular kind of crime, when in fact there has been no such change, or the real change has been on a much smaller scale than the statistics suggest.

Reporting of crimes

A third factor that may affect the picture of crime presented by official statistics is the way in which the police treat particular incidents. Many of the laws defining crimes are imprecise or ambiguous—concepts such as reckless driving, obscenity, and gross negligence may leave a great deal to interpretation. The result may be that conduct which is treated as a crime in one police district, and thus appears as such in statistics, may not be treated as the same crime in another, because the law is interpreted in a different manner. Another practice that may have the same result is the way in which a particular incident is broken down into different crimes. The theft of a number of items may be recorded as a single theft of all of them or as a series of thefts of the individual items.

Criminologists have for many years endeavoured to obtain a more accurate picture of the incidence of crimes and the trends and variations from one period to another. Two research methods have usually been employed—the victim survey and the self-report study. The victim survey requires the researcher to identify a sample of the population at risk of becoming victims of the kind of crime in which the researcher is interested, or of crimes generally, and to ask them to disclose any crime of which they have been victims during the period specified in the research. The information obtained from the survey, after a large number of people have been questioned, is compared with the statistics for reported crime for the same period and locality, giving an indication of the relationship between the actual incidence of the type of crime in question and the number of cases of that type reported to the police. Although criminologists have developed sophisticated procedures for interviewing victim populations, such projects are subject to a number of limitations. Results depend entirely on the victim's recollection of the incidents, his ability to recognize that a crime has been committed, and his willingness to disclose it. The method can be applied only to crimes that have victims; it does not help to identify the incidence of victimless crimes. Research of this kind, however, undertaken with an awareness of the limitations of the process, undoubtedly extends knowledge of the actual incidence of particular types of crime and, if it is carried on over a period of years, may provide a clearer picture of trends in crime than official statistics. One major survey of this kind, the British Crime Survey, is expected to last for many years and, as it obtains information from a very large sample of households, may be more representative than smaller ones.

Victim survey

An alternative approach favoured by some criminologists is the self-report study, in which a sample of the general population is asked, under assurances of confidentiality, if they have committed any offenses of a particular kind. This type of research is subject to the same difficulty as the victim survey—the researcher has no means of verifying the information given to him, and the subject can easily conceal the fact that he has committed an offense at some time—but surveys of this kind have often confirmed that large numbers of offenses have been committed without being reported and that crime is much more widespread than official statistics suggest.

Self-report study

## Analysis of crime

### CHARACTERISTICS OF OFFENDERS

Knowledge of the types of people who commit crimes is subject to one overriding limitation: it is generally based

on studies of those who have been detected, prosecuted, and convicted. The populations of penal institutions are not necessarily representative of the whole range of criminals—in one sense, they are by definition the unsuccessful criminals. Despite this limitation, some basic facts emerge that probably give a reasonably accurate picture of those who commit crimes. The first is that crime is predominantly a male activity. In all criminal populations, whether of offenders passing through the courts or of those sentenced to institutions, men outnumber women by a high proportion. In Britain in 1984, for instance, of 449,000 offenders found guilty of criminal offenses, 387,400 (86 percent) were males; in the same year, the daily average population of the prisons consisted of 41,822 men and 1,473 women. In most Western societies the incidence of recorded crime by women, and the number of women passing through the penal systems, is on the increase; in the United States, for instance, the number of women arrested for property crimes between 1960 and 1976 increased by 276 percent—a significantly higher rate of increase than that exhibited by other groups. A similar trend is shown in English prison statistics: the number of women in prison under sentence rose from 538 in 1974 to 941 in 1984, an increase of 75 percent in 10 years. A number of explanations have been offered for this trend. One suggestion is that it reflects a real trend in the commission of crimes by women—that the changing social role of women, with more women leaving the home and taking employment, expecting and achieving financial independence, leads to greater opportunity for crime and to greater temptation. An alternative explanation is that the change in the apparent rate of female criminality merely reflects a change in the operation of the criminal justice system—that crimes committed by women are less likely than was previously the case to be ignored by law enforcement agencies out of a sense of chivalry. Even though female criminality appears to be increasing faster than male criminality, it will be many years before women reach the same level of crime as men.

'redominance f male riminals

A second aspect of criminality about which there is a reasonable measure of agreement is that crime is predominantly an activity of the young. In both Britain and the United States, for example, the peak period for involvement in relatively minor property crime is adolescence—from 15 to 21. For involvement in more serious crimes the peak age is likely to be rather higher, from the late teenage years through the 20s. Criminality tends to decline steadily after the age of 30. Criminologists have sought explanations of this phenomenon—whether it is a natural effect of aging, the consequence of taking on family responsibilities, or the effect of experiencing penal measures imposed by the law for successive convictions—but the evidence is inconclusive. Not all types of crime are subject to decline with aging. Fraud and certain kinds of theft, as well as crimes requiring a high level of businesslike organization, are more likely to be committed by older men, and sudden crimes of violence, committed for emotional reasons, may occur at any age.

'he youth ıctor

The relationship between social class or economic status and crime has been studied extensively by criminologists. Studies carried out in the United States in the 1920s and '30s claimed to show that a higher incidence of criminality was concentrated in deprived and deteriorating neighbourhoods of large cities, and studies of penal populations revealed that the level of educational and occupational attainments was generally lower than in the wider population. Early studies of juvenile delinquents dealt with by courts disclosed a high proportion of lower-class offenders. Later research has called into question the assumption that criminality is closely associated with social origin; in particular, self-report studies have suggested that offenses are more widespread across the social spectrum than the figures based on identified criminals would suggest.

ocial and conomic tatus

The relationship between racial or ethnic origin and criminality is a difficult and controversial question. Penal populations probably contain a disproportionately high number of persons from some minority racial groups (although precise figures are difficult to obtain), in the sense that the proportion of minority group members in prison is greater than the group's proportion in the general population. Criminologists have pointed out that this may be the result of the high incidence among minority racial groups of characteristics that are commonly associated with identified criminality—*e.g.,* unemployment and low economic status—and the fact that in many cities racial minority groups inhabit areas that have traditionally been high crime areas, perhaps as a result of their shifting populations and general lack of social cohesion. Further explanations are differential enforcement practices on the part of the police and the adherence of members of some minority groups to cultural standards that are in conflict with the general law (*e.g.,* the widespread use of cannabis [marijuana] by members of the predominantly black Rastafarian sect).

## THEORIES OF CAUSATION

Few modern criminologists would claim that any single theory constitutes a universal explanation of criminality or a valid predictor of future criminal behaviour in a particular population. A more common view is that many of the different theories offered may help to explain particular aspects of criminality and that different types of explanation may all contribute to the understanding of the problem of crime.

**Biologic theories.** Some theories attribute the tendency toward criminality to innate biologic factors. The most famous of these is probably that of the Italian Cesare Lombroso (1835–1909), one of the first scientific criminologists, whose theories were related to Darwinian theories of evolution. His investigations of the skulls and facial features of robbers led him to the hypothesis that serious or persistent criminality was associated with atavism, or the reversion to a primitive stage of human development. Another biologic theory related criminality to body types, suggesting that it was more common among muscular, athletic persons (mesomorphs) than among tall, thin persons (ectomorphs) or soft, rounded individuals (endomorphs). These theories have little support today, but there is some interest in the idea that criminality may be related to chromosomal abnormalities—in particular, the idea that so-called XYY males (characterized by the presence of a surplus Y chromosome) may be more likely to be involved in criminal behaviour than the general population.

Lombroso's theory

Some criminologists have endeavoured to answer the question of whether biologic factors are more important than social factors in criminal behaviour by studying the behaviour of twins. Various studies have shown that twins are more likely to exhibit similar tendencies toward criminality if they are identical (monozygotic) than if they are fraternal (dizygotic). The suggestion of genetic influences in criminal behaviour is supported by studies of adopted children carried out to determine the influence of the biologic parent on criminality. One such study showed that the rate of criminality was higher among those adopted children who had one biologic parent who was criminal than among those who had one adoptive parent who was criminal but whose biologic parents were not. The highest rates of criminality were found among those children who had both biologic parents and adoptive parents who were criminal.

Twin studies

**Sociological theories.** Sociologists have proposed a variety of theories that explain criminal behaviour as a normal adaptation to the offender's social environment. One such theory, known as differential association, proposed that all criminal behaviour is learned behaviour and that the process of learning criminal behaviour depends on the extent of the individual's contact with other persons whose behaviour reflects varying standards of legality and morality. The more the individual is exposed to contact with persons whose own behaviour is unlawful, the more likely he is to learn and adopt their values as the basis for his own behaviour. The theory of anomie, proposed by the American Robert K. Merton, suggested that criminality is a result of the offender's inability to attain by socially acceptable means the goals that society expects of him; faced with this inability, the individual is likely to turn to other, not necessarily socially acceptable, objectives or to pursue the original objectives by unacceptable

Theory of anomie

means. A development from this theory is the concept of the subculture—an alternative set of moral values and conventional expectations to which the person can turn if he cannot find acceptable routes to the objectives held out for him by the broader society. This theory, developed particularly with reference to delinquent gangs in U.S. cities, has been disputed by other sociologists who deny the existence of any subculture of delinquency among the lower classes of society; the behaviour of gangs is for these latter sociologists an expression of widespread lower-class values emphasizing toughness and excitement.

A further group of sociological theories denies the existence of subcultural value systems and portrays the delinquent as an individual who subscribes generally to the morals of society but who is able to justify to himself particular forms of delinquent behaviour by a process of "neutralization," in which the behaviour is redefined in moral terms to make it acceptable. Control theory emphasizes the links between the offender and his social group—the individual's bond to society. According to this theory, the ability of the individual to resist the inclination to commit crime—which may be an easy way to satisfy a particular desire—depends on the strength of his attachment to parents, his involvement with conventional activities and avenues of progress, and his commitment to orthodox moral values that prohibit the conduct in question. Labeling theory, by contrast, portrays criminality as a product of the reaction of society to the individual, rather than of his own inclinations and personality. It assumes that the criminal is not substantially different from any other individual, except that he has become involved in the processes of the criminal justice system and has acquired a "criminal" identity. Through a process of rejection by law-abiding persons and acceptance by other delinquents, which is a consequence of the criminal identity conferred on him by the courts, the offender becomes more and more socialized into criminal behaviour patterns and estranged from law-abiding behaviour. Eventually he comes to see himself cast by society into the role of a criminal, and he acts out society's expectations. Each time he passes through the court system, the process is extended to form a process described as "amplification of deviance." Radical criminologists change the focus of inquiry, looking for the causes of delinquency not in the individual but in the structure of society, in particular its political and legal systems. The criminal law is seen as an instrument by which the powerful and affluent maintain their position and coerce the poor into patterns of behaviour that preserve the status quo.

**Psychological theories.** Psychologists have approached the task of explaining delinquent behaviour by examining in particular the processes by which behaviour and restraints on behaviour are learned. Psychoanalytical theories emphasize the instinctual drives for gratification and the control exercised through the more rational aspect of personality, the superego. Criminality is seen to result from the failure of the superego, as a consequence either of its incomplete development or of unusually strong instinctual drives. The empirical basis for such a theory is necessarily thin. Behaviour theory views all behaviour—criminal and otherwise—as learned and thus manipulable by the use of reinforcement and punishment. Social learning theory examines the manner in which behaviour is learned from contacts within the family and other intimate groups, from social contacts outside the family, particularly from peer groups, and from exposure to models of behaviour in the media, particularly television.

Behavioral theory of crime

Mental illness is the cause of a relatively small proportion of crimes, but its importance as a causative factor may be exaggerated by the seriousness of some of the crimes committed by persons suffering from mental disorders. Severe depression or psychopathy (sometimes described as sociopathy or personality disorder) may lead to grave offenses of violence. On a less serious level, depression may lead to theft or other uncharacteristic behaviour. (D.A.T.)

**A non-Western perspective: China.** The Chinese have in general adopted a Marxist interpretation of the causes of crime. Crime is viewed as a product of class society, of exploitative systems founded upon the institution of private property. Because the socialist system is considered by its proponents as incapable of producing crime, official theory has always looked outside of post-1949 Chinese society to find the causes of contemporary crime. A number of specific sources of criminal activity have been suggested: (1) external enemies and remnants of the overthrown reactionary classes (the latter referring to the government of the Republic of China in Taiwan) who infiltrate the country with spies and conduct sabotage; (2) remains of the old (pre-1949) society, such as gangsters and hooligans, who refuse to reform; (3) lingering aspects of bourgeois ideology that prize profit, cunning, selfishness, and decadence and thus encourage crime; and (4) the poverty and cultural backwardness that is seen as the legacy of the old society. The Cultural Revolution (1966–76) has also been cited as a cause of crime; it is said to have confused notions of right and wrong and to have destroyed respect for authority.

While Chinese criminology thus adopts a social explanation of crime in capitalist society, it has little sympathy for the view that society is to blame for crime in contemporary China. The two main causes are seen to be backward thinking and ignorance. For this reason, crime is ideally to be fought, and ultimately eliminated, by thought reform and by education. (D.C.C.)

## Detection of crime

In most countries the detection of crime is the responsibility of the police, although special law enforcement agencies may be responsible for the discovery of particular types of crime (customs departments, for instance, may be responsible for the detection of smuggling and related offenses). Crime detection falls into three distinguishable phases: the discovery that a crime has been committed, the identification of a suspect, and the collection of sufficient evidence to indict the suspect before the court. Criminologists have shown that a high proportion of crimes are discovered and reported by persons other than the police (such as victims or witnesses), but certain types—in particular crimes that may involve a subject's assent, such as dealing in drugs or prostitution, or those in which there may be no identifiable victim, such as obscenity—are often not discovered unless the police take active steps to determine whether these crimes are being committed. This may require controversial methods, such as surveillance, interception of communications, infiltration of gangs, and entrapment (*e.g.*, by making a purchase from a suspected drug dealer). Once the commission of a crime has been discovered, the identification of the suspect becomes essential.

### THE ROLE OF FORENSIC SCIENCE

Forensic science has come to play an increasingly important part in the investigation of serious crimes. One of the first significant developments was identification by fingerprints. It was discovered in the 19th century that almost any contact between a finger and a fixed surface left a latent mark that could be exposed by a variety of procedures, the most common being the use of a fine powder. It was accepted in 1893, by the Troup Committee established by the Home Secretary, that no two individuals had the same fingerprints, and this proposition has never been seriously refuted. Fingerprint evidence was accepted for the first time in an English court in 1902. The original purpose of recording and collecting fingerprints was to establish and to make readily available the criminal record of particular offenders, but fingerprinting is now widely used as a means of identifying the perpetrators of particular offenses. Most major police forces maintain collections of fingerprints taken from known criminals at the time of their conviction, for use in identifying these individuals should they commit later crimes. Fingerprints (which may be incomplete) found at the scene of the crime are matched with fingerprints in the collection. According to the British standard, if the sets of fingerprints share at least 16 characteristics, it is considered virtually certain that they are from the same person. Searching fingerprint collections had historically been a time-consuming manual task, based on various systems of classification,

Fingerprinting

but systems for electronic storage and rapid searching of fingerprint collections were developed and implemented in the 1980s.

A broad range of other scientific techniques is available to law enforcement agencies attempting to identify suspects or to establish beyond doubt the connection between a suspect and the crime in question. Examples include the analysis of bloodstains and traces of other body fluids (such as semen or spittle) that may indicate some of the characteristics of the offender. Fibres can be analyzed by microscopy or chemical analysis to show, for instance, that fibres found on the victim or at the scene of the crime are similar to those in the clothing of the suspect. Hair samples, and particularly skin cells attached to hair roots, can be compared chemically and genetically to those of the suspect. Many inorganic substances, such as glass, paper, and paint, can yield considerable information under microscopic or chemical analysis. Examination of a document in question may reveal it to be a forgery, on the evidence that the paper on which it is written was manufactured by a technique not available at the time to which it allegedly dates. The refractive index of even small particles of glass may be measured to show that a given item or fragment of glass was part of a particular batch manufactured at a particular time and place. Such information may help to identify the kind of automobile involved in a hit-and-run accident. Computer networks allow investigators to search increasingly large bodies of data on material samples, but the creation of the necessary data bases is a lengthy process.

### MODUS OPERANDI AND SUSPECT IDENTIFICATION

The method by which an offense was committed may also help to identify the suspect, as many offenders repeatedly commit offenses in much the same way. The burglar's method of entry into the house, the type of property stolen, or the kind of deception practiced on the victim of a fraud may all suggest to the police who is responsible for the crime. Visual identification of a stranger by the victim is often possible, but experience has shown that such identifications are often mistaken and have frequently led to miscarriages of justice. If the victim or witness believes that he can recognize the offender, the police may show him an album containing photographs of a large number of known criminals, in the hope that one can be picked out. A suspect identified in this way is usually asked to take part in a lineup, in which the witness is asked to pick the suspect out of a group of people with similar characteristics.

### GATHERING EVIDENCE

The identification of the suspect is not the final stage of the process: it is essential that the investigating agency gather sufficient legally admissible evidence to convince the judge or jury that the suspect is guilty before a conviction can be expected. It is common for the police to be reasonably certain that a particular individual is responsible for a crime but to remain unable to establish his guilt by legally admissible evidence. In order to secure the necessary evidence, the police employ a variety of powers and procedures; because these potentially involve interference with the freedom of the suspect (who must at this stage be treated as an innocent person), they are normally subject to close control either by legislation or by the courts.

The use of search warrants

One important procedure is a search of the person of the suspect or of premises or vehicles. Most jurisdictions within the common-law tradition allow a search to be carried out only if there is "probable cause for believing" or "reasonable ground for suspecting" that the evidence will be found. In some cases a person may be stopped on the street and searched, subject to various requirements that the police officer identify himself and state the reasons for the search. A search of private premises usually requires a search warrant issued by a magistrate or judge. The law generally permits a search warrant to be issued only if the issuing authority is satisfied after hearing evidence on oath that there is good reason to suspect that the evidence, which the warrant usually defines specifically, will be found on the premises. The warrant may be subject to time limits and normally permits only one search to be carried out. In most countries the judge or magistrate who issues the warrant must be told of the outcome of the search. Material seized as a result of a search under the authority of a search warrant is usually detained by the police for production as exhibits at any subsequent trial. In the United States the law has imposed strict consequences on any abuse of this procedure; evidence discovered as a result of any search that does not comply with the procedures and standards laid down by the Supreme Court and by other courts, interpreting the various amendments to the U.S. Constitution collectively known as the Bill of Rights, is not admitted in the trial, even though it clearly establishes the guilt of the accused person, and even though the suppression of the evidence may prevent the conviction of a person who is plainly guilty. This rule, known as the exclusionary rule, has given rise to controversy in the United States and has not generally been adopted in other English-speaking countries.

### INTERROGATION AND CONFESSION

The interrogation of suspected persons is an important aspect of the investigation of offenses. Usually the aim of the questioning is to obtain an admission of the offense that will lead eventually to a plea of guilty and avoid the need for a contested trial. All English-language countries place restrictions on the scope and methods of interrogation in order to ensure that suspects are not coerced into confessions by unacceptable means. In the United States any suspect who is being interrogated in custody must be offered the services of a lawyer, at the expense of the state if he cannot afford to pay, and failure to advise the suspect of this right (known as the Miranda warnings, after the case of *Miranda* v. *Arizona*) results in the rejection of a confession as evidence.

The Miranda warnings

English law follows the same general principle, that a person suspected or accused of a criminal offense is not at any stage in the process of investigation or trial obliged to answer any question or to give evidence. (There are a few minor exceptions; for instance, the owner of a motor vehicle is required by law to disclose the identity of the person who was driving the vehicle on any particular occasion, and drivers of motor vehicles may be required to give samples of breath, blood, or urine in certain circumstances.) For many years the law relating to confessions in England consisted of a simple rule prohibiting the admission as evidence at trial of any involuntary statement made by an accused person. This rule was supplemented by more detailed rules governing the questioning of suspected persons by the police, formulated by the judges of the High Court and known as the Judges' Rules. The principal effect of the Judges' Rules was to impose an obligation on the investigating police officer to administer to the suspect a caution to the effect that he was not obliged to answer any question and that anything he did say might be given in evidence at his trial. This caution was required to be given at the beginning of any period of interrogation and immediately before the suspect began to make a full statement or confession. Failure to give the caution at the right time or in the right form did not necessarily mean that the statement would be excluded from evidence, but it did give the trial judge the discretion to exclude the evidence if he considered it just to do so. The operation of the Judges' Rules was a source of controversy for many years, and they have been replaced by a comprehensive series of provisions under the Police and Criminal Evidence Act, 1984. This act provides that a confession by an accused person may be admitted in evidence provided that the court is satisfied that the confession was not obtained by oppression of the person who made it or as a result of anything said or done that was likely to render the confession unreliable. Oppression is defined to include torture, inhuman or degrading treatment, and the use or threat of violence, but there is no doubt that it includes other matters as well (such as excessively prolonged periods of questioning). This broad principle is supplemented by a much more detailed code of practice. For full treatment of trial procedures prior to sentencing, see the articles CRIMINAL LAW and PROCEDURAL LAW.

The Judges' Rules

## Prosecution

In countries whose legal system follows the English tradition, the function of prosecution is usually distinguished from that of investigation on the one hand and adjudication on the other. In most countries (although not in England until recently) the function of prosecution has been given to an official who is not part of either the police or the judicial system; a wide variety of terms are used to designate this official—district attorney in the state jurisdictions of the United States, procurator-fiscal in Scotland, and crown attorney in Canada are examples. The prosecutor may be an elected local official (as in the United States in most cases) or a member of an organization responsible to a minister of the national government. The first tasks of the prosecutor are to assess the information collected by the investigators, to determine whether there is sufficient evidence to justify the institution of criminal proceedings, and to decide whether there are any reasons why the public interest requires that a prosecution should not be undertaken.

The prosecutor

In common-law systems the prosecutor is usually entrusted with extensive discretion in deciding whether to institute criminal proceedings. In part, this discretion arises out of the ambiguity of the criminal law; frequently a statute defining a particular criminal offense does not make absolutely clear what kind of behaviour it is intended to cover or includes a much wider range of circumstances than it was intended to prohibit. If this is so, the prosecutor must decide whether the case he is dealing with falls within what was intended to be the scope of the law. Changing attitudes in the community toward particular kinds of behaviour may mean that a criminal prohibition, while remaining on the statute books, no longer reflects the sentiment of the community, and the prosecutor is no longer expected to bring charges against people who infringe it. In other cases, laws may be enacted without the usual exemptions from responsibility for those who commit the act unintentionally (offenses of strict liability). In such cases the prosecutor may nevertheless feel justified in not bringing proceedings against those who are technically guilty if they are in his view morally innocent.

## The court system

Court systems and procedures reflect the history and culture of the country in which they have developed; there are many variations among different countries, or among different jurisdictions within the same country, regarding the way in which criminal cases are brought to trial.

### CRIMINAL PROCEDURE IN ENGLISH-SPEAKING COUNTRIES

Each state of the United States has its own legal system, and, within the United Kingdom, England and Wales, Scotland, and Northern Ireland all have different arrangements for the conduct and procedure of criminal trials. These countries, however, generally follow what is called "adversarial" procedure, in which allegations are made by the prosecution, resisted by the defendant, and determined by an impartial trier of fact—judge or jury—who is generally required to find in favour of the defendant by acquitting him if there is any significant doubt as to his guilt. English criminal procedure, employing the adversarial method, is the model from which the court systems of many English-language systems have been developed (although Scotland evolved its own distinctively different rules independently); over the years the differences between, for instance, the English criminal courts and those of the typical U.S. state have widened in some aspects, but the same basic principles are still reflected in both countries. The court systems of most English-speaking countries provide two or more sets of criminal procedure, to deal with the more serious and less serious cases, and a further set of procedures for hearing appeals against the decisions of courts of trial. (See also JUDICIAL AND ARBITRATIONAL SYSTEMS.)

Adversarial procedure

**England.** All criminal cases brought to trial in England begin in the magistrates' court. The magistrates' court has a number of different functions to perform—to determine the mode of trial, to try the case if summary trial is chosen, and to deal with ancillary matters such as bail and the granting of legal aid. Although the expression "examining magistrates" is still found in the statutes, the magistrates have long ago lost any function in the investigation of the alleged crime; their function is now wholly concerned with the adjudicatory phase of the process. The police investigation is normally completed by the time the case comes before the magistrates' court for the first time. The magistrates themselves are for the most part lay people chosen for their experience and knowledge of society. All are appointed by the central government on the advice of a committee (known as the Lord Lieutenant's Advisory Committee) for the particular county in which they are to sit. Magistrates, who are required to sit on an average of at least 14 days each year, develop considerable experience in their work, but they cannot be considered professionals. In large cities there are professional, legally qualified magistrates, known as stipendiary magistrates. The stipendiary magistrate can sit on his own, but lay magistrates may sit only as a bench of two or more. Lay magistrates are invariably attended by a legally qualified clerk to advise them on matters of law. The system of lay magistrates has existed in England and Wales since about 1360 and is generally an accepted part of the administration of justice.

Lay magistrates

**The United States.** Criminal procedure in U.S. states follows a pattern derived from English traditions and principles, but with many variations. The lay magistrates play an insignificant role, if any, in the U.S. system, and the prosecutor (the district attorney) is a key courtroom figure. He determines the charges, which in turn may well determine whether the accused appears before a lower court (dealing with misdemeanours) or a higher court (dealing with felonies). The accused is offered bail in almost every case, but he is not released unless he is able to deposit with the court either cash or security in the form of a bond, often posted on his behalf by a bondsman who charges a proportion of the amount of the bond. In some states it is common for an accused person to be released without bond on his own recognizance. The role of the examining magistrates in English criminal procedure may be played in the United States by the grand jury, whose task it is to examine the evidence produced by the prosecutor and, if warranted, to return an indictment. The deliberations and proceedings before the grand jury are normally conducted in private. When the case is brought before the trial court, it is often settled on the basis of a plea bargain made between the prosecutor and the defense lawyer, by which the accused pleads guilty to some of the charges and the prosecutor recommends a sentence that has been agreed upon beforehand. Plea bargaining, which can take many other forms, is more readily accepted in U.S. courts than in English courts as long as basic rules, designed to ensure fair dealing for the accused, are observed. If the case goes to trial before a jury, a major difference between the English and U.S. systems is seen in the procedure for the selection of the jurors. In a U.S. court the lawyers are allowed to question potential jurors about their beliefs and attitudes so as to exclude those who may be prejudiced. The selection of the jury in an important case may take almost as long as the hearing of the evidence. The rules of evidence are much the same as those followed in an English trial, with variations of detail, and the accused is normally represented by a lawyer paid for by the state if he cannot afford one himself. U.S. law allows a wider range of appeals, both within the state system and, if a question of constitutional rights is involved, by removal of the case to a federal court. It may be many years before the case is finally resolved beyond all dispute.

The grand jury

### OTHER SYSTEMS OF CRIMINAL PROCEDURE

**Continental Europe.** The jurisdictions of continental Europe follow methods of criminal procedure very different from those of the English-speaking world. Often described as the inquisitorial method, continental practice emphasizes the role of the judge, who is normally responsible for calling and questioning all witnesses and who does not separate the process into two distinct phases of trial of guilt and sentencing. The tribunal may consist of

Inquisitorial method

several judges, or a combination of judges and lay assessors, who deliberate together on both conviction and sentence. The rules of evidence are generally less restrictive; materials that would be considered hearsay in common-law countries are often admitted, and information about the accused person's record is available to the tribunal. A major difference between the two traditions is that most European jurisdictions do not permit conviction on the basis of a plea of guilty; even though the accused is willing to admit his guilt, the court must investigate the evidence fully (although the admission is part of that evidence). A second major difference is that the decision of the tribunal is normally accompanied by a statement of reasons, which are never given for the verdict of a jury. (D.A.T.)

An institution without parallel in English law is the French unified magistracy, whose members are divided into *assise* ("seated," or the members of the bench) and *parquet,* or *debout* ("standing" or the prosecuting attorneys). It is a state prosecuting system in which the state acts as a party to the prosecution of civil and criminal cases.

**Africa.** Prosecuting and sentencing systems in African countries in general follow those of the former colonial rulers from whom the legal systems are derived. In the common-law countries this means that, although there is everywhere state prosecution, considerable responsibility falls on the police forces to initiate prosecutions. Sentencing is the responsibility of the court that tries the case and convicts the defendant. In some countries, such as The Sudan and parts of Nigeria, where Indian legal influence was strong, versions of the Indian Criminal Procedure Code were adopted, in which the magistrate, rather than the police, takes charge of the investigation and levels charges.

**Islāmic countries.** Among Islāmic countries of English and French colonial heritage, the modern states have adopted the procedure of the colonial countries that ruled them. Pakistan, for instance, which originally inherited the Indian Criminal Procedure Code, now has a procedural system very similar to that of England. It is an accusatorial system in which both sides present their oral arguments to an impartial judge. There is a competent and independent bar from whose ranks judges are chosen. This was amended in 1980 with the introduction of special Islāmic courts and judges. On the other hand, Egypt's criminal system almost exactly mirrors that of France. The system is inquisitorial, and the judge has a much greater power to question and intervene and to determine the method of proceeding. There also exists the Niyaba, a system of state prosecutors very similar to the French *parquet.* Egyptian judges, unlike their English and Pakistani counterparts, are often career judges. In all categories of Islāmic states there are not only ordinary criminal courts but often also police courts, which tend to deal with lesser criminal offenses, and military courts, which hear questions affecting security and military matters. In those states, such as Saudi Arabia and Iran, that claim to adhere totally, or almost so, to traditional Islāmic law, Islāmic judges, called *qāḍīs,* exercise jurisdiction in Islāmic courts.

**Soviet Union.** Criminal procedure in the Soviet Union is based upon the inquisitorial model developed in continental Europe. Cases are tried by a court staffed with a professional judge and two lay assessors, who jointly and by majority vote render a verdict and decide the sentence. Soviet trials follow an extensive preliminary investigation during which the investigator (most often a lawyer working in the procuracy) examines witnesses and compiles a written analysis of the evidence. The record of this investigation is introduced as evidence at the trial, and the judge decides whether or not to summon witnesses for further testimony in the courtroom. The defense counsel have no access to their clients during the pretrial proceedings. Like defense attorneys in France until 1897 and Italy until 1945, they enter cases only at the end of the preliminary investigation.

At most trials the issue of guilt or innocence is not at question, and the proceedings focus upon the factors relevant to sentencing, such as the nature and circumstances of the crime and the character and background of the accused. On those occasions when the defense casts doubt upon the evidence needed to convict, judges tend not to acquit the accused but to return the case to the procuracy for "supplementary investigation." Unless new evidence is found, the procurator may drop the case.

**China.** The Chinese penal system broadly divides procedures and sanctions into criminal and administrative. "Crimes" are in China distinguished from "ordinary illegal acts." Crime is only that behaviour that is punishable by a court under the Criminal Law or other laws calling specifically for criminal punishment for violators. Ordinary illegal acts, however, can be punished administratively by nonjudicial bodies (such as the police) on their own initiative and according to their own less formal procedures. In general, administrative punishments cannot be appealed to a court. Milder sanctions, such as group criticism, may be imposed by neighbourhood-level organs of local government. Thus, disruptions of social order can meet with a response by the state well before they reach a level called criminal.

The concept of "circumstances" is of crucial importance to criminal procedure in China. Circumstances might, for example, mean the identity of the accused or the victim, the existence of an official campaign against the particular type of crime involved, or such matters as whether a robber also beat his victim or whether the accused showed repentance. It is common in many countries for such factors to be taken into account in sentencing. Chinese law, however, differs in its provision for allowing circumstances also to bring an act within or entirely outside the coverage of the Criminal Law and, more importantly, the associated Criminal Procedure Law. It is only the latter, for example, that provides for a public trial by a court and the right to a defense. A person suspected of selling pornographic books may, if the police deem the circumstances clearly minor, be judged by the police and punished by up to 15 days' detention in a police station. In such cases there is no right to a defense and no appeal to a court, but the maximum punishment that can be imposed is much less than would be possible were the accused to be prosecuted under the Criminal Law. (A.N.A./I.D.E./D.C.C.)

## Sentencing

In countries following the Anglo-American legal tradition, sentencing is a function that is distinguished from that of determining guilt or innocence and is normally the responsibility of the judge rather than of the jury, although in some parts of the United States the jury is empowered to determine the sentence. Most such systems of law traditionally give the judge a wide discretion in determining both the kind of penalty to be imposed (imprisonment, fine, probation) and its extent. As modern sentencing systems provide an increasingly wide range of forms of sentence, the choice of sentence becomes a more complex task. The extensive discretion involved in sentencing and the wide variety of different forms of sentence mean that in many cases there are complaints of disparity in the sentences passed on different offenders and of arbitrariness and idiosyncrasy in the decisions of individual judges; it is sometimes said that the sentence imposed on an offender may depend more on the judge before whom the offender appears than on the gravity of his offense or his record.

The problem of disparity

It has long been recognized that the quality of the decisions made by judges in sentencing depends on the information available to them. In a case in which there has been a contested trial, the judge will have heard all of the evidence related to the immediate background of the offense but will not necessarily know much about the background of the offender. This gap is filled in many jurisdictions by a report (a presentence report or social inquiry report) prepared by a probation officer and submitted to the court after the offender has been convicted or has pleaded guilty.

## Parole

The word *parole* in French means "word," and its use in connection with the release of prisoners was derived from the idea that they were released on their word of

honour. The practice of allowing prisoners to be released from prison before serving the sentence of imprisonment pronounced by the court goes back at least to 18th-century England. At that time almost all serious crimes (felonies) were punishable with death, but only a small proportion of those who were convicted of felonies were actually executed. The majority of those who were sentenced to death were pardoned by the king, but their pardon was granted on the condition that they consent to be transported to one of the colonies where labour was required—during the 17th and 18th centuries this was America and, following American independence, Australia. Eventually the courts were given power to pronounce sentences of transportation themselves, usually for a period specified in the sentence, but most sentences of transportation were modified by executive action. In particular, there developed the system of "ticket of leave," under which a convict detained under a sentence of transportation was allowed a measure of freedom, or the right to return to England, in return for good behaviour. When the sentence of transportation was abolished in the mid-19th century, the sentence that replaced it in English law, penal servitude, incorporated the same procedure under a different name, release on license. The prisoner sentenced to penal servitude could earn his release from the penitentiary, but not from the shadow of the sentence, by his good behaviour in custody. His release was decided by the executive government and was conditional on good behaviour outside prison; if another offense was committed, the prisoner could be returned to prison to serve out the rest of the sentence (known as the remanet). In England the system of release from sentences of penal servitude became almost inflexible by the later years of the 19th century, with the result that all prisoners serving the sentence were released after serving a fixed and predetermined portion of it; in the United States at that time, however, the principle of the indeterminate sentence became widely accepted and eventually formed the basis of the sentencing laws of many states.

Sentences of transportation

Indeterminate sentence

In those states where the indeterminate sentence was adopted, the law required a judge who decided to sentence an offender to a term of imprisonment to fix maximum and minimum limits of confinement; the actual date of his release, and the conditions, were then decided by an executive body usually known as the parole board, which had power also to revoke the offender's parole and return him to prison. The indeterminate sentence was seen to have a number of advantages over the more rigid form of sentence, in which the prisoner could work out his exact date of release from the moment he was sentenced. The indeterminate sentence allowed the authorities to observe the behaviour and attitudes of the offender while he was serving his sentence, and in particular the way in which these changed for the better; it provided an incentive to the prisoner to improve, in order to convince the authorities that he was ready for release. In addition to contributing in this way to the rehabilitation of the offender, the indeterminate sentence had a number of administrative advantages to the prison authorities. It provided a powerful sanction against misbehaviour—a prisoner who was violent or disruptive in prison knew that he risked losing the chance of release; it allowed the authorities to compensate for disparities in the sentences imposed by judges (often believed to be a source of friction and discontent among prisoners); and it provided a means by which the population of the prisons could be kept within limits.

Parole supervision

An essential feature of parole is the supervision of the offender during the remaining part of the sentence after his release from prison. A prisoner who has been released on license is not free from all restrictions; he is normally required to observe various conditions, which may be quite restrictive, dealing with such matters as where he lives and works or requiring him to undergo medical or psychiatric treatment. Failure to comply with these conditions can lead to the revocation of the parole, which means that the offender is returned to prison to serve out the remainder of the sentence. Enforcement of the conditions, as well as the provision of help and counseling, is usually the responsibility of a probation or parole officer, to whom the paroled offender is required to report at stated intervals and who may have considerable power over the offender. (The license may include a condition requiring the offender to live at a place approved by the probation or parole officer, for instance, or require the officer's consent to a change of employment.) In many countries the decision of the supervising officer or the parole board to terminate the offender's license and require him to return to prison are not subject to appeal or judicial review, even though the consequences for the offender may be serious.

## Theories and objectives of punishment

The proper objectives of a system of punishment administered by the official organs of the state have been the subject of debate among philosophers, lawyers, and legislators for centuries. A variety of different theories or objectives of punishment have been proposed, some differing only in minor degrees, some fundamentally in conflict with each other. The debate has often been confused by the fact that the same expression is frequently used to denote different ideas, which are not always clearly distinguished from one another; conversely, the same idea may be described by different names, resulting in increased confusion.

### RETRIBUTION

In modern judicial systems the term retribution has acquired various shades of meaning. The key principle that all theories of retribution share is that there should be relation between the gravity of the crime and the severity of the punishment.

One theory of retribution proposes that punishment is not imposed in order to achieve a social objective (such as law-abiding behaviour in the future by the offender or others who witness his example) but is rather an end in itself. Those who hold this view maintain that punishment does not require justification by intended results or effects. Retribution in this sense does not necessarily imply severity of punishment.

Punishment as an end in itself

According to a second theory of retribution, punishment must be justifiable in relation to the gravity of the offense itself, even though there are other reasons (such as deterrence or treatment) that indicate a severer penalty than the gravity of the offense itself would warrant. Some commentators divide this limiting principle of retribution into two subdivisions. One is the principle that punishment should not be inflicted unless the recipient of the punishment has been found guilty of an offense, and the other is that the offender should not be punished more severely than his offense warrants. The first subdivision prohibits such practices as collective punishment imposed on whole communities or the taking of hostages from the general population, as has been practiced by occupying forces at various times. It also requires that the proper forms of the law be observed before punishment is inflicted. The second limb of the principle assumes that some scale can be drawn equating particular crimes with particular punishments. This is extremely difficult to do without resorting to a crude system of inflicting on the offender exactly the damage he has inflicted on the victim (which is in any event impossible in relation to many modern crimes, which have no specific victim). All that can be done is to draw up a conventional scale relating penalties to offenses; this may allow the severer penalties to be preserved for the graver offenses, but it will not in itself justify the relationship between any particular penalty and any particular offense, except by reference to the conventional values of the scale.

Retribution as a limiting principle can be distinguished from retribution as an educational principle, in that in the latter case enactment and implementation of the criminal law, and particularly the imposition of sentences, has the effect of providing a concrete example of society's values, which serves to reinforce those values among those who hold them and to instill them in those who do not. The member of the community who sees his moral values expressed in the judgment of the court in a particular case may feel more strongly committed to them than previously; if he sees them ignored by the court, in the lenient treatment of an offender whose behaviour has violated the

Retribution as an educational principle

fundamental moral principles of society, he may come to question them himself and possibly feel less constrained by them. The principle assumes that a repeated failure of the courts to express such values would lead eventually to moral decline and the dissolution of society.

This aspect of retribution must be distinguished from yet another—the idea that the official organs of the state must punish offenders in order to satisfy the demand for punishment that is natural among members of the community, particularly among those who are the victims of the crime and who in the absence of official punishment administered by the state are likely to take the law into their own hands and seek revenge by direct violence.

Punishment as expiation

A variation on this interpretation of retribution is the concept of expiation—the idea that the offender should undergo punishment in his own interests, to discharge his guilt and to make himself acceptable to society again.

DETERRENCE

Like retribution, deterrence is a complex concept that informs several related theories of punishment. Criminologists distinguish between general and individual deterrence.

A general deterrent is a punishment the object of which is to deter other persons from following the example of the offender, by fear of the same consequences that have been inflicted on him. The theory of general deterrence is not concerned with the future behaviour of the offender himself. It assumes that most crimes are rational and that potential offenders will calculate the risk of being caught, prosecuted, and sentenced in the same manner. Demonstrating the validity of the theory has proved difficult; general trends in crime and their relationship to particular sentencing policies (such as changes in the incidence of a particular crime after it has been treated more or less severely) are seldom a true indication of the effect of penalties as deterrents, as many other factors may be at work. Occasionally it is claimed that particular sentences have had a strong deterrent effect, but usually closer inquiry reveals that this is open to doubt. One example of effective general deterrence is legislation designed to curtail driving after drinking alcohol; studies have suggested that mandatory penalties and a high probability of conviction of those detected have at least a temporary deterrent effect on a wide population.

Individual deterrence

Individual deterrence, unlike general deterrence, is aimed at the particular individual who is punished; the object is to teach him not to repeat the behaviour. This is the rationale of much informal punishment, such as is inflicted on children by parents as a part of their upbringing, as well as of formal sanctions administered by the authority of law. The effectiveness of this type of deterrence can be measured, at least in theory, by examining the conduct of the offender after the administration of the punishment to determine whether he has committed the offense again. Such studies can often be misleading, however, as in practice the only basis for determining that the offender has repeated the offense is a further conviction before a court. Because a high proportion of crimes do not result in convictions, many of those offenders who are not reconvicted after being punished in a particular way may have again committed offenses but avoided conviction.

Denunciation

A third sense of deterrence used by some writers coincides exactly with the idea of retribution as a form of moral education for the community as a whole, sometimes described as denunciation. Although this idea is closely associated with general deterrence through fear, and many sentences of the courts are intended to achieve both objectives simultaneously, there is an important difference between them. The idea of education through retribution, or denunciation, is aimed at the law-abiding person who is not tempted to commit crime; its object is to reinforce him in his rejection of lawbreaking behaviour of the kind in question. Most people do not steal because they know that stealing is dishonest and they consider themselves honest; a sentence on a thief reinforces them in this view. General deterrence through fear is aimed at those who do not necessarily reject the possibility of law-breaking behaviour on moral grounds but who do so on the basis of a careful calculation of the gains and risks involved. Those who are prepared to consider stealing if they thought they could get away with it are frightened off by the example of the thief who is punished.

The principle of proportionality

Deterrence shares with retribution the idea that punishments should be related in severity to the gravity of the crime. The principle of proportionality is central to the idea of deterrence, on practical grounds. If all punishments are the same, irrespective of the gravity of the crime, there may be no incentive to commit the lesser rather than the greater offense. The offender might just as well use violence against the victim of his theft, if the penalty for robbery is no severer than that for simple stealing.

INCAPACITATION

Incapacitation is an object of punishment that has been known since early times. The idea is simply that the offender should be dealt with in a manner that will make it impossible for him to repeat his offense—by execution or banishment in earlier times, in more modern times by execution or lengthy periods of incarceration. This is the only objective of punishment for which there is any certainty (although an offender incarcerated for this reason may escape or commit crimes within the prison). The difficulty arises in reconciling the idea with other principles, in particular those limiting retribution. In practice, it is limited to offenders who have committed crimes repeatedly (multiple recidivists) under what are known as habitual offender statutes (which permit courts to impose longer sentences on such offenders than are normally authorized for the offense) or to offenders who are designated as dangerous, in that there is reason to suppose them likely to commit grave crimes of violence in the future unless restrained. Because there is great difficulty in identifying with any certainty those offenders who are dangerous in this sense, the principle is controversial.

Reserved for habitual or dangerous offenders

REHABILITATION

The most recently formulated theory of punishment is rehabilitation—the idea that through treatment and training the offender should be rendered capable of returning to society and functioning as a law-abiding member of the community. This idea began to establish itself in legal practice in the 19th century. Although seen by many as a humane improvement on former practices, it did not always result in the offender's receiving a more lenient penalty than a retributive or deterrent philosophy would have given him. For many offenders, rehabilitation meant release on probation under some form of condition instead of a period in prison; for others it meant a longer period in custody undergoing treatment or training than would have been acceptable if it had been designated as punishment. One expression of the concept of rehabilitation was the indeterminate sentence popular in many U.S. jurisdictions, under which the length of detention was governed by the degree of reform exhibited.

Criticisms of the theory

Beginning in the 1970s the concept of rehabilitation came under considerable criticism, and it is no longer as widely accepted as previously. The reasons for this growing skepticism are, first, the failure of criminologists to demonstrate that rehabilitation can be achieved in any systematic way. A second objection to rehabilitation as a theoretical basis for penal treatment is that sentences based on the concept typically give too much authority to the administrator, who may be empowered to decide to release or to continue to detain the offender, depending on his assessment of the offender's progress, which may itself be a vaguely defined measure. There have been cases in which this has led to gross abuse and the detention of offenders guilty only of minor crimes for long periods, out of all proportion to the gravity of their offenses, simply because of their inability or refusal to accept or adopt a subservient attitude to those in authority. A more fundamental objection to the concept of rehabilitation rests on a challenge to the criminological and political assumptions on which it is based. The idea that the offender can be successfully treated assumes that he is in some respect inadequate to comply with the legitimate demands of society on its citizens and is in need of additional training to supplement the deficiencies of his upbringing,

Criminality as response to societal defects

education, or personality. Modern criminological theories that challenge this interpretation and portray criminal behaviour as a legitimate or at least predictable reaction to structural defects in society undermine the basis on which the concept of treatment is founded; this is reinforced by the similarities seen between the treatment of delinquents and that of political dissidents in some countries. Some modern criminologists assert "the right to be different" and question the right of society to seek to change the values and ideas of individuals who choose to behave in a manner that brings them into conflict with the law.

### THEORIES IN CONFLICT

It is obvious that, in the practical operation of a sentencing or penal system, these different theories often come into conflict. A lenient sentence (such as probation), designed to rehabilitate the offender, may fail to express society's rejection of his behaviour or to provide an effective deterrent to others. A sentence that requires the offender to submit to a compulsory program of treatment or training for a long period may in turn conflict with the idea of retribution as a limiting principle. A sentence of unusual severity, designed to make an example of the offender as a warning to others, is in conflict both with the principle of rehabilitation and with that of proportionality. A sentence whose object is incapacitation may fail to satisfy those who believe in rehabilitation or in proportionality. In practice the daily operation of any system of sentencing requires decision makers to choose between these different theories in different cases; no single theory provides a system suitable for all cases. The choice between them cannot yet be made on scientific grounds, and it may well be that criminology will never provide the information on which to base a scientific choice between the different objectives of punishment, if only because some of them rest on moral principles rather than on a supposed empirical effect. (D.A.T.)

### PUNISHMENT IN OTHER SYSTEMS

**Africa.** Sentencing courts in Africa stress the punitive and deterrent aspects of sentencing rather than the reformative. In consequence, sentences of imprisonment are often proportionately longer than is usual in Europe, for example, and fines are heavier. The legislation of African countries reflects the same approach to sentencing; capital punishment and in many cases corporal punishment are permitted and in some cases may be mandatory. A notable feature of the current legislation in Tanzania is the imposition of minimum sentences for a variety of offenses, including dishonesty and theft of stock; the court in practice has no option but to impose a specified minimum prison sentence, which from 1963 to 1972 was to commence and end with 12 strokes of the cane.

**Islāmic countries.** Traditional Islāmic law divides crimes into two main categories. Five so-called *ḥadd* crimes are specifically mentioned, along with their appropriate penalties, in the Qur'ān (see ISLĀM, MUḤAMMAD AND THE RELIGION OF: *Islāmic law, Sharī'ah*). All other crimes are called *ta'zīr* crimes, and the punishment of these is left to the discretion of the *qāḍī*, although the books of Sharī'ah law limit his discretion to certain traditional punishments. In Sharī'ah there is found very little power to impose fines. The general punishments are imprisonment or corporal punishment. The traditional requirement of eyewitnesses to crimes considerably limited the application of the severest penalties. The imposition of fines is a recent innovation.

**Soviet Union.** Under the influence of the humanist strain in the socialist tradition, the system of punishment established in the Soviet Union in the 1920s made the rehabilitation of the offender the central purpose of punishment. But during the Stalinist era (1929–53) deterrence and the traditional Russian emphasis on the economic value of convict labour became dominant factors in shaping the penal order. In the years since Joseph Stalin's death in 1953 punishment in the Soviet Union has reflected all three of these concerns.

Overall, the punishments meted out by Soviet judges beginning in the 1960s strike the Western observer as severe. Half of all offenders received custodial sanctions, and their terms of confinement were measured in years rather than in months. In this respect Soviet penal practice resembles that of the United States more than that of western Europe. (See also below *The death penalty.*)

Labour camps

In the 1970s a shortage of unskilled workers available for construction projects in remote parts of the country (*e.g.*, laying railroads, erecting oil rigs) led to a renewed emphasis on the use of convict labour. The introduction of a new penal sanction, "suspended sentence with compulsory labour assignment," and a new form of release, "early release with compulsory labour assignment," provided authorities with means for dispatching convicts to work sites where corrective labour colonies did not exist. The new sanction bore more than a superficial resemblance to the tsarist penalty "exile and hard labour," but for most of those convicts sent to construction sites the assignment represented an alternative to time spent in custody and therefore an improvement in their lot. The new form of early release was soon applied to the majority of inmates before they became eligible for parole.

**China.** The chief goal of criminal punishment in China is reform. Secondary goals are specific deterrence (deterring the offender from repeating his crime) and general deterrence (deterring other would-be criminals).

An authoritative Chinese textbook on criminal law states that the goal of reform in criminal punishment is founded upon the historical mission of the proletariat to reform society and mankind. The thoughts of citizens are not their own affair; the government has the right and the duty to see to it that all members of society become "new men." The commission of a criminal act is, in a sense, evidence that the offender is in particular need of reform and hence justifies the use of particularly coercive measures. The notion that an offender incurs a debt to society that can be paid merely by serving a prison term is alien to Chinese penology. The state is keenly interested in changes in the offender's thinking during imprisonment. Thus, reform through labour and political study generally accompanies imprisonment for criminal offenses.

The primacy of reform over deterrence is intimately connected with Chinese theories on the causes of crime. Chinese criminology holds that crime can be reduced and eventually eliminated through thought reform, education, and the perfection of socialist society. Criminal punishment is seen as merely a supplementary means to this end. According to this view, when the thought of all members of society has been reformed, there will be no more crime. (A.N.A./I.D.E./P.H.S./D.C.C.)

## Treatment of juvenile offenders

Early common law made no special provision for children who committed crimes. Provided that the child was over the minimum age for criminal responsibility—originally seven—and had "mischievous discretion"—the ability to tell right from wrong—the child was fully liable as an adult to the penalties provided by the law. During the 19th century, children who were liable criminally were imprisoned, and there are records of children being hanged as late as 1708. The need for special treatment of juvenile offenders was first recognized during the 19th century in the reformatory movement, the purpose of which was to establish institutions for young offenders, based on training, as an alternative to confinement in adult prisons. The idea of a special court system for juvenile offenders, the juvenile court, began to gain ground in the early years of the 20th century. Juvenile courts were established in England under legislation enacted in 1908, and most U.S. states had juvenile courts by the 1920s. Although there are significant differences between the concepts of the juvenile courts in the two countries, in both, the juvenile court deals with criminal and noncriminal cases. The English juvenile court is essentially a magistrates' court, exercising the ordinary criminal jurisdiction of the magistrates' court over a limited age group of offenders—from 10 years (the minimum age of criminal responsibility) to 17. (Those under 14 are designated as "children"; those over 14 and under 17 are "young persons.") Offenders aged 17 and

Reformatory movement

over appear in the normal adult courts, although special sentencing provisions apply to offenders under the age of 21.

The main difference between a juvenile court and an adult court in England is that the juvenile court has a much wider jurisdiction in terms of the offenses it can try. It can deal with a juvenile for any offense except homicide, although it is not bound to deal with a young person for a serious offense such as robbery or rape; on such a charge he can be committed to the Crown Court for trial in the same manner as an adult. A child may be committed to the Crown Court for trial on a charge of murder or manslaughter. A juvenile may also be sent to an adult court—magistrates' court or Crown Court—for trial, if charged jointly with an adult (as, for example, in the case of a parent and child jointly charged with shoplifting). In such cases the adult court normally returns the juvenile to the juvenile court for sentencing. In addition to its criminal jurisdiction, the juvenile court may deal with children of any age up to 17 in what are called care proceedings—proceedings that are based on the idea that the child is in need of care, protection, or control because one of a number of conditions is satisfied. These conditions include neglect or assault by parents but also include the fact that the juvenile has committed an offense (the "offense condition"). A juvenile who commits an offense can thus come before the juvenile court either in criminal proceedings or in care proceedings, although the court may not take action in care proceedings unless satisfied that the juvenile is in need of care, protection, or control; the fact that an offense has been committed is not in itself sufficient. This combination of two different roles in the juvenile court has been a source of difficulty and controversy for many years, particularly because the court in its criminal jurisdiction is required by law to "have regard to the welfare of the child or young person" and, if satisfied that it is necessary to do so, to remove him from unsatisfactory surroundings for his own good, irrespective of the gravity of his offense. A juvenile who appears before the juvenile court charged with a minor offense, if found to be in need of care or control on the basis of inquiries into his background, can therefore be removed from the care of his parents and possibly be required to reside in an institution (known as a community home), perhaps for a period of several years, and possibly under conditions of security. Proposals to change the image of the juvenile court so that its criminal aspect declined in importance and its welfare aspect was emphasized were enacted in 1969 but never fully implemented. The dual role of the juvenile court has thus been retained, but its authority has to some extent been transferred to an administrative body, the local authority. If the juvenile court makes a care order—in practice the most powerful sanction it has available—its effect is to transfer to the local authority the parental rights over the child, and it is for the local authority to decide whether to allow the child to live at home with his parents (but subject to local authority control), to board the child with foster parents, or to require the child to live in a community home. The court has only a limited degree of control over this decision.

Care proceedings

The care order is only one of the sanctions available to the English juvenile court and is used only in a minority of the cases that come before it. Another measure, the supervision order, places the juvenile under the general supervision of a social worker but may require him to take part in a wide range of organized, constructive activities as intermediate treatment. A supervision order may also include restrictive requirements prohibiting the juvenile from certain activities or a curfew in the form of a "night restriction," requiring him to remain at home during the evening for a specified period. Juveniles may also be fined (although the court must usually order the parent to pay the fine) or be ordered to pay compensation, as in the case of an adult.

Intermediate treatment

In U.S. jurisdictions, as in England, a high proportion of juvenile offenders are dealt with informally by means of cautions or counseling. Where a case is brought to court, the procedure is distinctively different from that of a criminal court. The court is not normally concerned with determining guilt or innocence so much as with making a finding of delinquency, which may be the basis for a disposition—either freedom in the community under supervision or confinement in a correctional facility for young people. In keeping with what was seen as the juvenile court's role as a welfare tribunal rather than a court of criminal jurisdiction, procedural standards in the United States were formerly rather elastic, but a series of decisions of the Supreme Court established that the basic rights granted by the Bill of Rights apply to proceedings in the juvenile court (although there is no right to jury trial). Most U.S. juvenile courts, like English courts, deal with cases of neglect as well as criminal cases and a further category of "status offenses"—behaviour amounting to an offense only when committed by a juvenile (such as running away from home), which in England would fall within the scope of care proceedings.

Status offenses

## Prisons

The idea of imprisonment as a form of punishment is relatively modern. Until the late 18th century, prisons were used primarily for the confinement of debtors who could not pay, of accused persons waiting to be tried, and of those convicted persons waiting for their sentences—death or transportation—to be put into effect. Although imprisonment was a sentence available to the courts in cases of misdemeanour, these were a small proportion of the cases tried, and the normal task of the assize courts in England was expressed in the fact that they were known as courts of "general gaol delivery"—delivery meaning the release of prisoners from the jails (to liberty or execution) rather than their commitment to the jails. The holding of accused persons awaiting trial remains an important function of prisons—in England about 20 percent of the prison population is unconvicted or unsentenced, while in some European countries (notably Italy) the proportion of the prison population that consists of unconvicted prisoners may be as high as 80 percent but since the late 18th century, with the decline of capital punishment, the prison has come to be used also as a place of punishment. The concept of the penitentiary was advocated in England during this period by Jeremy Bentham; at the same time in the United States penitentiaries were created first in Pennsylvania and then in New York. By the late 19th century the decline in the use of the death penalty and the abolition of transportation meant that a sentence of confinement (under a variety of names—in England it was known until 1948 as penal servitude) had become the principal sanction for most serious crimes.

### DEVELOPMENT OF THE PENITENTIARY

The development of the penitentiary in the late 18th century was in part a reaction to the conditions of the jails of the period. Sanitation in English prisons at this time was such that disease was widespread among prisoners, who were generally held without any segregation according to sex or classification; outbreaks of "jail fever" occasionally killed not only the prisoners but also the jailers, and even on occasions the judges and lawyers involved in their trials. At this time many prisoners in England were confined not in buildings but in the hulks of ships moored in the Thames River and elsewhere; in theory they were waiting to sail for Australia under a sentence of transportation, but in practice many of them served their sentences in the hulks and were released without ever leaving the country. The appalling conditions in the many local prisons of late 18th-century England were exposed by John Howard, who had traveled throughout Britain, and later throughout Europe, for the preparation of his book *The State of the Prisons in England and Wales* (1777). Reaction against the gross severity of the former penal system of death and transportation and against the physical conditions of the jails led in England to the building of "convict prisons" by the central government; the local jails remained under the control of local authorities until 1877. In that year the whole prison system of England and Wales was brought under central government control, to be administered by a body known as the Prison Commission.

"Jail fever"

Administration

In the United States the prison system is more complex; offenders who are sentenced by federal courts for crimes against the federal criminal code serve their sentences in federal penitentiaries managed by the federal government, but the majority of offenders who are in custody are in state or local institutions, which form part of the penal system of the particular state. This consists of one or more state penitentiaries, possibly supplemented by a number of institutions offering a lower degree of security, such as prison camps or farms, and local jails, each managed and financed by the local jurisdiction in which it is situated. The principal function of the jail is still that of holding persons awaiting trial, but usually short sentences (less than 12 months) are served in the local jail rather than the state penitentiary.

### THE PRISON POPULATION

Concern over prison conditions has not diminished over the years. Problems of security and the protection of prisoners from violence on the part of other prisoners have been compounded by the difficulties arising from overcrowding, as prison populations in most countries continue to grow. Increasing prison populations have been a common feature of most industrialized societies in the era since World War II. In England in 1880 the prison population stood at 32,000; as the prisons came under central government control, there began a long period of decline, probably the result of changes in sentencing laws and practices. By the end of World War I the daily average prison population had declined to about 10,000. It remained stable during the interwar years, rising and falling slightly from one year to the next, but after World War II there began a period of steady increase that has continued unabated. From a daily average figure of about 12,000 in 1945, the prison population grew relentlessly each year, despite a variety of changes in the law designed to contain it. By the 1960s it had reached 30,000—the level of the 1880s—and by 1976, despite the introduction of a parole system and suspension of sentences of imprisonment, the figure of 40,000 was exceeded. Since then it has never dropped below that figure, and by 1984 the population touched 45,000 on occasions. A similar situation has occurred in the United States. A total of 250,000 persons were incarcerated in 1975; by the end of 1980 this had grown to 315,000.

Increase in prosecution

The rise in prison populations is attributed to a variety of factors, but probably the most significant is the rise in the incidence of reported crime and in the number of offenders brought before the courts. In England, both reported crime and the number of persons convicted have risen faster than the prison population. One country that has followed a different trend in the prison population is The Netherlands, where the prison population was halved from 1950 to 1975. This reduction has been attributed by criminologists primarily to a shortening of sentences passed by courts in The Netherlands for common offenses.

The people who make up the populations of most prison systems have many characteristics in common. They are predominantly male—in England males outnumber females by 28 to 1 (although the number of women in prison is rising at a higher rate than the number of men)—and relatively young—nearly 70 percent of those in custody are under the age of 30. Most offenders in prison have a number of previous convictions; the offenses they have committed are most commonly burglary, theft, violence, or robbery. A similar picture is revealed by U.S. statistics; the most common offenses for which prisoners are in custody are burglary and robbery.

Types of institution

Prisoners are distributed among a variety of types of institution. In the United States most prisoners serving longer sentences are held in state prisons, which are usually large maximum-security buildings holding more than 1,000 offenders in conditions of strict security. Young offenders usually are detained in separate institutions, often designated under names that imply that their purpose is treatment or correction rather than punishment. Women are normally held in separate institutions. Prisoners who are not considered a danger to the community may be confined in low-security or open prisons.

In England, as in the United States, most prisoners are held in prisons constructed more than a century ago. Prisons are classified administratively as local or central prisons. Local prisons serve a variety of purposes—holding prisoners awaiting trial or sentencing and prisoners serving shorter sentences (up to about 18 months). There the worst overcrowding occurs. Prisoners serving longer sentences are detained in central prisons, dealing exclusively with similar cases. For security, prisoners are classified into four categories, from A (prisoners likely to attempt escape, and constituting, if successful, a significant danger to the public) to D (prisoners who can be trusted to work in conditions of minimal security). Central prisons cover a range from maximum-security institutions to medium-security prisons, where the degree of security is less intense; and to open prisons, where physical security is minimal and there is normally no obstacle to a prisoner's absconding. In some European countries a further category of institution is available to accommodate prisoners who are allowed to serve their sentences intermittently, usually over a series of weekends. Younger offenders in England (in the age group 15–21) were until recently held in Borstal institutions, named after the village in Kent where the first one was operated. For many years these institutions were admired as an example of practical rehabilitation through training, but declining enthusiasm for this concept, and disillusionment with its effectiveness, led to its replacement with that of "youth custody." Another feature of the English prison system is the detention centre. These institutions for young males serving sentences that must not exceed four months are based on the principle of vigorous discipline and physical activity, popularly known as the "short sharp shock"; research has failed to show, however, that it is an effective deterrent to further crime.

Borstal institutions

Prisons have been described as total institutions, in which every aspect of life is subject to control. In addition to daily routines such as mealtimes, times of rising and retiring, and bathing, many other aspects of the prisoner's life are subject to control. In part this control forms the deprivation of freedom that is the essence of imprisonment, and in part it is a necessary adjunct as a means of maintaining security, controlling the introduction of weapons or contraband substances, and preventing escapes. Most prisons limit the number of visits that a prisoner may receive from his family or friends. In England the Prison Rules allow a convicted prisoner one visit every four weeks, although the prison governor may increase or limit visits at his discretion. Only relatives and friends of the prisoner may visit him, although adequate facilities must be available for visits by legal advisers if the prisoner is engaged in any litigation (for instance, divorce proceedings). Visits normally take place within the sight of an officer, and in some cases within his hearing. In many prisons, visits are conducted with the prisoner sitting on one side of a table and his visitor on the other, with a wire mesh partition between them; the visitor may be searched for contraband. In other prisons the conditions for visiting may be less restrictive—the visitor and the prisoner may be allowed to meet in a room without any physical barrier but still in the sight of officers. Conjugal visits (in which the prisoner's spouse comes to stay with the prisoner for a period of several days) are not permitted in England, but some U.S. states do permit them. Correspondence of prisoners in England is subject to censorship by the prison authorities, and prisoners may not write more than one letter each week.

Visitation privileges

Disciplinary sanctions

Control of the prison is maintained by a number of disciplinary sanctions, which may include forfeiture of privileges, confinement within a punishment block or cell, or the loss of remission or good time (time deducted from the sentence as a reward for good behaviour). The procedures for the imposition of sanctions on prisoners have been improved in both England and the United States, in part as a result of actions taken through the courts. Generally, prisons are governed by rules setting out a code of conduct and listing prohibited behaviour; the code must be given to the prisoner on his arrival in the prison. Typically, the prohibited offenses include mutiny and violence to officers; escaping, or being absent from a place where

the prisoner is required to be; and possessing unauthorized articles. The rules may also include one or more generally defined offenses (such as the English "offence against good order and discipline") that leave much scope for interpretation. Disciplinary sanction may be imposed by the prison administrator or governor in minor cases, but the imposition of a more serious sanction—such as loss of remission or good time—usually requires a more formal disciplinary hearing before a committee or board, which will normally follow the basic rules of procedure in a court of law.

risoners' ghts

The idea that the prisoner has rights that may be protected by actions in the courts has been developed particularly in the United States, where actions brought under the provisions of the U.S. Constitution (notably the Eighth and Fourteenth amendments) have established that as a general principle prisoners are entitled to the protection of the Constitution and that interference with rights guaranteed by the Constitution to citizens in general requires special justification. In some cases, courts have ordered state prison administrators to make major improvements in prison conditions or to close down particular institutions, but not all of these decisions have been effectively enforced. In England, in the absence of a written constitution, prisoners resorting to the courts have relied on the general principles of administrative law, which require fair procedures by disciplinary bodies. Although many actions brought by prisoners have been unsuccessful, prison disciplinary procedures have been improved as a result of such litigation.

For many prisoners the worst pressures arise not from the prison authorities but from fellow prisoners, particularly in overcrowded institutions where prisoners are forced to share cells and supervision is limited. Some criminologists have claimed that there exists a prison subculture, standing opposed to the official hierarchy of the prisons, which demands the loyalty of the prisoner and expects him to conform to a series of informal rules, enforcing his compliance by violence and social pressures. Certain types of prisoner are particularly likely to be treated with violence by other prisoners—those who have committed sexual offenses against children (known in English prison argot as "nonces") or law enforcement officers sentenced for corruption or similar matters. In the English prison system such offenders may be allowed to go into solitary confinement for their own protection. Racial conflict is a major problem in many U.S. prisons, and riots have occurred in prisons in the United States and other countries, usually as a result of grievances over the management of the prison, disparity of sentencing, and the uncertainties of the parole system.

rison ıspection

One innovation of some importance in England is the creation of an Inspectorate of Prisons. The chief inspector of prisons and his assistants have the responsibility of inspecting all prisons in England and of reporting directly to the home secretary (rather than to the prison department of the Home Office). An annual report is published each year by the Inspectorate, dealing with the general problems of the system, and individual reports, some of them very critical, have been published on the conditions and management of particular institutions.

## The death penalty

IN ENGLISH LAW

Death was formerly the penalty for all felonies in English law. In practice the death penalty was never applied as widely as the law provided, as a variety of procedures were adopted to mitigate the harshness of the law. Many offenders who committed capital crimes were pardoned, usually on condition that they agreed to be transported (or to transport themselves) to what were then the American colonies; others were allowed what was known as benefit of clergy. The origin of benefit of clergy was that offenders who were ordained priests (clerks in Holy Orders) were subject to trial by the church courts rather than the secular courts; if the offender convicted of a felony could show that he had been ordained, he was allowed to go free, subject to the possibility of being punished by the ecclesiastical courts. In medieval times the only proof of ordination was literacy, and it became the custom by the 17th century to allow anyone convicted of a felony to escape the death sentence by giving proof of literacy. All that was required was the ability to read (or recite) one particular verse from Psalm 51 of the Bible, known as the "neck verse" (for its ability to save one's neck); most offenders learned the words by heart.

enefit of lergy

In 18th-century England concern with rising crime led to many statutes either extending the number of offenses punishable with death or doing away with benefit of clergy for existing felonies, which as a result became capital. By the end of the 18th century English criminal law contained about 200 capital offenses. The application of the death penalty was extremely erratic, as in any capital case the judge was entitled to reprieve the offender so that he could petition for mercy; but the judge was not obliged to do this, and if he decided to "leave the offender for execution," the death sentence was normally carried out immediately after the closing of the assize, without appeal. In practice, many offenders who were convicted of capital crimes escaped the gallows as a result of reprieves and royal pardons, usually on condition of transportation, and many others who were charged with capital crimes were acquitted against the evidence, because the jury was unwilling to see the death penalty applied in a minor case.

The erratic application of the death penalty in the late 18th and early 19th centuries led to demands for reform, both from humanitarian reformers, such as Sir Samuel Romilly, and from those who were more concerned with the effectiveness of the legal system and who could see that the very severity and arbitrariness of the law and its administration undermined its deterrent effect. Between 1820 and 1840 most of the capital statutes were repealed, and by 1861 only four offenses retained the death penalty—murder, treason, arson in a royal dockyard, and piracy with violence.

Public execution

Until the mid-19th century executions in England were public, and throughout the 18th century great crowds attended the regular executions in London and other cities. Often an execution was followed by scenes of violence and disorder in the crowd, and it was commonly believed that pickpockets were busy among the spectators at executions, even at those for the offense of larceny from the person. Public opinion eventually turned against the idea of executions as spectacles, and after 1868 executions were carried out in private in prisons.

Although treason remained a capital crime in England, and persons convicted of treason were executed after both world wars, in practice the only capital crime for which criminals remained liable to be executed was murder. (Arson in a royal dockyard ceased to be capital in 1971.) From the 1930s until the mid-1960s, reformers campaigned for the abolition of the death penalty for murder. One attempt came close to succeeding in 1947, and later the government appointed a royal commission to consider the question; the report of this commission, the Royal Commission on Capital Punishment 1953, remains one of the most significant and comprehensive accounts of the question. Following its publication, a number of controversial executions, and a further parliamentary attempt to abolish the death penalty altogether, Parliament enacted in 1957 a statute restricting the death penalty to certain types of murder, known as "capital murders"—murder in the course of theft, murder of a police or prison officer in the execution of his duty, murder by shooting or causing an explosion, and murder on a second occasion. All other murders were to be punished by a mandatory life sentence (although murderers sentenced to life imprisonment were eligible to be released on license at any time during their sentence).

Restriction and abolition in Britain

The operation of the system of capital murder created great dissatisfaction, as it led to some executions that the public viewed as unjustified, while other types of murderers escaped the death penalty simply because of the method used to commit the crime (in particular, deliberate poisoners were not subject to the death penalty, but the emotional murderer who had happened to seize a gun was liable to execution). Another objection was the

fact that the liability to the death penalty might depend on a narrow question of law—such as whether a murder committed by a burglar escaping from the scene of the burglary occurred "in the course or furtherance of theft" if the theft was already complete. These objections led to a further move for change, and in 1965 the Murder (Abolition of Death Penalty) Act was passed, abolishing the death penalty for all murders and replacing it with a mandatory life sentence in all cases. The judge was given the power to recommend that the offender sentenced to life imprisonment should not be released before he had served a certain minimum period. Despite two parliamentary motions to restore the death penalty, both of which failed, this remained the position in England and Wales and Scotland in the mid-1980s. Northern Ireland retained the death penalty under a law designating certain murders as capital but abolished it in 1973.

IN THE UNITED STATES

In the United States, where the existence of the death penalty is primarily a matter of state law, capital punishment was never as widely provided as in 18th-century England, but it was permitted by many states for murder and in some states for offenses such as rape and kidnapping. Executions were common; between 150 and 200 persons were executed each year in the decade before World War II. In the postwar years the number of executions declined to about 50 each year by the late 1950s. During the 1960s doubts grew as to whether the application of the death penalty was constitutional; the question was raised as to whether execution was "cruel and unusual punishment" of a kind forbidden by the Eighth Amendment to the Constitution, or whether it violated the requirement of the Fifth and Fourteenth amendments that all persons within the United States should be afforded equal protection under the law. These doubts led to a complete cessation of executions for a decade, until the constitutional issues were settled by the Supreme Court of the United States in 1972 in the case of *Furman* v. *Georgia,* although this turned out to be a confusing ruling. The Supreme Court ruled that the death penalty itself did not violate the Constitution but that the manner of its application in many states did. It was shown that capital punishment was likely to be imposed in a discriminatory way and in particular that blacks were far more likely to be executed than whites. The decision in *Furman* v. *Georgia* had left uncertain the precise requirements of the Constitution for a valid death penalty statute, except that it required a system for applying the death penalty that would not be discriminatory against any racial or other minority.

The question of constitutionality

Some states enacted legislation making the death penalty mandatory in all cases of convictions for the crime in question, on the assumption that, if there was no discretion in the application of the penalty, there could be no question of discrimination in its application. Other states enacted statutes that provided for the death penalty to be imposed only after a special hearing, at which matters of mitigation and aggravation were to be considered, so that the discretion would be exercised in a systematic rather than an arbitrary manner. The constitutionality of these new statutes was considered by the Supreme Court in a series of decisions in 1976, which decided that laws making the application of the death penalty automatic were unconstitutional but that those providing a framework for the exercise of discretion in a structured manner were constitutional. The decision upheld the death penalty statutes of some states; in the light of the decision, other states enacted new legislation providing for the application of the death penalty in a manner indicated by the Supreme Court. About two-thirds of the states now have provisions in their laws for the death penalty for murder; although some states provide the death penalty for other crimes, there is doubt about its constitutionality in these cases.

One effect of the doubts over the constitutionality of the death penalty, together with the length of time needed to exhaust the appeal procedures available in the United States, was that the informal moratorium on the carrying out of death sentences that had begun in 1967 came to an end. The first execution under the new legislation took place in 1977, but many who were sentenced to death after 1976 contested the validity of the convictions or sentences, with the result that a large number of convicted offenders were held on "death row" in U.S. prisons, waiting in some cases for years to know if they were to be executed. In 1986, for example, some 1,700 prisoners were held in U.S. prisons under sentence of death.

"Death row"

IN OTHER COUNTRIES

**Continental Europe.** Most European countries have abolished the death penalty for murder; it remains on the statute book for peacetime offenses only in Belgium but is never carried out in practice. (D.A.T.)

**Africa.** The death penalty has been generally retained by African countries, and (in default of an active public opinion) there are no movements to abandon it. Indeed, African governments increasingly resort to the death penalty as the sanction for new classes of offenses, such as armed robbery. Military takeovers often lead to the execution by firing squad of those found guilty of serious "crimes" committed before the takeover.

**Soviet Union.** The post-Stalin period witnessed an increase in the use of the death penalty by Soviet courts, to the extent that the Soviet Union assumed a place among the world's leaders in executions. Inflicted on several hundred convicts per year, capital punishment was directed not against political offenders (as had been the case under Stalin) but rather against vicious murderers and officials convicted of large-scale embezzlement of state property. The use of the death penalty and of lengthy terms of imprisonment reflected the potency of the idea of deterrence, if not among penologists, at least among the politicians who made the law.

**China.** Despite the penological ideal of reform, the death penalty is widely used in China. Most executions appear to be for murder, rape, and robbery with violence. In 1981 the number of offenses carrying a possible death penalty was increased to include theft, bribery, embezzlement, molesting women, gang fighting, drug trafficking, pimping, and teaching criminal methods. The Criminal Procedure Law adopted in 1979 originally provided for all death sentences to be approved by the Supreme People's Court, China's highest judicial organ, but in 1981 this general requirement was removed in cases of murder, rape, robbery, and a number of other crimes, such as breaching dikes, that involved danger to the public.

Death sentences in China can be for immediate execution or for suspended execution, whereby the condemned is given a two-year reprieve. If a person shows evidence of reform and repentance, the sentence may be commuted at the end of the two years to a life sentence or a fixed term of imprisonment.

A distinctive feature of the death penalty in China has been the use of "mass sentencing rallies" to publicize exemplary cases. Condemned prisoners have frequently been paraded in public before their execution. The Criminal Procedure Law provides that the execution itself not be public, but this rule has not been universally observed.

(A.N.A./P.H.S./D.C.C.)

THE CONTINUING CONTROVERSY

Arguments for and against the death penalty take many forms. Those in favour of its retention or reintroduction for murder claim that it has a uniquely potent deterrent effect on potentially violent offenders for whom the threat of imprisonment is not a sufficient restraint; that death is the only penalty that adequately reflects the gravity of murder; that prolonged detention over decades is a harsher penalty than death; that execution is the only sure means of preventing a murderer from being released or escaping and committing more murders. They may argue also that to keep murderers in prison for long periods is uneconomical. Those against the death penalty point out that there is no evidence of its being a more potent deterrent than the threat of a sentence of life imprisonment; that the death penalty tends to be imposed in a discriminatory manner on the poor and on members of minority groups; that it creates the risk that an innocent person may be executed; that it prevents any possible rehabilitation of the offender;

that it lowers the community and the state to the same level of behaviour as the criminal; that violence used by the state in the form of capital punishment breeds violence by criminals and brutalizes those who have to administer it; that the death penalty distorts the administration of the criminal law and sensationalizes trials. Criminologists have never succeeded in producing convincing evidence to resolve these issues. Many of the arguments involved in the debate are questions of morality and personal conviction that are not within the scope of criminological research.

## Alternatives to prison

In most criminal justice systems the majority of offenders are dealt with by means other than custody—by fines and other financial penalties, by probation or supervision, or by orders to make reparation in some practical form to the community.

Fines

The most common penalty is the fine. In 1984 in England, for instance, 42 percent of all criminal offenders (excluding motoring offenders) were fined, the same percentage were dealt with by various other means not involving custody, and 16 percent were imprisoned in one manner or another. The fine is a simple penalty that avoids the disadvantages of many other forms of sentence; it is inexpensive to administer and does not normally have the side effects, such as social stigma and loss of job, that may follow imprisonment. Fining has limitations, however. There are dangers that the imposition of financial penalties may result in more affluent offenders' receiving penalties that they can easily discharge, while less affluent offenders are placed under burdens that they cannot sustain. In some cases, it has been suggested, the more affluent offender, who is able to pay a very large fine, may be able to persuade the court to fine him in circumstances where any other offender would be sent to prison; such discrimination is likely to lessen respect for the legal system. Other problems arise when courts have to deal with offenders who have no financial resources or with those whose incomes are too small to allow them to pay anything more than a derisory fine. Some countries, notably Sweden, solve this problem by allowing the court to calculate the fine not in terms of a sum of money but as a number of days' earnings.

The problem of lack of means is to some extent related to that of the enforcement of fines; a significant number of offenders who are fined have to be brought back to court for nonpayment of the fines imposed on them. If the court is satisfied that the offender has failed to pay as a result of willful neglect or culpable default, and that other means of securing payment are unlikely to succeed, he may be committed to prison. The other means include seizure and sale of the offender's property (distress) or seizure of any funds he may have in a bank or savings account (garnishee order). The length of time for which an offender may be committed to prison for deliberate nonpayment of a fine depends on the amount outstanding. Each year about 20,000 offenders are imprisoned in England for deliberate nonpayment of fines (this represents about 1 percent of those who are fined), but most of these are imprisoned for very short periods—one or two weeks is typical. If the offender is able to pay the amount outstanding, he is entitled to immediate release; if he pays a part, the term of imprisonment is reduced proportionately.

Restitution or compensation

Related to the fine is an order to pay restitution (in some countries termed compensation). The principle of restitution is popular in some countries as an alternative to punitive sentencing, but there are some drawbacks. One is the possibility, as in the case of the fine, that the more affluent offender may receive favourable treatment from the court because he is able to pay restitution (particularly if he pleads that he should not be sent to prison in order to allow him to continue to earn the money with which to pay restitution to the victim). The second drawback is that such schemes do not help all victims of crime. Only those who are the victims of crimes for which the offender is caught and convicted and has the funds to pay restitution are likely to be recompensed. Even when an offender is ordered to pay restitution, it is often by installments over a long period. Victims of crimes of violence in some countries—such as England and Canada—are entitled to restitution from public funds, whether or not the offender is detected or has the resources necessary to compensate him. The English scheme is administered by the Criminal Injuries Compensation Board, to whom the victim applies, showing evidence of the violence against him and the extent of the loss he has suffered as a result of the crime. If the board is satisfied that the crime has occurred and that the claim is reasonable, the victim is compensated out of public funds in a single payment, sometimes in a large amount. The scheme has a number of limitations—the relatives of murdered people do not normally receive anything, unless they were financially dependent on the victim or the victim was a child under 17 (when a token amount is payable); nothing is paid at all if the total extent of the injury is such that the amount which would be awarded is less than £400; and nothing is payable at all if the victim in any way provoked the crime or has himself a record of criminal offenses.

Probation and suspended sentence

There are many ways of dealing with offenders that do not involve the payment of money. One is probation, a system that takes many different forms in different jurisdictions but that essentially involves the suspension of sentence on the offender subject to the condition that he is supervised while living in the community by a probation officer and possibly agrees to comply with such other requirements as the court may think appropriate. Usually, if the offender complies with the probation order and commits no further offense while it is in force, no other penalty is imposed, but if he breaks the requirement of the order or commits another offense, he can be brought back before the court and punished for the original offense as well as the later one. In many U.S. states probation is combined with a suspended sentence, so that the sentence the offender will have to serve if he breaks the order is fixed in advance; in England the sentence is not fixed in advance, and the court has complete discretion if there is a breach to sentence the offender for the original crime in light of his later behaviour. English law also allows suspended sentence of imprisonment for a specified period (not more than two years), on condition that the offender commit no further offense during the period of suspension. This is different from a probation order, as no supervision is required and no other conditions may be included in the order.

Offenders who are found to be suffering from mental illness may be committed to a mental hospital rather than a penal institution. English law, for instance, allows a criminal court to make a hospital order against an offender whose mental condition warrants his detention for treatment, provided that there is medical evidence to that effect before the court. Alternatively, the court may make a probation order with a condition that the offender undertake psychiatric treatment. An offender who is detained in a hospital may apply to the Mental Health Review tribunal, which can order his release if it considers that his detention is no longer justified.

Reparation

The concept of reparation has gained in popularity in a number of jurisdictions. Under this method, the offender makes good the damage he has done through his crime, not by paying money but by providing services to the victim directly or indirectly through the community. In England this takes the form of the community service order, under which the court is empowered to order anyone who is convicted of an offense that could be punished with imprisonment to perform up to 240 hours of unpaid work for the community, usually over a period of not more than 12 months. The consent of the offender is necessary before the court can make such an order, to avoid allegations that it amounts to forced labour. Typically, the offender carries out work in his leisure time, under the direction of the probation service. The kind of work involved varies according to the area, the time of year, and the abilities of the offender; in some cases it may involve heavy physical labour, but in others it may require such work as the provision of help to handicapped people. If the offender completes the hours of work ordered by the court, he receives no further penalty, but if he fails to carry out the work, without reasonable excuse, he can be resentenced

for the original offense. Although follow-up studies of offenders given community service orders have not shown that this method is more effective than other forms of sentence in preventing further offenses (the same proportion of offenders who receive community service orders are convicted of later offenses as are those sentenced in other ways), the community service order is widely judged to be a successful innovation, and several other countries have adopted systems based on the same principle. It is less expensive to administer than imprisonment, less damaging to the offender and his family, and more useful to the community, and the majority of offenders complete the order satisfactorily, whatever their subsequent behaviour may be. There are some doubts about the extent to which the availability of community service as an alternative to prison weakens the deterrent effect of the criminal law, but there can be no doubt that community service has become an established sentencing alternative.

Other alternatives

Other alternatives to prison are based on the idea of preventing an offender from committing further offenses, without necessarily confining him in a prison. The most familiar power of this kind is that of disqualifying an offender from driving a motor vehicle or from holding a driver's license. This power is available under the laws of most countries to deal with those offenders who either commit serious driving offenses, in particular driving after consuming an excessive amount of alcohol, or who commit repeated but less serious offenses such as speeding. In many countries there exists a system in which the offender is awarded a number of points each time he commits a motoring offense; when the number of points accumulated reaches a certain figure, he is automatically disqualified for a specified period. Some countries allow courts to disqualify from driving those offenders who have used motor vehicles in commission of the crime for which they are being sentenced, with the aim of hindering the offender from committing further such offenses. Although attractive in the abstract, this seldom works well in practice, as the absence of a driver's license may well prevent an offender from finding work after release from prison; as a result he may be likely to commit further crimes. Other forms of disqualification may be imposed on offenders convicted of particular types of crimes: a fraudulent company director may be disqualified from being involved in the direction of a company, a corrupt politician may be disqualified from holding public office, or a parent who sexually abuses his children may be deprived of parental authority over them.

## Crime and social policy

Increasing crime appears to be a feature of all modern industrialized societies, and no developments in law or penology can be shown to have had a significant impact on the problem. The effect of crime on the quality of life cannot be measured simply in terms of the actual incidence of crime, as the fear of crime affects far more people than are likely to become victims and forces them to accept limitations on their freedom of action. Paradoxically, many social changes that are seen as progress may lead to further escalation in the incidence of crime—economic progress, producing greater wealth, almost always leads to greater opportunities for crime in the form of more goods to steal or enhanced possibilities for successful fraud—and an increase in individual liberty may have similar effects, as the older constraints on behaviour are discarded. Crime is least likely to be a serious problem in a society that is economically undeveloped and subject to restrictive religious or similar restraints on behaviour. For modern urbanized society, in which economic growth and personal success are dominant values, there is no reason to suppose that crime rates will not continue to increase. (D.A.T.)

**BIBLIOGRAPHY.** General coverage of topics in crime and punishment and in criminology is provided by SANFORD H. KADISH (ed.), *Encyclopedia of Crime and Justice,* 4 vol. (1983), which includes extensive, detailed subject bibliographies. See also LEON RADZINOWICZ and JOAN KING, *The Growth of Crime: The International Experience* (1977), an introductory text on crime and punishment; A. KEITH BOTTOMLEY, *Criminology in Focus: Past Trends and Future Prospects* (1979), a sociopolitical examination of the explanations of crime; HERMANN MANNHEIM, *Comparative Criminology: A Text Book* (1965, reissued 1973), a classic; JAMES Q. WILSON and RICHARD J. HERRNSTEIN, *Crime and Human Nature* (1985), an anthropological study; HERBERT L. PACKER, *The Limits of the Criminal Sanction* (1968), a discussion of the application of legal punishment; ARTHUR L. STINCHCOMBE *et al., Crime and Punishment—Changing Attitudes in America* (1980), a report on public opinion; and SAMUEL WALKER, *Popular Justice: A History of American Criminal Justice* (1980), a brief overview, with an especially good analysis of prisons.

Criminal policy considerations are the subject of JOHANNES ANDENAES, *Punishment and Deterrence* (1974); ELLIOTT CURRIE, *Confronting Crime: An American Challenge* (1985), a review of crime policy literature; HAROLD E. PEPINSKY and PAUL JESILOW, *Myths That Cause Crime,* 2nd ed. (1985), an attempt at demystifying crime; SAMUEL WALKER, *Sense and Nonsense About Crime: A Policy Guide* (1985), an examination of various positions on crime control; JAMES Q. WILSON (ed.), *Crime and Public Policy* (1978), a collection of essays; and FRANKLIN E. ZIMRING and GORDON J. HAWKINS, *Deterrence: The Legal Threat in Crime Control* (1973, reprinted 1976). See also FRANCESCO CORDASCO and DAVID N. ALLOWAY, *Crime in America: Historical Patterns and Contemporary Realities: An Annotated Bibliography* (1985).

Works on crime detection and criminal procedure include FRANK LUNDQUIST and A.S. CURRY (eds.), *Methods of Forensic Science,* 4 vol. (1962–65); H.J. WALLS, *Forensic Science: An Introduction* (1974); RUDOLF VOM ENDE, *Criminology and Forensic Sciences: An International Bibliography, 1950–1980,* 3 vol. (1981–82); GREAT BRITAIN. ROYAL COMMISSION ON CRIMINAL PROCEDURE, *Report* (1981); and CHRISTOPHER J. EMMINS, *A Practical Approach to Criminal Procedure,* 3rd ed. (1985). For works on rehabilitation and sentencing see FRANCIS A. ALLEN, *The Decline of the Rehabilitative Ideal: Penal Policy and Social Purpose* (1981); FRANCIS T. CULLEN and KAREN E. GILBERT, *Reaffirming Rehabilitation* (1982); ANDREW ASHWORTH, *Sentencing and Penal Policy* (1983); ALFRED BLUMSTEIN *et al.* (eds.), *Research on Sentencing: The Search for Reform,* 2 vol. (1983); LOIS G. FORER, *Criminals and Victims: A Trial Judge Reflects on Crime and Punishment* (1980); MARVIN E. FRANKEL, *Criminal Sentences: Law Without Order* (1973); JEFFREY H. REIMAN, *The Rich Get Richer and the Poor Get Prison: Ideology, Class and Criminal Justice* (1979); JOANNA SHAPLAND, *Between Conviction and Sentence: The Process of Mitigation* (1981); and NIGEL WALKER, *Sentencing: Theory, Law and Practice* (1985). Works on capital punishment include JAN GÓRECKI, *Capital Punishment: Criminal Law and Social Evolution* (1983); DAVID PANNICK, *Judicial Review of the Death Penalty* (1982); and THORSTEN SELLIN, *The Penalty of Death* (1980). Prisons and parole are discussed in ERVING GOFFMAN, *Asylums: Essays on the Social Situations of Mental Patients and Other Inmates* (1960, reissued 1973); NORVAL MORRISS, *The Future of Imprisonment* (1974); MICHAEL SHERMAN and GORDON HAWKINS, *Imprisonment in America: Choosing the Future* (1981); GRESHAM M. SYKES, *The Society of Captives: A Study of a Maximum Security Prison* (1958, reissued 1971); MARVIN E. WOLFGANG (ed.), *Prisons, Present and Possible* (1979); VERGIL L. WILLIAMS, *Dictionary of American Penology: An Introductory Guide* (1979); and ANDREW VON HIRSCH and KATHLEEN J. HANRAHAN, *The Question of Parole: Retention, Reform, or Abolition?* (1979). Crime victims are the subject of ROBERT ELIAS, *Victims of the System: Crime Victims and Compensation in American Politics and Criminal Justice* (1983); DAVID MIERS, *Responses to Victimisation: A Comparative Study of Compensation for Criminal Violence in Great Britain and Ontario* (1978); and JOANNA SHAPLAND, *Victims in the Criminal Justice System* (1985). The relation between mental health and criminal justice is treated in NORVAL MORRIS, *Madness and the Criminal Law* (1982, reprinted 1984); and KENT S. MILLER, *The Criminal Justice and Mental Health Systems: Conflict and Collusion* (1980).

Juvenile delinquency, policy, and justice are discussed in MICHAEL RUTTER and HENRI GILLER, *Juvenile Delinquency: Trends and Perspectives* (1983, reissued 1984); PAUL LERMAN, *Delinquency and Social Policy* (1970); JAMES C. HACKLER, *The Prevention of Youthful Crime: The Great Stumble Forward* (1978); and LAMAR T. EMPEY, *Juvenile Justice: The Progressive Legacy and Current Reforms* (1979).

(D.A.T.)

Organized crime is the focus of WILLIAM J. CHAMBLISS, *On the Take: From Petty Crooks to Presidents* (1978, reprinted 1982), a lively analysis of the interaction between gangsters, business people, politicians, and labour racketeers; MICHAEL LEVI, *The Phantom Capitalists: The Organisation and Control of Long-Firm Fraud* (1981), a study of the criminal organization and techniques of bankruptcy fraud and of the impact of control measures upon them; MARY MCINTOSH, *The Organisation of*

*Crime* (1975), a comparative and historical analysis of the way in which criminal organizations adapt to changes in law enforcement and the economy; WAYNE MOQUIN and CHARLES VAN DOREN (eds.), *The American Way of Crime: A Documentary History* (1976), a collection of historical readings; PETER REUTER, *Disorganized Crime: The Economics of the Visible Hand* (1983), an examination of the structure and profitability of crime in the United States; LAURIE TAYLOR, *In the Underworld* (1984, reissued 1985), a racy account of the London underworld; and UNITED STATES. PRESIDENT'S COMMISSION ON ORGANIZED CRIME, *The Edge: Organized Crime, Business, and Labor Unions* (1986), a review of the nature and impact of organized crime in America. See also EUGENE DOLESCHAL, ANNE NEWTON, and WILLIAM HICKEY, *A Guide to the Literature on Organized Crime: An Annotated Bibliography Covering the Years 1967–81.*

White-collar crime is examined in AUGUST BEQUAI, *White-Collar Crime: A 20th-Century Crisis* (1978), a synopsis of the main techniques of fraud in America; STEVEN BOX, *Power, Crime, and Mystification* (1984), an account of corporate, female, sexual, and police crime in the United Kingdom; M. DAVID ERMANN and RICHARD J. LUNDMAN, *Corporate Deviance* (1982), an overview of some of the more sensational cases of malpractice by U.S. corporations against owners, employees, customers, and the general public; GILBERT GEIS, "White-Collar and Corporate Crime," ch. 7 in ROBERT F. MEIER (ed.), *Major Forms of Crime* (1984), pp. 137–166, a summary of the development of research and thinking about white-collar crime; GILBERT GEIS and EZRA STOTLAND (eds.), *White-Collar Crime: Theory and Research* (1980), an important source; MICHAEL LEVI, *Fraudulent Justice: White-Collar Crime and the Criminal Process* (1987); and EDWIN H. SUTHERLAND, *White Collar Crime* (1949, reissued 1983), a seminal work—the reissue containing the names of corporations not printed in the original edition and an excellent introduction.

(Mi.L.)

Books dealing with international terrorism from the Western viewpoint are JOHN BARRON, *KGB: The Secret Work of Soviet Secret Agents* (1974), and *KGB Today: The Hidden Hand* (1983, reissued 1985); JILLIAN BECKER, *The Soviet Connection: State Sponsorship of Terrorism* (1985); OVID DEMARIS, *Brothers in Blood: The International Terrorist Network* (1978); SAMUEL T. FRANCIS, *The Soviet Strategy of Terror,* rev. ed. (1985); ROBERTA GOREN, *The Soviet Union and Terrorism* (1984); and CLAIRE STERLING, *The Terror Network: The Secret War of International Terrorism* (1981, reprinted 1982). The following give the Soviet view of international terrorism and what lies behind it: I. BLISHCHENKO and N. ZHDANOV, *Terrorism and International Law,* trans. from Russian (1984); ERNST HENRY, *Stop Terrorism* (1982; originally published in Russian, 1981); and BORIS SVETOV *et al., International Terrorism and the CIA: Documents, Eyewitness Reports, Facts* (1983; originally published in Russian, 1981).

A general work dealing in part with the philosophical and sociological background of modern terrorism is WALTER LAQUEUR, *Terrorism* (1977). Comprehensive case histories of two terrorist groups are JILLIAN BECKER, *Hitler's Children: The Story of the Baader-Meinhof Terrorist Gang,* rev. ed. (1978), and *The PLO: The Rise and Fall of the Palestine Liberation Organization* (1984). A work describing the attack of terrorism on liberal democracy in the West is PAUL WILKINSON, *Terrorism and the Liberal State,* 2nd rev. ed. (1986). See also YONAH ALEXANDER (ed.), *International Terrorism: National, Regional, and Global Perspectives* (1976); CHRISTOPHER DOBSON and RONALD PAYNE, *The Terrorists: Their Weapons, Leaders, and Tactics* (1982); EDWARD F. MICKOLUS, *Transnational Terrorism: A Chronology of Events, 1968–1979* (1980), and *The Literature of Terrorism: A Selectively Annotated Bibliography* (1980); PETER JANKE and RICHARD SIM, *Guerrilla and Terrorist Organisations: A World Directory and Bibliography* (1983).

(Ji.B.)

Crime and punishment in other countries are treated in general in DANE ARCHER and ROSEMARY GARTNER, *Violence and Crime in Cross-National Perspective* (1984); GEORGE F. COLE *et al.* (eds.), *Major Criminal Justice Systems* (1981); DAE H. CHANG (ed.), *Criminology: A Cross-Cultural Perspective,* 2 vol. (1976); and LOUISE I. SHELLEY, *Crime and Modernization: The Impact of Industrialization and Urbanization on Crime* (1981). Specific major world regions are discussed in the following: for Australia, see DAVID BILES (ed.), *Crime and Justice in Australia* (1977); and DUNCAN CHAPPELL and PAUL WILSON (eds.), *The Australian Criminal Justice System,* 2nd ed. (1977); and on Africa, see ALAN MILNER (ed.), *African Penal Systems* (1969), the leading reference work, which contains comprehensive articles reporting on the penal systems of African countries. There has been little change in policies or provision since then, apart from the consequences of armed coups referred to in the text above. W. CLIFFORD, *An Introduction to African Criminology* (1974), perceptively analyzes common criminological problems from an African perspective. ALAN MILNER, *The Nigerian Penal System* (1972), is a study of one country's provision for crime and punishment. Also useful are the country studies on the law of criminal procedure prepared by the Library of Congress Law Library; these include CARLETON W. KENYON, *The Sudan* (1984), *Mauritius* (1983), and *South Africa* (1982); and CHARLES MWALIMU, *Kenya* (1983).

(A.N.A.)

On crime in Islāmic countries, see M. CHERIF BASSIOUNI (ed.), *The Islamic Criminal Justice System* (1982). See also relevant sections of N.J. COULSON, *A History of Islamic Law* (1964, reissued 1978), an excellent guide to the elements of Islāmic law, giving its sources and their development and explaining how the law has fared in the modern age; JOSEPH SCHACHT, *An Introduction to Islamic Law* (1964, reprinted 1982), the best introduction to the subject in English, dealing with history and development and with legal ideas and including a superb bibliography; MAJID KHADDURI and HERBERT J. LIEBESNY (eds.), *Law in the Middle East,* vol. 1, *Origin and Development of Islamic Law* (1955, reprinted 1984), a collection of essays on topics of Islāmic law both traditional and modern; and *The Encyclopaedia of Islam,* 5 vol. (1913–36), and a new edition, of which five volumes appeared from 1960 to 1986, the best reference work in English. Specific topics in Islāmic law are treated in J.N.D. ANDERSON, "Homicide in Islamic Law," *Bulletin of the School of Oriental and African Studies,* 13 (part 4):811–828 (1951), an excellent article on Islāmic law's attitude toward unlawful killing; and M.J.L. HARDY, *Blood Feuds and the Payment of Blood Money in the Middle East* (1963), a good account of the traditional law relating to blood feuds (*tha'r*).

(I.D.E.)

On the Soviet Union, see HAROLD J. BERMAN, *Justice in the USSR: An Interpretation of Soviet Law,* rev. ed. enl. (1963), a classic work delineating the influence of Marxism-Leninism and of Russian national tradition upon Soviet criminal law and its administration. A comprehensive compilation of the available but usually fragmentary data on crime and punishment in the Soviet Union, contemporary and historical, is found in GER P. VAN DEN BERG, *The Soviet System of Justice: Figures and Policy* (1985). VALERY CHALIDZE, *Criminal Russia: Essays on Crime in the Soviet Union* (1977; originally published in Russian, 1977), describes the criminals and criminal subcultures of the Soviet Union. On the revival of criminology in the Soviet Union and its implications for policy and practice, see PETER H. SOLOMON, JR., *Soviet Criminologists and Criminal Policy: Specialists in Policy-Making* (1978). Soviet theories of crime and the system of punishment receive detailed treatment in WALTER D. CONNOR, *Deviance in Soviet Society: Crime, Delinquency, and Alcoholism* (1972). For economic crime and undetected and unsolved crime in Soviet Russia, see A. KATSENELINBOIGEN, "Coloured Markets in the Soviet Union," *Soviet Studies,* 29(1):62–85 (January 1977); and FRIDRIKH NEZNANSKY, *The Prosecution of Economic Crimes in the USSR, 1954–1984* (1985). JOHN N. HAZARD, *Communists and Their Law: A Search for the Common Core of the Legal Systems of the Marxian Socialist States* (1969), is a work explaining variations in the approach to crime among eastern European states.

(P.H.S.)

For China, see JEROME ALAN COHEN, *The Criminal Process in the People's Republic of China, 1949–1963: An Introduction* (1968), indispensable for an understanding of many current issues and their historical background; SHAO-CHUAN LENG and HUNGDAH CHIU, *Criminal Justice in Post-Mao China: Analysis and Documents* (1985); "Concepts of Law in the Chinese Anti-Crime Campaign," *Harvard Law Review,* 98(8):1890–1908 (June 1985), an outline of the structure of punitive sanctions in China, suggesting a broad theoretical framework for understanding Chinese law, especially criminal law; CHINA, *The Criminal Law and the Criminal Procedure Law of the People's Republic of China* (1984), in English and Chinese, including several subsequent decisions that amend or supplement the two laws; WILLIAM C. JONES, "The Criminal Law of the People's Republic of China," *Review of Socialist Law,* 6(4):405–423 (December 1980), a readable and insightful exploration of the significance of the promulgation of the Criminal Law; BAO RUOWANG (JEAN PASQUALINI) and RUDOLPH CHELMINSKI, *Prisoner of Mao* (1973, reissued 1976), a fascinating account of Bao's years of imprisonment in China and the practice of thought reform; TIMOTHY A. GELATT, "The People's Republic of China and the Presumption of Innocence," *Journal of Criminal Law and Criminology,* 73(1):259–316 (Spring 1982), an analysis of theoretical debates in China since 1969 and their practical ramifications; and HUNGDAH CHIU, "Structural Changes in the Organization and Operation of China's Criminal Justice System," *Review of Socialist Law,* 7(1):53–76 (March 1981), which includes an overview of the pre-1977 criminal justice system and an appendix of excerpts from a Peking trial transcript.

(D.C.C.)

# Criminal Law

Criminal law, in a broad sense, is the body of law that defines criminal offenses, regulates the apprehension, charging, and trial of suspected persons, and fixes penalties and modes of treatment applicable to convicted offenders. Criminal law is only one of the devices by which organized societies protect the security of individual interests and assure the survival of the group. There are, in addition, the standards of conduct instilled by family, school, and religion; the rules of the office and factory; the regulations of civil life enforced by ordinary police powers; and the sanctions available through tort actions. The distinction between criminal law and tort law is difficult to draw with real precision, but in general one may say that a tort is a private injury while a crime is conceived as an offense against the public, although the actual victim may be an individual.

This article treats the principles of criminal law and the influence of the social sciences on criminal legislation and law enforcement. For treatment of the law of criminal procedure, see PROCEDURAL LAW. For the larger context of crime, criminal behaviour, and punishment, see CRIME AND PUNISHMENT.

The article is divided into the following sections:

Principles of criminal law 817
Common law and code law 817
Substantive criminal law 818
The definition of criminal conduct
The elements of crime
Some particular offenses
Criminal law and the social sciences 820
Behavioral norms
The effectiveness of statutes
The effectiveness of punishment
Treatment of mental deficiency and juvenile delinquency
Bibliography 821

## Principles of criminal law

The traditional approach to criminal law has been that crime is an act that is morally wrong. The purpose of criminal sanctions was to make the offender give retribution for harm done and expiate his moral guilt; punishment was to be meted out in proportion to the guilt of the accused. In modern times more rationalistic and pragmatic views have predominated. Writers of the Enlightenment such as Cesare Beccaria in Italy, Montesquieu and Voltaire in France, Jeremy Bentham in Britain, and P.J.A. von Feuerbach in Germany considered the main purpose of criminal law to be the prevention of crime. With the development of the social sciences, there arose new concepts, such as those of the protection of the public and the reform of the offender. Such a purpose can be seen in the West German criminal code of 1975, which provides that the court "has to consider the consequences of the sentence upon the future life of the offender in society." In the United States, a Model Penal Code proposed by the American Law Institute in 1962 states that an objective of criminal law should be "to give fair warning of the nature of the conduct declared to constitute an offense" and "to promote the correction and rehabilitation of offenders." Since that time there has been renewed interest in the concept of general prevention, including both the deterrence of possible offenders and the stabilization and strengthening of social norms.

### COMMON LAW AND CODE LAW

Important differences exist between the criminal law of most English-speaking countries and that of other countries. The criminal law of England and the United States derives from the traditional English common law of crimes and has its origins in the judicial decisions embodied in reports of decided cases. England has consistently rejected all efforts toward comprehensive legislative codification of its criminal law; even now there is no statutory definition of murder in English law. Some Commonwealth countries, however, notably India, have enacted criminal codes that are based on the English common law of crimes.

*Comparisons of Anglo-American and continental European criminal law*

The criminal law of the United States, derived from the English common law, has been adapted in some respects to American conditions. In the majority of the U.S. states the common law of crimes has been repealed by legislation. The effect of such statutes is that no person may be tried for any offense that is not specified in the statutory law of the state. But even in these states the common-law principles continue to exert influence, for the criminal statutes are often simply codifications of the common law, and their provisions are interpreted by reference to the common law. In the remaining states, prosecutions for common-law offenses not specified in statutes do sometimes occur. In a few states the so-called penal, or criminal, codes are simply collections of individual provisions with little effort made to relate the parts to the whole or to define or implement any theory of control by penal measures.

In western Europe the criminal law of modern times has emerged from various codifications. By far the most important were the two Napoleonic codes, the *Code d'Instruction Criminelle* of 1808 and the *Code Pénal* of 1810. The latter constituted the leading model for European criminal legislation throughout the first half of the 19th century, after which, although its influence in Europe waned, it continued to play an important role in the legislation of certain Latin-American and Middle Eastern countries. The German codes of 1871 (penal code) and 1877 (procedure) provided the models for other European countries and have had significant influence in Japan and South Korea, although after World War II the U.S. laws of criminal procedure were the predominant influence in the latter countries. The Italian codes of 1930 represent one of the technically most developed legislative efforts in the modern period. English criminal law has strongly influenced the law of Israel and that of the English-speaking African states. French criminal law has predominated in the French-speaking African states. Italian criminal law and theory have been influential in Latin America.

*Movements for codification and reform*

In the last few decades the movement for codification and law reform has made considerable progress everywhere. The American Law Institute's Model Penal Code stimulated a thorough reexamination of both federal and state criminal law, and new codes were enacted in most of the states. England has enacted several important reform laws (including those on theft, sexual offenses, and homicide), as well as modern legislation on imprisonment, probation, suspended sentences, and community service. Sweden enacted a new strongly progressive penal code in 1962. In West Germany (Federal Republic of Germany) a revised version of the criminal code was published in 1975 and subsequently often amended. In the same year a new criminal code came into force in Austria. New criminal codes have also been published in Portugal (1982) and Brazil (1984). France enacted important reform laws in 1958, 1970, 1975, and 1982, as did Italy in 1981 and Spain in 1983. Other reforms have been under way in Finland, The Netherlands, Belgium, Switzerland, and Japan. The Soviet Union's constituent republics began enacting revised criminal codes in 1960, as did Czechoslovakia and Hungary (1961), East Germany (German Democratic Republic), Bulgaria, and Romania (1968), and Poland (1969). After Yugoslavia became a federal state in 1974, a

number of local penal codes came into force in addition to the federal code of 1977.

Comparisons between the systems of penal law developed in the western European countries and those having their historical origins in the English common law must be stated cautiously. Substantial variations exist even among the nations that adhere generally to the Anglo-American system or to the law derived from the French, Italian, and German codes. In many respects, however, the similarities of the criminal law in all states are more important than the differences. Certain forms of behaviour are everywhere condemned by law. In matters of mitigation and justification, the continental law tends to be more explicit and articulate than the Anglo-American law, although modern legislation in countries adhering to the latter has reduced these differences. Contrasts can be drawn between the procedures of the two systems, yet even here there is a common effort to provide fair proceedings for the accused and protection for basic social interests.

## SUBSTANTIVE CRIMINAL LAW

Substantive criminal law is composed of the following elements: the definitions of the types of offenses that are held to be punishable; the classification of crimes (as, for example, felonies and misdemeanours in the United States, or *crime, délit,* and *contravention* in continental law); the principles and doctrines applied to the judgment of crime that qualify the provisions of criminal legislation (such as self-defense, necessity, insanity, and so forth); and principles determining national jurisdiction over crimes with an international aspect (crimes committed by foreigners, nationals abroad, or on ships and aircraft outside the national territory and waters).

**The definition of criminal conduct.** *Legality.* The principle of legality is recognized in almost all civilized countries as the keystone of the criminal law. It is employed in four senses. The first is that there can be no crime without a rule of law; thus immoral or antisocial conduct not forbidden and punished by law is not criminal. The law may be customary, as in common-law countries; in most countries, however, the only source of criminal law is a statute (*nullum crimen sine lege,* "no crime without a law").

Second, the principle of legality directs that criminal statutes be interpreted strictly and that they not be applied by analogical extension. If a criminal statute is ambiguous in its meaning or application, it is often given a narrow interpretation favourable to the accused. This does not mean that the law must be interpreted literally, if to do so would defeat the clear purpose of the statute. The Model Penal Code of the American Law Institute incorporates a provision that has been enacted in some U.S. state laws. It recommends that its provisions be construed "according to the fair import of their terms," which comes closer to the European practice.

Defining the application of the law

Third, the principle of legality forbids the application of the law retroactively. In order that a person may be convicted, a law must have been in effect at the time the act was committed. This aspect of the principle is embodied in the ex post facto provisions of the U.S. Constitution and such international treaties as the European Convention for the Protection of Human Rights and Fundamental Freedoms (1950) and the International Covenant on Civil and Political Rights (1966).

Fourth, the language of criminal statutes must be as clear and unambiguous as possible in order to provide fair warning to the potential lawbreaker. In some countries statutes may be considered inapplicable if they are vague.

*Protection against double jeopardy.* Legal systems generally include some restriction against prosecuting a person more than once for the same offense. In Anglo-American law the most difficult problems of double jeopardy involve the question of whether the second prosecution is for the "same" or a "different" offense. It is held that acquittal or conviction of an offense prohibits subsequent prosecution of a lesser offense that was included in the first. According to the U.S. Supreme Court in *Blockburger* v. *U.S.,* 284 U.S. 299, 304 (1932), the test to be applied to determine whether there are two offenses or only one is whether each provision requires proof of a fact that the other does not. In continental European law, on the other hand, the question is whether or not the second prosecution concerns the same "material fact" or "historical event," and the state cannot subject a person to a second trial for any offense arising out of the same factual situation.

A problem under the federal system of the United States is whether or not an offender may be prosecuted under both state and federal law for the same conduct (the specific offenses being different). A number of state laws have prohibited state prosecutions after acquittals or convictions in a federal court or in the court of another state for offenses involving the same conduct.

*Statutes of limitation.* All systems of law have statutes restricting the time within which legal proceedings may be brought. The periods prescribed may vary according to the seriousness of the offense. In German law, for example, the periods range from six months for breaches of administrative regulations to 30 years for crimes involving a life sentence. General statutes limiting the times within which prosecutions for crimes must be begun are common in continental Europe and the United States. In England there is no general statute of limitation applicable to criminal actions, although statutes for specific crimes frequently have included time limits.

In many countries there are no statutes of limitation for particularly heinous offenses, including capital felonies in the United States, genocide and murder in West Germany, and—in the Soviet Union and other Communist-governed countries—crimes against the peace, war crimes, and crimes against humanity. In 1968 the UN General Assembly adopted a Convention on the Non-applicability of Statutes of Limitation on War Crimes and Crimes against Humanity, despite strong opposition among the majority of Western members on the ground that it was retroactive.

*Requirements of jurisdiction.* The jurisdiction of a court refers to its capacity to take valid legal action. Many governments claim jurisdiction over the acts of their own nationals, even when these acts have occurred abroad. Accordingly, most states decline any obligation to surrender their nationals to other countries. The constitutions of Brazil, West Germany, and The Netherlands prohibit extradition of their nationals; and in other states extradition is prohibited by statute, as in Belgium, France, and Switzerland. The Italian constitution permits extradition of nationals only if agreed upon in international conventions.

In Anglo-American practice, on the other hand, the jurisdiction of the courts is generally limited to acts occurring in whole or part within the boundaries of the state. Nationals who commit crimes in foreign countries may be extradited, but only if required or authorized by treaty with the country concerned. Within the United States, jurisdiction over criminal conduct was formerly limited, under the common law, to acts occurring within the territorial limits of a particular state. Thus, if a person fired a bullet across a state line and killed someone in another state, sometimes only the latter state was considered to have jurisdiction. Many states have, however, by statute extended their jurisdictions to cover offenses in which either the relevant result or the relevant conduct, or even only part of it, occurred in the state. In addition, federal statutes confer jurisdiction on U.S. courts in cases involving treason, forgery of ship's papers, enticing to desertion from military service, bribery of a U.S. official, and other acts, even though the conduct occurred outside the national boundaries. The United States also claims jurisdiction over crimes committed on U.S. vessels and aircraft on or over the high seas. The Tokyo Convention on Offenses and Certain Other Acts Committed on Board Aircraft (1963) and the Hague Convention for Unlawful Seizure of Aircraft (1970) recognize that states have the right and even the duty of jurisdiction with respect to any crime committed upon aircraft bearing its national character.

**The elements of crime.** It is generally agreed that the essential ingredients of any crime are (1) a voluntary act or omission (*actus reus*) accompanied by (2) a certain state of mind (*mens rea*). An act may be any kind of voluntary human behaviour. Movements made in an

The nature of a criminal act

epileptic seizure are not acts, nor are movements made by a somnambulist before awakening, even if resulting in the death of another. Criminal liability for the result also requires that the harm done must have been caused by the accused. The test of causal relationship between conduct and result is that the event would not have happened the same way without participation of the offender.

Criminal liability may also be predicated on a failure to act when the accused was under a legal duty to act and was reasonably capable of doing so. The legal duty to act may be imposed directly by statute, such as the requirement to file an income tax return, or it may arise out of the relationship between the parties, as the obligation of parents to provide their child with food.

*The mental element.* Although most legal systems recognize the importance of the guilty mind, or *mens rea,* the statutes have not always spelled out exactly what is meant by this concept. The American Law Institute's Model Penal Code has attempted to clarify the concept by reducing the variety of mental states to four. Guilt is attributed to a person who acts "purposely," "knowingly," "recklessly," or "negligently." These terms correspond roughly to those used in continental European legal theory. Singly or in combination, they appear adequate to deal with most of the common *mens rea* problems. Their general adoption would clarify and rationalize the substantive law of crimes.

*Liability without mens rea.* Some penal offenses do not require the demonstration of culpable mind on the part of the accused. These include statutory rape, in which knowledge that the child is below the age of consent is not necessary to liability. There is also a large class of "public welfare offenses," involving such things as economic regulations or laws concerning public health and safety. The rationale for eliminating the *mens rea* requirement in such offenses is that to require the prosecution to establish the defendant's intent, or even recklessness, would render such regulatory legislation largely ineffective and unenforceable. Such cases are known in Anglo-American law as strict liability offenses, and in French law as *infractions purement matérielles.* In German law they are excluded because the requirement of *mens rea* is considered a constitutional principle.

Public welfare offenses

There has been much criticism of statutes that create liability without moral fault. To expose citizens to the condemnation of a criminal conviction without a showing of moral culpability raises issues of justice. In many instances the objectives of such legislation can more effectively be achieved by civil sanctions such as suits for damages, injunctions, and the revocation of licenses.

*Ignorance and mistake.* In most countries the law recognizes that a person who acts in ignorance of the facts of his action should not be held criminally responsible. Thus, one who takes and carries away the goods of another person, believing them to be his own, does not commit larceny, for he lacks the intent to steal. Ignorance of the law, on the other hand, is generally held not to excuse the actor; it is no defense that he was unaware that his conduct was forbidden by criminal law. This doctrine is supported by the proposition that criminal acts may be recognized as harmful and immoral by any reasonable adult. The matter is not so clear, however, when the conduct is not obviously dangerous or immoral; a substantial body of opinion would permit mistakes of law to be asserted in defense of criminal charges in such cases, particularly when the defendant has in good faith made reasonable efforts to discover what the law is. In West Germany the Federal Court of Justice in 1952 adopted the proposition that if a person engages in criminal conduct but is unaware of its criminality he cannot be fully charged with a criminal offense; this has been incorporated as rule in the West German criminal code. Law and practice in Switzerland are quite similar. In Austria mistake of law is a legal defense.

*Responsibility.* It is universally agreed that, in appropriate cases, persons suffering from serious mental disorders should be relieved of the consequences of their criminal conduct. Much controversy has arisen, however, as to the appropriate legal tests of responsibility. Most legal definitions of mental disorder are not based on modern concepts of medical science, and the psychiatrist accordingly finds it difficult to make his knowledge relevant to the requirements of the court. Various attempts have been made to formulate a new legal test of responsibility. The American Law Institute's Model Penal Code has endeavoured to meet the manifold difficulties of this problem by requiring that the defendant be deprived of "substantial capacity either to appreciate the criminality of his conduct or to conform his conduct to the requirements of the law" as a result of mental disease or defect. This resembles the Soviet formulation of 1958, which requires a mental disease as the medical condition and incapacity to appreciate or control as the psychological condition resulting from it. The same may be said of the West German law, although the latter includes in mental illness such disorders as psychopathy and neurosis in addition to psychoses and provides for various gradations of diminished responsibility. Several U.S. jurisdictions, including federal law, have abandoned the volitional prong of the insanity test and returned to the ancient English rule laid down in *M'Naghten's Case,* 8 Eng. Rep. 718, 722 (1843). According to that case an insane person is excused only if he did not know the nature and quality of his act or could not tell right from wrong. The English Homicide Act of 1957 also recognizes diminished responsibility, though to less effect. It provides that a person who kills another shall not be guilty of murder "if he was suffering from such abnormality of mind . . . as substantially impaired his mental responsibility for his acts or omissions in doing or being a party to the killing." The primary effect of this provision is to reduce an offense of murder to one of manslaughter.

Intoxication is usually not treated as mental incapacity. Soviet law is especially harsh, holding that the mental-disease defense is not applicable to persons who commit a crime while drunk and that drunkenness may even be an aggravating circumstance. In West German law, on the other hand, intoxication like any other mental defect is acceptable as a defense in criminal cases.

Intoxication

*Mitigating circumstances and other defenses.* The law generally recognizes a number of particular situations in which the use of force, even deadly force, is excused or justified. The most important body of law in this area is that which relates to self-defense. In general, in Anglo-American law, one may kill an assailant when the killer reasonably believes that he is in imminent peril of losing his life or of suffering serious bodily injury and that killing the assailant is necessary to avoid imminent peril. Some jurisdictions require that the party under attack must try to retreat when this can be done without increasing the peril. Under many continental European laws, however, the defendant may stand his ground unless he has provoked his assailant purposely or by gross negligence, or unless the assailant has some incapacity such as infancy, inebriation, mistake, or mental disease. Other situations in which the use of force is generally justifiable, both in Anglo-American law and in continental European law, include the use of force in defense of others, law enforcement, and protection of property.

The use of force may also be excused if the defendant reasonably believed himself to be acting under necessity. The doctrine of necessity in Anglo-American law relates to situations in which a person, confronted by the overwhelming pressure of natural forces, must make a choice between evils and engages in conduct that would otherwise be considered criminal. In the oft-cited case of *U.S.* v. *Holmes,* in 1842, a longboat containing passengers and members of the crew of a sunken American vessel was cast adrift in the stormy sea. To prevent the boat from being swamped, members of the crew threw some of the passengers overboard. In the trial of one of the crew members, the court recognized that such circumstances of necessity may constitute a defense to a charge of criminal homicide, provided that those sacrificed be fairly selected, as by lot. Because this had not been done, a conviction for manslaughter was returned. The leading English case, *Regina* v. *Dudley and Stephens,* 14 Q.B.D. 273 (1884), appears to reject the necessity defense in homicide cases. In German or French courts, however, the defendants would probably have been acquitted.

arying efinitions f criminal ehaviour

In general the use of force may be excused if the defendant reasonably believed himself to be acting under duress or coercion, or to be carrying out military orders believed by the defendant to be lawful.

**Some particular offenses.** All advanced legal systems condemn as criminal the sorts of conduct described in the Anglo-American law as treason, murder, aggravated assault, theft, robbery, burglary, arson, and rape. With respect to minor police regulations, however, substantial differences in the definition of criminal behaviour occur even among jurisdictions of the Anglo-American system. Comparisons of the continental European criminal law with that based on the English common law of crimes also reveal significant differences in the definition of certain aspects of more serious crimes. Continental European law, for example, frequently articulates grounds for mitigation involving considerations that are taken into account in the Anglo-American countries only in the exercise of discretion by the sentencing authority or by lay juries. This may be illustrated with respect to so-called mercy killings. The Anglo-American law of murder recognizes no formal grounds of defense or mitigation in the fact that the accused killed to relieve someone of suffering from an apparently incurable disease. Many continental European and Latin-American codes, however, provide for mitigation of offenses prompted by such motives and sometimes even recognize in such motives a defense to the criminal charge.

*Degrees of participation.* The common-law tradition distinguishes four degrees of participation in crime. One who commits the act "with his own hand" is a principal in the first degree. His counterpart in French law is the *auteur* (literally, "author"), or *coauteur* when two or more persons are directly engaged. A principal in the second degree is one who intentionally aids or abets the principal in the first degree, being present when the crime occurs; this is comparable to the French concept of *complicité par aide et assistance,* although in some countries, such as West Germany and the Soviet Union, that have adopted a wider (more subjective) interpretation of the concept it includes the activity of *coauteurs.* In Anglo-American law one who instigates, encourages, or counsels the principal without being present during the crime is called an accessory before the fact; in continental law this third degree of participation is covered partly by the concept of *instigation* and partly by the above-mentioned *aide et assistance.* The fourth and last degree of participation is that of accessory after the fact, who is punishable because he receives, conceals, or comforts one known by him to have committed a crime so as to obstruct his apprehension or to impede his punishment. In continental legal systems this has become a separate offense. Italian and Austrian law treat all participants in a crime as principals in the first degree, with the exception of accessories after the fact. The American Law Institute's Model Penal Code proposes the same simplifications.

ıccessory efore or fter the ıct

*Conspiracy.* Under the common law, conspiracy is usually described as an agreement between two or more persons to commit an unlawful act or to accomplish a lawful end by unlawful means. This definition is delusively simple, however, for each of its terms has been the object of extended judicial exposition. Criminal conspiracy is perhaps the most amorphous area in the Anglo-American law of crimes. In some jurisdictions, for example, the "unlawful" end of the conspiracy need not be one that would be criminal if accomplished by a single individual; but courts have not always agreed as to what constitutes an "unlawful" objective for these purposes. Statutory law in some American states, following the lead of the American Law Institute's Model Penal Code, have limited conspiracy offense to the furtherance of criminal objectives. The European codes have no conception of conspiracy as broad as that found in the Anglo-American legal system. In some of the continental European countries, such as France or Germany, punishment of crimes may be enhanced when the offense was committed by more than one person acting in concert. In most countries the punishment of agreements to commit offenses, irrespective of whether the criminal purpose was attempted or executed, is largely confined to political offenses against the state. Some extension of the conspiracy idea to other areas has occurred, however. Thus in the Italian code of 1930 association for the purpose of committing more than one crime was made criminal. None of these continental European provisions, however, has the generality of the original Anglo-American concept. None, for example, condemns agreements to commit acts not otherwise criminal.

*Attempt.* In Anglo-American law there is a class of offenses known as inchoate, or preliminary, crimes because guilt attaches even though the criminal purpose of the parties may not have been achieved. Thus the offense of incitement or solicitation consists of urging or requesting another to commit a crime. Certain specified types of solicitation may be criminal, such as solicitation of a bribe or solicitation for immoral purposes, or inciting members of the armed forces to mutiny.

Inchoate offenses

The most important category of inchoate offenses is attempt, which consists of any conduct intended to accomplish a criminal result that fails of consummation but goes beyond acts of preparation to a point dangerously close to completion of the intended harm. The line between acts of mere preparation and attempt is difficult to draw in many cases. In Soviet law acts of preparation are punishable if they are socially dangerous. Attempt is a further step and is generally defined in Soviet legal theory as an act that contains all the elements of an offense except the result. In continental European and some Anglo-American legal systems, attempt may also consist of conduct that would be criminal if the circumstances were as the actor believed them to be. A defense of "impossibility" is recognized only if the mistake is shown to be absolutely unreasonable. Unlike the law of some continental European countries, including the Soviet Union, no defense has traditionally been granted to an offender who voluntarily desists from committing the intended harm after his conduct has reached a point beyond mere preparation. The American Law Institute's Model Penal Code and several American state codes, however, provide for an affirmative defense if it can be shown that the actor "abandoned his effort to commit the crime or otherwise prevented its commission, under circumstances manifesting a complete and voluntary renunciation of his criminal purpose."

## Criminal law and the social sciences

Criminal law has been strongly influenced in the past century or two by the social sciences, especially criminology, sociology, and psychology. The empirical methods of the social sciences have been introduced into legal research and have done much to improve legislation and the courts' approach to sentencing, as well as the planning methods of law-enforcement agencies.

**Behavioral norms.** The fact that the crime rates in many countries have risen faster than the population has brought into question the relevance of the law itself and whether or not laws against crime actually influence behaviour. Various large-scale inquiries have been made into the relation between law and civil order: in the United States, the President's Commission on Law Enforcement and Administration of Justice; in Europe, several research studies sponsored by the Council of Europe; in Germany, the hearings of the Criminal Reform Commission of the Bundestag. One conclusion emerging from these inquiries is that criminal legislation ought to be restricted to acts that pose a serious threat to public order and that can be effectively dealt with by the police, the courts, and various correctional institutions. The effort to punish all behaviour that is considered immoral or deviant, such as drunkenness, gambling, disorderly conduct, vagrancy, and petty sex offenses, multiplies the number of crimes without changing the norms of behaviour.

Criminal law and human behaviour

**The effectiveness of statutes.** Human conduct is determined by many factors that are not responsive to criminal statutes. Thus it appears that introducing or abolishing the death penalty does not have any appreciable effect on the murder rate. Much depends also on the way in which laws are enforced. Any inquiry into the effectiveness of criminal statutes must examine the way in which police,

attorneys, and the courts operate—for example, the manner in which they investigate suspects, gather evidence, instruct juries, and use their powers in "plea bargaining" and in sentencing.

**The effectiveness of punishment.** It is difficult to determine the extent to which punishment serves to deter convicted offenders from committing further crimes. Studies into the effectiveness of various forms of the treatment of criminals has led some researchers to the conclusion that "nothing works." In a more positive light, available studies seem to indicate that lenient penalties (such as fines, probation, suspended sentences) and severe measures are about equally effective in preventing future criminality. Accordingly, there has been an international trend away from custodial treatment. Short-term sentences are seen as particularly harmful because they tear the offender away from his family and occupation and expose him to criminal indoctrination in prison and to social obloquy after his release. Long-term sentences are also viewed with growing skepticism, despite more than 150 years of prison reform, because of the adverse side effects of even the best institutions. These ill effects include acclimatization to the prison atmosphere, association with prison subcultures, infantilism, mental illness, and in general a decline in fitness for responsible life in a free community. It is now considered preferable to treat the convicted criminal in open institutions if possible.

**Treatment of mental deficiency and juvenile delinquency.** A large area of criminal behaviour involving mental deficiency and diminished responsibility cannot be dealt with through sentencing that does not also serve to rehabilitate the offender. Mentally disabled offenders require hospitalization and psychiatric treatment; this is usually handled through the probation mechanism or by commitment to a hospital for the criminally insane. Similar problems arise in the case of crimes resulting from narcotics addiction; prison terms for addicts make no sense unless some effort is made to treat the underlying condition. The same applies to juveniles, who are generally dealt with through separate courts and sent to detention centres, training centres, and part-time homes.

BIBLIOGRAPHY

*General and comparative works:* RICHARD J. TERRILL, *World Criminal Justice Systems: A Survey* (1984), provides a comprehensive overview of the topic. Other works include: M. CHERIF BASSIOUNI, *International Criminal Law: A Draft International Criminal Code* (1980); GEORGE P. FLETCHER, *Rethinking Criminal Law* (1978); HOWARD JONES, *Crime and the Penal System,* 3rd ed. (1965); LUIS JIMENEZ DE ASUA, *Tratado de derecho penal,* 7 vol. (1950–70), editions vary; HERMANN MANNHEIM, *Comparative Criminology: A Text Book,* 2 vol. (1965, reprinted 1973); EDWARD M. WISE and GERHARD O.W. MUELLER (eds.), *Studies in Comparative Criminal Law* (1975).

*Treatises on selected nations:* JOHANNES ANDENAES, *The General Part of the Criminal Law of Norway,* rev. trans. from Norwegian (1965); GUISEPPE BETTIOL, *Diritto penale: parte generale,* 11th ed. (1982); T.S. BATRA, *Criminal Law in India* (1981); JOSÉ CEREZO MIR, *Curso de derecho penal español: parte general,* 3rd rev. ed., vol. 1, part 1 (1985); M. CHERIF BASSIOUNI and V.M. SAVITSKI (eds.), *The Criminal Justice System of the USSR* (1979); F.J. FELDBRUGGE, *Soviet Criminal Law: General Part* (1964); CHRISTOPHER H.W. GANE and CHARLES N. STODDART, *A Casebook on Scottish Criminal Law* (1980); D. HAZEWINKEL-SURINGA, *Inleiding tot de studie van het Nederlandse strafrecht,* 9th ed. by J. REMMELINK (1984); HANS-HEINRICH JESCHECK, *Lehrbuch des Strafrechts: Allgemeiner Teil,* 3rd rev. ed. (1978); WAYNE R. LAFAVE and AUSTIN W. SCOTT, *Handbook on Criminal Law* (1972); ROGER MERLE and ANDRÉ VITU, *Traité de droit criminel,* 5th ed., vol. 1 (1984); FRANCISCO MUÑOZ CONDE, *Derecho penal: parte especial,* 6th ed. (1985); ANTONIO PAGLIARO, *Principi di diritto penale: parte generale,* 2nd ed. (1980); GRAHAM PARKER, *An Introduction to Criminal Law,* 2nd ed. (1983); IGOR ANDREJEW (ed.), *Droit pénal polonais: ouvrage collectif,* trans. from Polish (1982); LEON RADZINOWICZ, *A History of English Criminal Law and Its Administration from 1750,* 4 vol. (1948–68); JOSÉ MARÍA RODRÍGUEZ DEVESA and ALFONSO SERRANO GÓMEZ, *Derecho penal español: parte general,* 9th ed. (1985); HANS SCHULTZ, *Einführung in den Allgemeinen Teil des Strafrechts: Ein Grundriss,* 4th ed., 2 vol. (1982); C.R. SNYMAN, *Criminal Law* (1983); JAMES FITZJAMES STEPHEN, *A History of the Criminal Law of England,* 3 vol. (1883, reprinted 1985); IVAR STRAHL, *Allmän straffrätt i vad angår brotten* (1976); OTTO TRIFFTERER and KURT SCHMOLLER, *Österreichisches Strafrecht-Allgemeiner Teil* (1985); R.J. WALKER and M.G. WALKER, *The English Legal System,* 6th ed. (1985); GLANVILLE WILLIAMS, *Textbook on Criminal Law,* 2nd ed. (1983).

*Works on substantive criminal law:* M. CHERIF BASSIOUNI, *Substantive Criminal Law* (1978); JEROME HALL, *General Principles of Criminal Law,* 2nd ed. (1960); SANFORD H. KADISH, STEPHEN J. SCHULHOFER, and MONRAD G. PAULSEN, *Criminal Law and Its Processes: Cases and Materials,* 4th ed. (1983); C.S. KENNY, *Kenny's Outlines of Criminal Law,* 19th ed. by J.W. CECIL TURNER (1966); ROLLIN M. PERKINS and RONALD N. BOYCE, *Criminal Law,* 3rd ed. (1982); PAUL H. ROBINSON, *Criminal Law Defenses,* 2 vol. (1984); J.C. SMITH and BRIAN HOGAN, *Criminal Law,* 5th ed. (1983); GLANVILLE WILLIAMS, *Criminal Law: The General Part,* 2nd ed. (1961).

*Works in criminology and sociology:* MARC ANCEL, *Social Defence: A Modern Approach to Criminal Problems* (1965; originally published in French, 1954; 3rd rev. ed., 1981); G. AVANESOV, *The Principles of Criminology* (1982; originally published in Russian, 1981); ROBERT L. BONN, *Criminology* (1984); G. CANEPA, *Personalità e delinquenza: problemi di antropologia criminale e di criminologia clinica* (1974); JOHN E. CONKLIN, *Criminology* (1981); HANS GÖPPINGER, *Kriminologie,* 4th rev. ed. (1980); J.E. HALL WILLIAMS, *Criminology and Criminal Justice* (1982); HANS VON HENTIG, *Das Verbrechen,* 3 vol. (1961–63); NORMAN JOHNSTON, LEONARD SAVITZ, and MARVIN E. WOLFGANG (eds.), *The Sociology of Punishment and Correction,* 2nd ed. (1970); GÜNTHER KAISER, *Kriminologie: Eine Einführung in die Grundlagen,* 6th rev. ed. (1983); JOHN LEKSCHAS *et al., Kriminologie: Theoretische Grundlagen und Analysen* (1983); NORVAL MORRIS, *The Future of Imprisonment* (1974); EDWIN H. SUTHERLAND and DONALD R. CRESSEY, *Criminology,* 10th ed. (1978); MICHAEL TONRY and FRANKLIN E. ZUMRING (eds.), *Reform and Punishment: Essays on Criminal Sentencing* (1983).

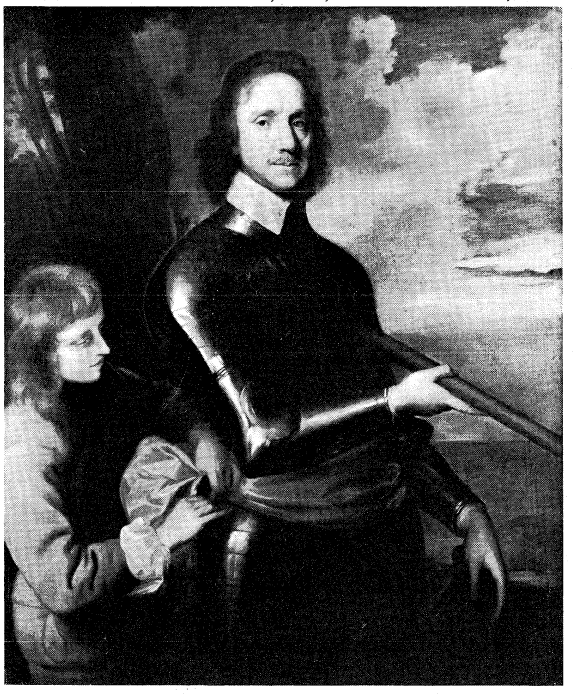
*Encyclopaedias:* SANFORD H. KADISH (ed.), *Encyclopedia of Crime and Justice,* 4 vol. (1983), a work of international scope noted for its comprehensive coverage; ALEXANDER ELSTER and HEINRICH LINGEMANN, *Handwörterbuch der Kriminologie,* 2nd rev. ed. edited by RUDOLF SIEVERTS, 3 vol. (1966–77), the most important German criminology encyclopaedia.

(H.-H.J.)

# Oliver Cromwell

Oliver Cromwell, an English soldier and statesman of outstanding gifts and a forceful character shaped by a devout Calvinist faith, was lord protector of the republican Commonwealth of England, Scotland, and Ireland from 1653 to 1658. One of the leading generals on the parliamentary side in the English Civil War against King Charles I, he helped to bring about the overthrow of the Stuart monarchy, and, as lord protector, he raised his country's status once more to that of a leading European power from the decline it had gone through since the death of Queen Elizabeth I. Cromwell was one of the most remarkable rulers in modern European history; for although a convinced Calvinist, he believed deeply in the value of religious toleration. At the same time his victories at home and abroad helped to enlarge and sustain a Puritan attitude of mind, both in Great Britain and in North America, that continued to influence political and social life until recent times.

By courtesy of the National Portrait Gallery, London



Cromwell, painting by Robert Walker. In the National Portrait Gallery, London.

## YOUTH AND EARLY PUBLIC CAREER

Upbringing

Cromwell was born at Huntingdon in eastern England on April 25, 1599, the only son of Robert Cromwell and Elizabeth Steward. His father had been a member of one of Queen Elizabeth's parliaments and, as a landlord and justice of the peace, was active in local affairs. Robert Cromwell died when his son was 18, but his widow lived to the age of 89. Oliver went to the local grammar school and then for a year attended Sidney Sussex College, Cambridge. After his father's death he left Cambridge to look after his widowed mother and sisters but is believed to have studied for a time at Lincoln's Inn in London, where country gentlemen were accustomed to acquire a smattering of law. In August 1620 he married Elizabeth, daughter of Sir James Bourchier, a merchant in the City of London. By her he was to have five sons and four daughters.

**Formative influences.** Both his father and mother came from Protestant families who had profited from the destruction of the monasteries during the reign of King Henry VIII, and it is probable that they influenced their son in his religious upbringing. Both his schoolmaster in Huntingdon and the Master of Sidney Sussex College were enthusiastic Calvinists and strongly anti-Catholic. In his youth Cromwell was not notably studious, being fond of outdoor sports, such as hunting; but he was an avid reader of the Bible, and he admired Sir Walter Raleigh's *The History of the World.* From his teachers and from his reading Cromwell learned that the sins of man were punishable on earth but that God, through His Holy Spirit, could guide the elect into the paths of righteousness.

During his early married life Cromwell, like his father, was profoundly conscious of his responsibilities to his fellow men and concerned himself with affairs in his native fenlands, but he was also the victim of a spiritual and psychological struggle that perplexed his mind and damaged his health. He does not appear to have experienced conversion until he was nearly 30; later he described to a cousin how he had emerged from darkness into light. Yet he had been unable to receive the grace of God without feeling a sense of "self, vanity and badness." He was convinced that he had been "the chief of sinners" before he learned that he was one of God's Chosen.

Career as a country squire

Cromwell also had financial worries until, at the age of 39, he inherited property at Ely from his mother's brother. Like other lesser gentry, he contended with bad harvests and a variety of taxes and impositions, such as ship money, exacted by the monarchy not only to pay for the upkeep of the navy but to sustain the lavish tastes of the court. Though in 1628 he had been elected a member of Parliament for the borough of Huntingdon, King Charles I dissolved this Parliament in 1629 and did not call another for 11 years.

During the interval, country gentlemen like Cromwell accumulated grievances. The Cromwell family was but one of a network of dissatisfied gentry who belonged to what one might call the political nation: for example, John Hampden, the wealthy Buckinghamshire squire who brought a test case against the crown over the levying of ship money, was Cromwell's first cousin. Thus, when in the spring of 1640 Cromwell was elected member of Parliament for the borough of Cambridge, partly because of the important social position he held in Ely and partly because of his fame as "Lord of the Fens," he found himself among a host of friends at Westminster who, led by John Pym, a veteran politician from Somerset, were highly critical of the monarchy. Little was achieved by the Short Parliament (dissolved after three weeks), but, when in November 1640 Cromwell was again returned by Cambridge to what was to be known as the Long Parliament, which sat until 1653, his public career began.

**Cromwell in Parliament.** Cromwell had already become known in the Parliament of 1628–29 as a fiery and somewhat uncouth Puritan, who had launched an attack on Charles I's bishops. He believed that the individual Christian could establish direct contact with God through prayer and that the principal duty of the clergy was to inspire the laity by preaching. Thus he had contributed out of his own pocket to the support of itinerant Protestant preachers or "lecturers" and openly showed his dislike of his local bishop at Ely, a leader of the High Church party, which stood for the importance of ritual and episcopal authority. He criticized the bishop in the House of Commons and was appointed a member of a committee to investigate other complaints against him. Cromwell, in fact, distrusted the whole hierarchy of the Church of England, though he was never opposed to a state church. He therefore advocated abolishing the institution of the episcopate and the banning of a set ritual as prescribed in *The Book of Common Prayer.* He believed that Christian congregations ought to be allowed to choose their own ministers,

Attacks on the church

who should serve them by preaching and extemporaneous prayer. Though he shared the grievances of his fellow members over taxes, monopolies, and other burdens imposed on the people, it was his religion that first brought him into opposition to the King's government. When in November 1641 John Pym and his friends presented to King Charles I a "Grand Remonstrance," consisting of over 200 clauses, among which was one censuring the bishops "and the corrupt part of the clergy, who cherish formality and superstition" in support of their own "ecclesiastical tyranny and usurpation," Cromwell declared that had it not been passed by the House of Commons he would have sold all he had "the next morning, and never have seen England more."

The Remonstrance was not accepted by the King, and the gulf between him and his leading critics in the House of Commons widened. A month later Charles vainly attempted to arrest five of them for treason: Cromwell was not yet sufficiently prominent to be among these. But when in 1642 the King left London to raise an army, and events drifted toward civil war, Cromwell began to distinguish himself not merely as an outspoken Puritan but also as a practical man capable of organization and leadership. In July he obtained permission from the House of Commons to allow his constituency of Cambridge to form and arm companies for its defense, in August he himself rode to Cambridge to prevent the colleges from sending their plate to be melted down for the benefit of the King, and as soon as the war began he enlisted a troop of cavalry in his birthplace of Huntingdon. As a captain he made his first appearance with his troop in the closing stages of the Battle of Edgehill (October 23, 1642) where Robert Devereux, 3rd earl of Essex, was commander in chief for Parliament in the first major contest of the war.

### MILITARY AND POLITICAL LEADER

Military organization

During 1643 Cromwell acquired a reputation both as a military organizer and a fighting man. From the very beginning he had insisted that the men who served on the parliamentarian side should be carefully chosen and properly trained, and he made it a point to find loyal and well-behaved men regardless of their religious beliefs or social status. Appointed a colonel in February, he began to recruit a first-class cavalry regiment. While he demanded good treatment and regular payment for his troopers, he exercised strict discipline. If they swore, they were fined; if drunk, put in the stocks; if they called each other Roundheads—thus endorsing the contemptuous epithet the Royalists applied to them because of their close-cropped hair—they were cashiered; and if they deserted, they were whipped. So successfully did he train his own cavalrymen that he was able to check and re-form them after they charged in battle. That was one of Cromwell's outstanding gifts as a fighting commander.

Throughout 1643 he served in the eastern counties that he knew so well. These formed a recognized centre of parliamentary strength, but, unwilling to stay on the defensive, Cromwell was determined to prevent the penetration of Yorkshire Royalists into the eastern counties and decided to counterattack. By re-forming his men in a moment of crisis in the face of an unbeaten enemy, he won the Battle of Gainsborough in Lincolnshire on July 28. On the same day he was appointed governor of the Isle of Ely, a large plateau-like hill rising above the surrounding fens, that was thought of as a possible bastion against advancing Royalists. In fact, however, Cromwell, fighting alongside the parliamentary general Sir Thomas Fairfax, succeeded in stemming the royalist attacks at Winceby in Lincolnshire and then successfully besieged Newark in Nottinghamshire. He was now able to persuade the House of Commons, well pleased with these victories, to create a new army, that would not merely defend eastern England but would march out and attack the enemy.

This new army was formed under the command of Edward Montagu, 2nd earl of Manchester, early in 1644. Appearing in the House of Commons, Cromwell, besides commending Manchester for the command, accused some of his fellow officers as incompetents or as being "profane" and "loose" in their conduct. Although not all members of the House of Commons approved of Cromwell's using his political position to defame other officers, his friends rallied round him, and in 1644 he was appointed Manchester's second in command, with the rank of lieutenant general, and paid five pounds a day. After an alliance had been concluded with the Scots, he was also appointed a member of the Committee of Both Kingdoms, which became responsible for the overall strategy of the Civil War. But since he was engaged at the front during the campaigning season, Cromwell took little part in its deliberations.

After Manchester's army had stormed Lincoln in May 1644, it marched north to join the Scots and the Yorkshire parliamentarians at the siege of York. But Charles I's commander in chief, Prince Rupert, raised the siege. He was, however, defeated in the Battle of Marston Moor, July 2, 1644, that in effect gave the north of England to Parliament. Cromwell had again distinguished himself in the battle, and when Manchester's army returned to eastern England to rest on its laurels, Cromwell criticized his superior officer for his slowness and lethargy. He did not believe that Manchester really wanted to win the war, and in mid-September he laid his complaints before the Committee of Both Kingdoms. The quarrel between the two commanders was patched up, but after the defeat at Newbury, caused largely by the earl of Manchester's refusal to support Cromwell's cavalry with his infantry, it broke into the open once more.

Removal of Lord Manchester

Cromwell now expounded his detailed complaint about Manchester's military conduct in the House of Commons. Manchester retorted by attacking Cromwell in the House of Lords. It was even planned to impeach Cromwell as "an incendiary." Once again, however, these quarrels were patched up. In December 1644, Cromwell proposed that in the future no members of either house of Parliament should be allowed to hold commands or offices in the armed forces; his proposal was accepted, and it was also agreed that a new army should be constituted under Sir Thomas Fairfax. Cromwell, an admirer of Fairfax, put forward his name and then busied himself with planning the new army, from which, as a member of Parliament, he himself was excluded. But, significantly, the post of second in command was left open, and, when the Civil War reached its climax in the summer of 1645, Fairfax insisted that Cromwell should be appointed to it. He then fought at the battles of Naseby and Langport, where Charles I's last two field armies were destroyed. In January 1646 the House of Commons awarded Cromwell £2,500 a year in confiscated Royalist land for his services and renewed his commission for a further six months. Thus he was able to join Fairfax in the siege of Oxford, from which Charles I escaped before it surrendered.

Cromwell was delighted with the way in which the war had gone since Fairfax had taken command of the new army and the lethargic earls of Essex and Manchester had been removed from their commands. He attributed these victories to the mercy of God and demanded that the men who had served the country so faithfully should have their due reward. After Naseby he wrote to the Speaker of the House of Commons urging that such "honest men" should not meet with discouragement: "He that ventures his life for the liberty of his country, I wish he trust God for the liberty of his conscience, and you for the liberty he fights for."

Postwar discord

But once the war was over the House of Commons wanted to disband the army as cheaply and quickly as possible. Disappointed, Cromwell told Fairfax in March 1647 that "never were the spirits of men more embittered than now." He devoted himself to trying to reconcile the Parliament with the army and was appointed a parliamentary commissioner to offer terms on which the army could be disbanded except for those willing to take part in a campaign in Ireland. As late as May he thought that the soldiers might agree to disband but that they would refuse to serve in Ireland and that they were "under a deep sense of some sufferings." When the civilian leaders in the House of Commons decided that they could not trust the army and ordered it disbanded, while they hired a Scottish army to protect them, Cromwell, who never liked the Scots and thought that the English soldiers were

being disgracefully treated, left London and on June 4, 1647, threw in his lot with his fellow soldiers.

Cromwell as mediator

**Mediation and the Second Civil War.** For the remainder of this critical year he attempted to find a peaceful settlement of the kingdom's problems, but his task seemed insoluble; and soon his good faith was freely called into question. The army was growing more and more restive, and on the day Cromwell left London, a party of soldiers seized Charles I. Cromwell and his son-in-law, Henry Ireton, interviewed the King twice, trying to persuade him to agree to a constitutional settlement that they then intended to submit to Parliament. At that time Cromwell, no enemy of the King, was touched by his devotion to his children. His main task, however, was to overcome the general feeling in the army that neither the King nor Parliament could be trusted. When, under pressure from the rank and file, General Fairfax led the army toward the houses of Parliament in London, Cromwell still insisted that the authority of Parliament must be upheld; and in September he also resisted a proposal in the House of Commons that no further addresses should be made to the King. Just over a month later he took the chair at meetings of the General Council of the Army (which included representatives of the private soldiers known as Agitators) and assured them that he was not committed to any particular form of government and had not had any underhand dealings with the King. On the other hand, fearing anarchy, he opposed extremist measures such as the abolition of the monarchy and the House of Lords and the introduction of a more democratic constitution.

But all Cromwell's efforts to act as a mediator between army, Parliament, and King came to nothing when Charles I escaped from Hampton Court Palace, where he had been kept in honourable captivity, and fled to the Isle of Wight to open negotiations with Scottish commissioners offering to restore him to the throne on their terms. On January 3, 1648, Cromwell abandoned his previous position and, telling the House of Commons that the King was "an obstinate man, whose heart God had hardened," agreed to a vote of no addresses, which was carried. The Royalists, encouraged by the King's agreement with the Scots and the failure of Cromwell to unite Parliament and the army, took up arms again and the Second Civil War began.

The arrest of Charles I

General Fairfax first ordered Cromwell into Wales to crush a rising there and then sent him north to fight the Scottish army that invaded England in June. Though his army was inferior in numbers to that of the Scots and northern Royalists, he defeated them both in a campaign in Lancashire; then he entered Scotland and restored order there; finally he returned to Yorkshire and took charge of the siege of Pontefract. The correspondence he conducted during the siege with the governor of the Isle of Wight, whose duty it was to keep watch on the King, reveals that he was increasingly turning against Charles. Parliamentary commissioners had been sent to the island in order to make one final effort to reach an agreement with the King. But Cromwell told the governor that the King was not to be trusted, that concessions over religion must not be granted, and that the army might be considered a lawful power capable of ensuring the safety of the people and the liberty of all Christians.

While Cromwell, still not entirely decided on his course, lingered in the north, his son-in-law Ireton and other officers in the southern army took decisive action. They drew up a remonstrance to Parliament complaining about the negotiations in the Isle of Wight and demanding the trial of the King as a Man of Blood. While Cromwell still felt uncertain about his own views, he admitted that his army agreed with the army in the south. Fairfax now ordered him to return to London; but he did not arrive until after Ireton and his colleagues had removed from the House of Commons all members who favoured continuing negotiations with the King. Cromwell asserted that he had not been acquainted with the plan to purge the House, "yet since it was done, he was glad of it, and would endeavour to maintain it." Hesitating up to the last moment, Cromwell, pushed on by Ireton, by Christmas Day finally accepted Charles's trial as an act of justice. He was one of the 135 commissioners in the High Court of Justice and, when the King refused to plead, he signed the death warrant.

**First chairman of the Council.** After the British Isles were declared a republic and named the Commonwealth, Oliver Cromwell served as the first chairman of the Council of State, the executive body of a one-chamber Parliament. During the first three years following Charles I's execution, however, he was chiefly absorbed in campaigns against the Royalists in Ireland and Scotland. He also had to suppress a mutiny, inspired by a group known as Levellers, an extremist Puritan party said to be aiming at a "levelling" between rich and poor, in the Commonwealth army. Detesting the Irish as primitive, savage, and superstitious, he believed they had carried out a huge massacre of English settlers in 1641. As commander in chief and lord lieutenant, he waged a ruthless campaign against them, though when he refused quarter to most of the garrison at Drogheda near Dublin in September 1649, he wrote that it would "tend to prevent the effusion of blood for the future, . . . which otherwise cannot but work remorse and regret." On his return to London in May 1650 Cromwell was ordered to lead an army into Scotland, where Charles II had been acknowledged as its new king. Fairfax had refused the command; so on June 25 Cromwell was appointed captain general in his place. He felt more tender toward the Scots, most of whom were fellow Puritans, than toward the Catholic Irish. The campaign proved difficult, and during the winter of 1650 Cromwell was taken ill. But he defeated the Scots with an army inferior in numbers at Dunbar on September 3, 1650, and a year later, when Charles II advanced into England, Cromwell destroyed his army at Worcester.

Cromwell in Ireland

This battle ended the civil wars. Cromwell now hoped for pacification, a political settlement, and social reform. He pressed through an "act of oblivion" (amnesty), but the army became more and more discontented with Parliament. It believed that the members were corrupt and that a new Parliament should be called. Once again Cromwell tried to mediate between the two antagonists, but his sympathies were with his soldiers. When he finally came to the conclusion that Parliament must be dissolved and replaced, he called in his musketeers and on April 20, 1653, expelled the members from the House. He asserted that they were "corrupt and unjust men and scandalous to the profession of the Gospel"; two months later he set up a nominated assembly to take their place. In a speech on July 4 he told the new members that they must be just, and, "ruling in the fear of God," resolve the affairs of the nation.

The "Little Parliament"

Cromwell seems to have regarded this "Little Parliament" as a constituent body capable of establishing a Puritan republic. But just as he had considered the previous Parliament to be slow and self-seeking, he came to think that the Assembly of Saints, as it was called, was too hasty and too radical. He also resented the fact that it did not consult him. Later he described this experiment of choosing Saints to govern as an example of his own "weakness and folly." He sought moderate courses and also wanted to end the naval war begun against the Dutch in 1652. When in December 1653, after a coup d'etat planned by Major General John Lambert and other officers, the majority of the Assembly of Saints surrendered power into Cromwell's hands, he decided reluctantly that Providence had chosen him to rule. As commander in chief appointed by Parliament, he believed that he was the only legally constituted authority left. He therefore accepted an "Instrument of Government" drawn up by Lambert and his fellow officers by which he became lord protector, ruling the three nations of England, Scotland, and Ireland with the advice and help of a council of state and a Parliament, which had to be called every three years.

### ADMINISTRATION AS LORD PROTECTOR

Before Cromwell summoned his first Protectorate Parliament on September 3, 1654, he and his Council of State passed more than 80 ordinances embodying a constructive domestic policy. His aim was to reform the law, to set up a Puritan Church, to permit toleration outside it, to promote education, and to decentralize administration.

The resistance of the lawyers somewhat dampened his enthusiasm for law reform, but he was able to appoint good judges both in England and Ireland. He was strongly opposed to severe punishments for minor crimes, saying: "to see men lose their lives for petty matters . . . is a thing that God will reckon for." For him murder, treason, and rebellion alone were subject to capital punishment. During his Protectorate, committees known as Triers and Ejectors were set up to ensure that a high standard of conduct was maintained by clergy and schoolmasters. In spite of resistance from some members of his council Cromwell readmitted Jews into the country. He concerned himself with education, was an excellent chancellor of Oxford University, founded a college at Durham, and saw to it that grammar schools flourished as they had never done before.

**Foreign and economic policies.** In 1654 Cromwell brought about a satisfactory conclusion to the Anglo-Dutch War, which, as a contest between fellow Puritans, he had always disliked. The question then arose of how best to employ his army and navy. His Council of State was divided, but eventually he resolved to conclude an alliance with France against Spain. He sent an amphibious expedition to the Spanish West Indies, and in May 1655 Jamaica was conquered. As the price for sending an expeditionary force to Spanish Flanders to fight alongside the French he obtained possession of the port of Dunkirk. He also interested himself in Scandinavian affairs; although he admired King Charles X of Sweden, his first consideration in attempting to mediate in the Baltic was the advantages that would result for his own country. In spite of the emphasis Cromwell laid on the Protestant interest in some of his speeches, the guiding motive in his foreign policy was national and not religious benefit.

Conquest of Jamaica

His economic and industrial policy followed mainly traditional lines. But he opposed monopolies, which were disliked by the country and had only benefitted the court gentry under Queen Elizabeth and the first two Stuarts. For this reason the East Indian trade was thrown open for three years, but in the end Cromwell granted the company a new charter (October 1657) in return for financial aid. Satisfactory methods of borrowing had not yet been discovered; hence—like those of practically all European governments of his time—Cromwell's public finances were by no means free from difficulties.

**Relations with Parliament.** When Cromwell's first Parliament met he justified the establishment of the Protectorate as providing for "healing and settling" the nation after the civil wars. Arguing that his government had prevented anarchy and social revolution, he was particularly critical of the Levellers who, he said, wished to destroy well-tested institutions "whereby England hath been known for hundreds of years." He believed that they wanted to undermine "the 'natural' magistracy of the nation" as well as "make the tenant as liberal a fortune as the landlord." He also thought that the spiritual anarchy that followed the destruction of the Anglican Church had gone too far, for now ordained preachers were frequently interrupted or shouted down in their pulpits. A radical in some directions, such as in seeking the reform of the laws, Cromwell now adopted a conservative attitude because he feared that the overthrow of the monarchy might lead to political collapse.

The four fundamentals

But vociferous republicans, who became leaders of this newly elected Parliament, were unwilling to concentrate on legislation, questioning instead the whole basis of Cromwell's government. Cromwell insisted that they must accept the "four fundamentals" of the new constitution that, he argued, had been approved both by "God and the people of these nations." The four fundamentals were government by a single person and Parliament; the regular summoning of parliaments, which must not be allowed to perpetuate themselves; the maintenance of "liberty of conscience"; and the division of the control of the armed forces between the protector and Parliament. Oliver said that he would sooner be "rolled into my grave and buried with infamy, than I can give my consent" to the "wilful throwing away of this Government, . . . so owned by God, so approved by men." He therefore required all members of Parliament, if they wished to keep their seats, to sign an engagement to be faithful to a protector and Parliament and to promise not to alter its basic character. Except for 100 convinced republicans, the members agreed to do so but were still more concerned with rewriting the constitution than reforming the laws as desired by the protector. As soon as he could legitimately do so (January 22, 1655), Cromwell dissolved Parliament.

But with his second Parliament, which he convened in 1656, he encountered exactly the same difficulty in the end, for the republican leaders, when they were allowed to resume their seats, tried to destroy the Protectorate on the ground that they were being forced to return to "an Egyptian bondage." Once again Cromwell emphasized that he had been "called" to power and that anarchy or an invasion from abroad would follow if his authority were not upheld. Thus in February 1658 he felt himself driven again to dissolve Parliament even though, as a former member, he understood only too well the gravity of his action.

Death and burial

Ever since the campaign in Ireland, Cromwell's health had been poor. In August 1658, after his favourite daughter, Elizabeth, died of cancer, he was taken ill with malaria and taken to London with the intention of living in St. James's Palace. But he died in Whitehall at three o'clock on September 3, the anniversary of two of his greatest victories. His body was secretly interred in Westminster Abbey on November 10, 13 days before his state funeral. In 1661, after the restoration of King Charles II, Cromwell's embalmed remains were dug out of the Westminster tomb and hung up at Tyburn where criminals were executed. His body was then buried beneath the gallows. But his head was stuck on a pole on top of Westminster Hall, where it is known to have remained until the end of Charles II's reign.

**Assessment.** Oliver Cromwell was by no means an extreme Puritan. By nature he was neither cruel nor intolerant. He cared for his soldiers, and when he differed from his generals he did not punish them severely. (For example, when he dismissed John Lambert he gave him a generous pension.) He was devoted to his old mother, his wife, and family. (The stories spread by Royalists that he was an admirer of a number of ladies have little substance to them.) While he concerned himself with the spiritual welfare of his children because he believed that "often the children of great men have not the fear of God before their eyes," he committed the mistake of not preparing for the practical tasks of government his eldest son, Richard, whom in the last days of his life he nominated to succeed him as protector. Music and hunting were among his recreations. He delighted in listening to the organ and was an excellent judge of horses. He was known to smoke, to drink sherry and small beer, and to prefer English food; he permitted dancing at the marriage of his youngest daughter. In his younger days he indulged in horseplay with his soldiers, but he was a dignified ruler. Sir Peter Lely, the famous Dutch painter, pictured him as he was in his prime (although the portrait was apparently not painted from life); the numerous paintings from life by Robert Walker dating from the beginning of the Civil War show him looking more of a fanatic.

Private life and religious beliefs

As lord protector, Cromwell was much more tolerant than in his fiery Puritan youth. Once bishops were abolished and congregations allowed to choose their own ministers, he was satisfied. Outside the church he permitted all Christians to practice their own religion so long as they did not create disorder and unrest. He allowed the use of *The Book of Common Prayer* in private houses and even the English Roman Catholics were better off under the protectorate than they had been before.

Although many Quakers were kept in prison for disturbing the peace, Cromwell was on friendly terms with George Fox, the founder of the society of Friends, and explored religious questions with him. When in the winter of 1656 a Quaker entered Bristol in imitation of Christ's entry into Jerusalem, Cromwell tried, though unsuccessfully, to save him from the fury of Parliament, which voted heavy punishments on the blasphemer. The year before, Cromwell interviewed two of the leaders of the Fifth Monarchy Men, an extreme sect: he pointed out to them that they had

been imprisoned for sedition but emphasized that no one would hinder them from preaching the Gospel of Christ.

Political views

In politics Cromwell held no fixed views except that he was opposed to what he called arbitrary government. Before the execution of Charles I he contemplated the idea of placing one of Charles's sons upon the throne. Cromwell also resisted the abolition of the House of Lords. In 1647 he said that he was not "wedded and glued" to any particular form of government. After the Assembly of Saints failed, he summoned two elected parliaments (1654–55 and 1656–58), but he was never able to control them. His failure to do so has been attributed to "lack of that parliamentary management by the executive which, in correct dosage, is the essential nourishment of any sound parliamentary life" (H.R. Trevor-Roper). In between these two parliaments (1655–56) he sanctioned the government of the country by major generals of the Horse Militia who were made responsible for law and order in groups of counties. But he soon abandoned this experiment when it met with protests and reverted to more normal methods of government. In the spring of 1657 he was tempted by an offer of the crown by a majority in Parliament on the ground that it fitted in better with existing institutions and the English common law. In the end he refused to become king because he knew that it would offend his old republican officers. Nevertheless, in the last year and a half of his life he ruled according to a form of government known as "the Petition and Advice." This in effect made him a constitutional monarch with a House of Lords whose members he was allowed to nominate as well as an elected House of Commons. But he found it equally difficult to govern either with or without parliaments.

Reputation

Although in the late 17th century Cromwell was execrated as a brave bad man, it was admitted that he had made his country great. In the 18th century, on the other hand, he was considered a nauseating hypocrite, while the 19th century, under the influence of the writer and historian Thomas Carlyle, regarded him as a constitutional reformer who had destroyed the absolutism of Charles I. Modern critics are more discriminating. His belief in God's providence is analyzed in psychological terms. Marxists blame him for betraying the cause of revolution by suppressing the radical movement in the army and resisting the policy of the Levellers. On the whole, he is regarded only in a very limited sense as a dictator, but rather as a patriotic ruler who restored political stability after the civil wars and contributed to the evolution of constitutional government and religious toleration. (M.As.)

**BIBLIOGRAPHY.** SIR CHARLES H. FIRTH, *Oliver Cromwell and the Rule of the Puritans in England* (1900, many times reprinted), is an excellent and fair biography. Other modern biographies include ROBERT S. PAUL, *The Lord Protector* (1955, reissued 1964), excellent on Cromwell's religious attitude; MAURICE ASHLEY, *The Greatness of Oliver Cromwell* (1957, reissued 1969), embodying some of the results of research since Firth's day; and ANTONIA FRASER, *Cromwell: The Lord Protector* (U.K. title, *Cromwell: Our Chief of Men*, 1973), a detailed study. ROY E. SHERWOOD, *The Court of Oliver Cromwell* (1978), is an examination of his court. IVAN ROOTS (ed.), *Cromwell: A Profile* (1973), is an anthology of essays by specialists, including an essay covering the 17th–20th centuries. J.E. CHRISTOPHER HILL, *God's Englishman* (1970), places Cromwell in perspective from a Marxist point of view. Other books on Cromwell include three by distinguished statesmen, French, American, and British: FRANÇOIS GUIZOT, *History of Oliver Cromwell . . .*, 2 vol. (1854; trans. from and also published in French, 1854); THEODORE ROOSEVELT, *Oliver Cromwell* (1900); and JOHN MORLEY, *Oliver Cromwell* (1900). Cromwell's letters and speeches are collected in *The Writings and Speeches of Oliver Cromwell*, 4 vol., ed. by WILBUR C. ABBOTT (1937–47); these volumes also contain a sound and detailed narrative of his career.

(M.As./Ed.)

# The Crusades

The term crusade is commonly used to refer to military expeditions organized by Western Christians against Muslim powers in order to take possession of or maintain control over the Holy City of Jerusalem and the places associated with the earthly life of Jesus Christ. Between 1095, when the First Crusade was launched, and 1291, when the Latin Christians were finally expelled from their bases in Syria, historians have formally enumerated eight major expeditions. Many other lesser ventures also took place, and even after 1291 there were attempts to recover what had been lost.

This period of roughly two centuries was one of significant social, economic, and institutional growth in western Europe. As a consequence, each of the Crusades reflected the particular conditions prevailing in Europe at the time, and their influence on Europe varied as new situations developed in the East.

This article is divided into the following sections:

The crusading movement and the first four Crusades 827
The First Crusade and the establishment of the Latin states 827
Background of the First Crusade
Preparations for the Crusade
The siege of Antioch
The siege of Jerusalem
The establishment of the crusader states
The era of the Second and Third Crusades 830
The Second Crusade
The crusader states to 1187
The institutions of the First Kingdom
The Third Crusade
The Fourth Crusade and the Latin Empire of Constantinople 834
The decline of the crusading movement 835
The later Crusades and the decline of the Latin enclaves 835
The Latin East after the Third Crusade
The Fifth Crusade
The Sixth Crusade
The Crusade of Louis IX of France
The final loss of the crusader states
The later Crusades and the Kingdom of Cyprus
The results of the Crusades 838

## The crusading movement and the first four Crusades

### THE FIRST CRUSADE AND THE ESTABLISHMENT OF THE LATIN STATES

**Background of the First Crusade.** Politically, western Europe in the last quarter of the 11th century was composed of several kingdoms loosely describable as feudal. While certain monarchies were already developing better integrated systems of government, many problems associated with feudalism, such as vassalage and inheritance, endemic private warfare, and brigandage, were still present. There were, of course, many variations, and in certain cases special conditions in a country precluded its significant participation in a crusade. Such, for example, were the problems of England's readjustment after the Norman Conquest in 1066, Spain's preoccupation with Muslim incursions from Africa, and civil strife in Germany consequent on long controversy with the popes. In fact, the principal support for the First Crusade came from France; and it was in France that the most typical aspects of feudalism, both negative and positive, could be found.

The economic revival of the 11th century

Europe in the 11th century was feeling the effects of a population growth that had begun toward the end of the 10th century and would continue well into the 13th century. Younger sons of the nobility, unless they married advantageously or entered the religious life, were more than ever disposed to adventure, organized or otherwise. At the same time, an economic revival was in full swing well before the First Crusade. Forest lands were being cleared, frontiers pushed forward, and markets organized, and the Muslim predominance in the Mediterranean was being challenged by Italian shipping. Nobles, bourgeois, and peasants were all seeking new outlets.

Especially significant for the Crusade was a general overhauling of the ecclesiastical structure, which enabled the popes to assume a more active role in society. In 1095, for example, Urban II, though still meeting resistance from the German Emperor, who opposed papal reform policies, was in a strong enough position to convoke two important ecclesiastical councils.

Thus it was that in the closing years of the 11th century western Europe was abounding in energy and full of confidence. What is more, as is evident in such achievements as the Norman Conquest of England, Europeans possessed the capacity to launch a major military undertaking. Why this energy was channelled into a holy war against Muslims in a distant land is uncertain, though the answer in part involves a new challenge from Eastern Islām presented by the incursions of the Seljuq Turks. The Seljuqs, originally one of several Turkish tribes on the northeastern borders of the Muslim world, had embraced Islām and by the 11th century were moving south and west into Iran and beyond. In large part, however, the answer also lies in the nature of popular religious life and feeling in the West. The two developments are not unconnected.

Factors in the growth of religious enthusiasm

The impact of religion on the lay person of the time is not easy to determine: his religion was unsophisticated, and he was moved by tales of signs and wonders and attributed natural disasters to supernatural intervention. At the same time, lay people were not indifferent to movements of reform, and there were instances of townsmen agitating against clergy whom they regarded as unworthy. There also developed, especially in France, under the leadership of certain bishops but with considerable popular support, what could be described as a peace movement. Many local areas proclaimed the Truce of God and the Peace of God, designed to halt or at least limit fratricidal strife and protect the lives of clergy, travellers, and other persons unable to protect themselves against brigandage. It is particularly interesting to note that the Council of Clermont at which Urban II preached the First Crusade in 1095 renewed and generalized the Truce of God.

While it may seem paradoxical that a council both promulgated peace and officially sanctioned war, it must be remembered that the peace movement was designed to protect those in distress, and a strong element in the Crusade was the idea of giving aid to fellow Christians in the East. Furthermore, there had developed in western Europe, partly under the influence of the Reconquista (the movement to reconquer Spain from the Muslims), the idea that war to defend Christian society was not only justifiable but a holy work, and therefore pleasing to God.

Closely associated with this Western concept of holy war was another popular religious practice, pilgrimage to a holy shrine. Eleventh-century Europe abounded in local shrines housing relics of saints, but three great centres of pilgrimage stood out above all the others: Rome, with the tombs of SS. Peter and Paul; Santiago de Compostela, in northwestern Spain; and Jerusalem, with the Holy Sepulchre of Christ's entombment. Moreover, pilgrimage, which had always been considered an act of devotion, had come to be regarded as a more formal expiation for serious sin, even occasionally prescribed as a penance to the sinner by his confessor.

The importance of Jerusalem

Yet another element in the popular religious consciousness of the 11th century, one associated with both Crusade

and pilgrimage, was a widespread belief that the end of the world was imminent. This belief was once thought to have been associated with the year 1000 (the millennium). Scholars now, however, tend to discount the idea that contemporaries had so precise a date in mind and emphasize the continuance of the idea through the 11th century and even beyond. Moreover, in certain late 11th-century portrayals of the end of all things, the "last emperor," now popularly identified with the "king of the Franks," the final successor of Charlemagne, was to lead all the faithful to Jerusalem and there await the Second Coming of Christ. Manifestly, Jerusalem, as the earthly symbol of the heavenly city, figured prominently in Western consciousness. And it was becoming clear, as the number of pilgrimages to Jerusalem increased in the 11th century, that any interruption would have serious repercussions.

By the middle of the 11th century the Seljuqs had wrested political authority from the ʿAbbāsid caliphs of Baghdad. Their policy, originally directed southward against the Fāṭimids of Egypt, was increasingly diverted by the pressure of Turkman raids into Anatolia and Byzantine Armenia. A badly organized Byzantine army was defeated at Manzikert in 1071, opening Asia Minor to eventual Turkish occupation. Meanwhile, many Armenians south of the Caucasus migrated south to join others in the region of the Taurus Mountains and to form a colony in Cilicia.

Seljuq expansion southward continued, and in 1085 the capture of Antioch in Syria was another blow to Byzantine prestige. Thus, although the Seljuq Empire was never successfully held together as a unit, the Byzantines, driven from all but the coastal regions of Asia Minor, and with Nicaea in Turkish hands by 1092, faced a resurgent Islām perilously close to the capital. It was this danger that prompted the Byzantine emperor Alexius Comnenus to seek aid from the West, and by 1095 the West was ready to respond.

The turmoil of these years disrupted normal political life and made the pilgrimage to Jerusalem difficult and often impossible. Stories of dangers and molestation reached the West and remained in the popular mind even after conditions improved. Further, informed authorities began to realize that the revived power of the Muslim world now seriously menaced the West as well as East. It was this realization that stimulated official and organized action.

Alexius' appeal came at a time when relations between the Eastern and Western branches of the Christian world were improving. Difficulties between the two in the middle years of the century had resulted in a de facto, though not formally proclaimed, schism, and ecclesiastical disagreements had been accentuated by Norman occupation of formerly Byzantine areas in southern Italy. A campaign of the Norman adventurer Robert Guiscard against the Greek mainland further embittered the Byzantines, and it was only after Robert's death in 1085 that conditions for a renewal of normal relations between East and West were reasonably favourable. Envoys of Emperor Alexius Comnenus thus arrived at the Council of Piacenza in 1095 at a propitious moment; and it seems probable that Pope Urban II viewed military aid as a means toward restoring ecclesiastical unity.

**The Council of Clermont**

The Council of Clermont convoked by Urban on November 18, 1095, was attended largely by bishops of southern France. There was also a scattering of representatives from the north and elsewhere. After the transaction of many important items of ecclesiastical business, which resulted in a series of canons, among them one renewing the Truce of God, there was included a canon that granted a plenary indulgence (the remission of all penance for sin) to those who undertook to aid Christians in the East. Then in a great outdoor assembly the pope, himself a Frenchman, addressed a large crowd. We are not certain what language Urban spoke to these people, but apparently it was a language they could understand.

Precisely what the Pope said can never be known, since the only surviving accounts of his speech were written years later. But he apparently stressed the plight of Eastern Christians, the molestation of pilgrims, and the desecration of the holy places. He urged those of his hearers who were guilty of disturbing the peace to turn their warlike energies toward a holy cause. He emphasized the need for penance, for the acceptance of suffering. No one should undertake this pilgrimage for any but the most exalted of motives.

The response was immediate and overwhelming, probably far greater than Urban had anticipated. Cries of "Deus volt" ("God wills it") were heard everywhere, and it was decided that those who agreed to go were to wear a cross. Moreover, it was not only warrior knights who responded; a popular element, apparently unexpected and probably not desired, also came forward.

The era of Clermont witnessed the concurrence of three significant developments: first, there existed as never before a popular religious fervour that was not without marked eschatological tendencies in which the Holy City of Jerusalem figured prominently; second, war against the infidel had come to be regarded as a religious undertaking, a work pleasing to God; and finally, from an organizational standpoint, ecclesiastical as well as secular, western Europe possessed the capacity to plan such an enterprise and carry it through.

**Preparations for the Crusade.** Following Pope Urban's speech, preparations began in both East and West. Emperor Alexius, who had doubtless anticipated some sort of auxiliary force, apparently soon realized that he would have to provide for and police a much larger influx of warriors. In the West, as the leaders began to assemble their armies, persons who took the cross sought to raise money, often by selling or mortgaging property, both for the immediate purchase of equipment and for the long-term needs of the journey.

**The People's Crusade**

As this was being done, several less organized and generally miscellaneous bands, including some armed men, commonly known as the "People's Crusade," set out across Europe. The most famous of these, brought together by a remarkable popular preacher, Peter the Hermit, and his associate Walter the Penniless, arrived at Constantinople after causing considerable disorder in Hungary and Bulgaria. Alexius received Peter cordially and advised him to await the arrival of the main Crusade force. But the rank and file grew unruly, and on August 6, 1096, they were ferried across the straits. As Peter returned to beg for aid, they were ambushed at Cibotus (called Civetot by the Franks) and all but annihilated by the Turks. Other, even less well-disciplined bands, largely from northwest Germany, where they committed atrocities against the Jewish communities, were dispersed before reaching the Byzantine frontier.

The main Crusading force, which began to move as Urban had directed in August 1096, consisted of four major contingents. Hugh of Vermandois, brother of King Philip I of France, left first with a small following that was reduced by shipwreck while crossing the Adriatic from Bari to Dyrrhachium. Godfrey of Bouillon, duke of Lower Lorraine since 1089, the only major prince from the German kingdom, though he and his associates largely spoke French, was joined by his brothers, Eustace and Baldwin, and a kinsman, Baldwin of Le Bourg. Taking the land route, Godfrey crossed Hungary without incident. Markets and provisions were supplied in Byzantine territory, and, except for some pillaging, the army reached Constantinople without serious incident in late December.

Second, Bohemond, a Norman from southern Italy and the son of Robert Guiscard, was thus on familar ground when he crossed the Adriatic, where he had fought with his father and was understandably feared by the Byzantines. But he was now 40 years old and determined to come to profitable terms with his former enemy. He arrived at Constantinople on April 9, 1097.

Third, the largest army was that assembled by Raymond of Saint-Gilles, the count of Toulouse. Fifty-five years of age, he was the oldest and most prominent of the Crusading princes, and he aspired and perhaps expected to become the leader of the entire expedition. He was accompanied by Adhémar, bishop of Le Puy, whom the Pope had named as legate for the Crusade. Raymond led his followers, including a number of noncombatant pilgrims whom he supported at his own expense, across north Italy, around the head of the Adriatic, thence southward into Byzantine territory. This large body caused considerable

trouble in Dalmatia and with Byzantine troops policing the area nearer the capital, where Raymond arrived on April 27.

Meanwhile, the fourth army, under Robert of Flanders, had crossed the Adriatic from Brindisi. Accompanying Robert were his cousin Robert of Normandy (brother of King William II Rufus of England) and Stephen of Blois (the son-in-law of William I the Conqueror). No king took part in the First Crusade, and the predominantly French-speaking participants came to be known as Franks.

Latin–Byzantine friction

The presence near Constantinople of massive military forces, numbering perhaps 4,000 mounted knights and 25,000 infantry, posed a serious problem for Alexius, and there was occasional disorder. Forced to consider the permanent imperial interests, which, it soon became evident, were different from the objective of the crusaders, the Emperor required each Crusade leader to promise under oath to restore any conquered territory that had belonged to the empire before the Turkish invasions and to swear allegiance to him for any lands occupied beyond the former frontiers. It was a reasonable demand, and all the leaders took the oath before crossing the straits, though Godfrey and Raymond did so only after some hesitation. Raymond, in fact, agreed only to a modified form of oath common in southern France, and in the end he remained of all the Crusade leaders the most loyal to Alexius.

**The siege of Antioch.** Late in May 1097, crusaders, accompanied by a Byzantine contingent, appeared before Nicaea. On June 19, when the city fell, it was surrendered, as had been agreed, to the Byzantines. Crossing arid and mountainous Anatolia proved a difficult task. At Dorylaeum, on July 1, 1097, Turks attacked the advance column of the crusader army. Despite the heat and a rain of arrows that fell on them, the crusaders held their ground, and, when the rest of the army drew up, the Turks were routed. A major victory in open warfare had been achieved by cooperation among the separate contingents and the Greeks.

Further advance across Anatolia was even more arduous, and it was only after suffering many casualties, especially in the region of the Taurus Mountains, that the crusaders arrived near Antioch on October 20. Meanwhile, Godfrey's brother Baldwin left the main army to become involved in Armenian politics and he became ruler of Edessa.

One of the great cities of the Levant, Antioch was surrounded by a great circle of walls, and despite reinforcements and supplies by Genoese and English ships and later by the Patriarch of Jerusalem, then in Cyprus, the siege proved long and difficult. Spring brought the threat of attack by a relief force under Karbūgah of Mosul, and some of the crusaders, among them Stephen of Blois, became discouraged and left for home. In Asia Minor, Stephen met Alexius, who had set out as he had agreed to do, and convinced him that Antioch's cause was hopeless. The Emperor's decision to turn back, however justified tactically, was a diplomatic blunder; when the crusaders learned of the Emperor's move, they felt free from any obligation to return the city to him.

Bohemond meanwhile had proposed that whoever first entered the city should have possession, provided the Emperor did not make an appearance. The Norman had, in fact, already made contact with a discontented commander within, who proceeded to admit him over a section of the walls. The other crusaders followed into the city on June 3, 1098, and massacred the Muslim inhabitants. Only the citadel held out.

Thus, Antioch was occupied, and it is significant that the Greek Patriarch who had been imprisoned was restored, suggesting that neither Adhémar nor presumably Urban II had any plan to substitute a Latin incumbent. The victory, however, was incomplete, and Karbūgah soon blockaded the city. Disagreements among the leaders persisted and were accentuated by unseemly arguments over the validity of what had come to be called the Holy Lance, which a Provençal priest found below the cathedral and insisted was the lance that, according to the Gospels, had pierced the side of Christ when he hung on the cross. Nonetheless, on June 28 the crusader army moved out of the city and after a day of fierce fighting defeated Karbūgah's forces. The citadel then surrendered to Bohemond, and its garrison was permitted to leave. Rejoicing was tempered by a devastating epidemic that took many lives, including that of the legate, Adhémar of Le Puy, who, as the spiritual leader of the Crusade, had been a wise counsellor and a stabilizing influence whom the leaders could ill afford to lose.

The death of Adhémar of Le Puy

As the leaders continued to confer on the final disposition of Antioch, the rank and file of the army grew impatient. In January, following the siege of Maʿarrat an-Nuʿmān, some miles to the south, there was a popular demonstration in favour of moving on. The army then set out for Jerusalem under the leadership of Raymond of Saint-Gilles. As they moved south, Tancred and Robert of Normandy and, later, Godfrey and Robert of Flanders joined them. Bohemond, fearing for the safety of Antioch, turned back soon after leaving.

**The siege of Jerusalem.** Not far from Beirut, the army entered the territory of the Fāṭimid caliphs of Cairo, who, as Shīʿite Muslims, were enemies of the Sunnite Seljuqs and the caliphs of Baghdad. In August 1098 the Fāṭimids had occupied Jerusalem. The final drive of the First Crusade, therefore, was against the Fāṭimids of Egypt, not the Seljuqs.

On June 7, 1099, the Christian army, by then considerably reduced to perhaps 1,200–1,500 cavalry and 12,000 foot soldiers, encamped before Jerusalem, the governor of which was well supplied and confident that he could withstand a siege until a relief force arrived from Egypt. The crusaders, on the other hand, were short of supplies and would be until six vessels arrived at Jaffa (Yafo) and managed to unload before the port was blockaded by an Egyptian squadron. On July 8 a strict fast was ordered, and, with the Muslims scoffing from the walls, the entire army, preceded by the clergy, marched in solemn procession around the city, thence to the Mount of Olives, where Peter the Hermit preached with his former eloquence.

Siege towers were brought up to the walls on July 13–14, and on July 15 Godfrey's men took a sector of the walls, and others followed on scaling ladders. When the nearest gate was opened, Tancred and Raymond entered, and the Muslim governor surrendered to the latter in the Tower of David. The governor along with his bodyguard was escorted out of the city. Tancred promised protection in the Aqṣā Mosque, but his orders were disobeyed. All Muslims, men, women, and children, as well as Jews, perished in the general slaughter that followed.

Massacre of Jerusalem's inhabitants

Thus it was that, three long years after they had set out, the crusaders attained their goal. Doubtless some participants had been moved more by hopes of material gain or sheer love of adventure than by religious devotion. Nonetheless, it is difficult to believe that the Crusade could have succeeded without an extraordinary dedication to the ultimate objective, the liberation of the Holy City of Jerusalem.

**The establishment of the crusader states.** After a successful surprise attack on the Egyptian relief army, which confirmed the occupation of Palestine, most of the crusaders, having fulfilled their vows of pilgrimage, departed for their homes and left the problem of governing the conquered territories in the hands of the few who remained. Initially, there was disagreement as to the nature of the government to be established, and some held that the Holy City should be ruled under ecclesiastical authority. As an interim measure, Godfrey was elected to govern, and took the modest title of defender of the Holy Sepulchre. Nevertheless, Raymond, who had previously declined the offer of the crown, chose to leave immediately.

In December 1099, in the midst of this somewhat confused situation, Bohemond and Baldwin of Edessa arrived in Jerusalem to fulfill their crusader vows. Accompanying Bohemond was Daimbert, the archbishop of Pisa, who was chosen patriarch and proceeded to receive the homage of both Godfrey and Baldwin. If Daimbert had ambitions to govern Jerusalem, they were thwarted when, on Godfrey's death, his brother Baldwin was summoned back to Jerusalem, where he assumed the title of king (November 11, 1100). Thus, there had come into being not a church state but a feudal Kingdom of Jerusalem.

Though the Crusade of 1101, which Pope Paschal II had organized to bring reinforcements, collapsed in Asia Minor, King Baldwin profited from the chronic rivalries among his Muslim neighbours. He was further aided in extending the coastline by Italians and in one instance by a Norwegian squadron that arrived under King Sigurd in 1110. By 1112 Arsuf, Caesarea, Acre, Beirut, and Sidon had been taken, and the entire coast except for Ascalon and Tyre was in Latin hands. Meanwhile, castles had been built in Galilee and the frontier pushed southward.

The County of Edessa

These same years witnessed the formation of three other crusader states to the north. The County of Edessa, an ill-defined domain extending into the upper Euphrates region with a population consisting mainly of Armenians and Syrians, had been established by Godfrey's brother Baldwin. When Baldwin left to become ruler of Jerusalem he bestowed the county, under his suzerainty, on his cousin Baldwin of Le Bourg.

Antioch had not been returned to the Emperor, and Bohemond had consolidated his position there. The city was predominantly Greek in population, though there were also Syrians and Armenians, and the latent Greek-Latin friction was intensified when Bohemond replaced the Greek Patriarch by a Latin. When Bohemond was captured by the Muslims in 1100, his nephew Tancred became regent and expanded the frontiers of the principality to include the important port of Latakia, taken from the Byzantines in 1103. Not long after his release in 1103, Bohemond travelled to Europe, where he succeeded in winning over Pope Paschal II to the idea of a new Crusade. Whatever the original intention, there resulted not an expedition against Muslims but an attack on the Byzantine city of Dyrrhachium. Like its predecessor, the ill-fated campaign of 1081, the enterprise failed, and in 1108 Bohemond was forced to take an oath of vassalage to the Emperor for Antioch and to return to Italy, where he died in 1111. Though Tancred, again in power, disregarded his uncle's oath, Antioch and its patriarchate remained a source of controversy.

A fourth crusader state was established on the coast in the vicinity of Tripoli (Arabic: Tarābulus) by Raymond of Saint-Gilles, who had been outmanoeuvred in Jerusalem and had returned to Constantinople hoping for aid from the Byzantine Emperor to whom he had always been loyal. But the Crusade of 1101, which he led, had been another blow to his prestige. In 1102 he returned to Syria, took Tortosa (Ṭarṭūs), and began the siege of Tripoli. But he died in 1105, and it remained for his descendants to finish the task. Tripoli fell in 1109.

The establishment and protection of the frontiers was, for the new states, a problem conditioned by two factors, geography and the politics of Levantine Islām. From Antioch south, the crusaders had occupied a narrow strip of coastland bounded by mountains to the north and by the Jordan Valley in the south. To the east beyond the Syrian desert lay the Muslim cities of Aleppo, Ḥamāh, Homs, and Damascus. Though the Franks did push southward to Aylah (or Elim, modern al-ʿAqabah), all attempts to move eastward failed, and it was necessary to erect castles at vulnerable points along the eastern frontier as well as along the coast and inland. Among the most famous of these were Krak de Montréal, in the Transjordan, and Krak des Chevaliers, in the County of Tripoli. Meanwhile, the hostility between Shīʿite Egypt and Sunnite Baghdad continued for some time, while the emirates in between remained divided in their allegiance and those in the north fearful of the Seljuqs of Iconium.

After Baldwin I's death in 1118, the throne passed to his cousin Baldwin of Le Bourg, who left Edessa to another cousin, Joscelin of Courtenay. In 1124, Tyre, the last great city north of Ascalon still in Muslim hands, was taken with the aid of the Venetians, who, as was customary, demanded a section of the city. Baldwin II was succeeded by Fulk of Anjou, a newcomer recommended by Louis VI of France, who married Baldwin's daughter Melisend. In 1131 Baldwin and Joscelin both died; they were the last of the first generation of crusaders, and with their passing the formative period in the history of the crusader states came to an end.

End of the period of expansion

Fulk's policies represented an end of expansion and a stabilization of gains made, a wise course, because his reign coincided with the rise of Zangī, *atabeg* (governor) of Mosul, whose achievements earned him the reputation of a great champion of the *jihād* (holy war) against the Franks. When Zangī moved against Damascus, the recognition of their common danger resulted in a Jerusalem-Damascus alliance, a kind of diplomacy by then common with the second-generation Franks.

The north, however, was in great danger. The Byzantines had recovered their influence in Anatolia and were putting pressure on Armenia and Antioch. Emperor Manuel Comnenus forced Prince Raymond of Antioch to acknowledge the imperial suzerainty. But the greater danger to both was dramatically brought home by Zangī's capture of Edessa in 1144. Attempts at recovery failed, and the northernmost crusader state was subsequently overrun.

## THE ERA OF THE SECOND AND THIRD CRUSADES

**The Second Crusade.** It had long been apparent that Edessa was vulnerable, but its loss came as a shock to Christians both East and West. Urgent pleas for aid soon reached Europe, and Pope Eugenius III in 1145 issued a formal crusade bull, the first of its kind, with precisely worded provisions designed to protect crusaders' families and property and reflecting contemporary advances in canon law. Legates were designated. The Crusade, energetically supported by Louis VII, was preached by St. Bernard of Clairvaux in France and, with interpreters, even in Germany. As in the First Crusade, many simple noncombatant pilgrims responded. But, since Emperor Conrad III, though at first reluctant to leave Germany, had been won over by Bernard's eloquence, the Second Crusade, unlike the First, was led by two of Europe's major rulers.

Conditions in the East in mid-12th century

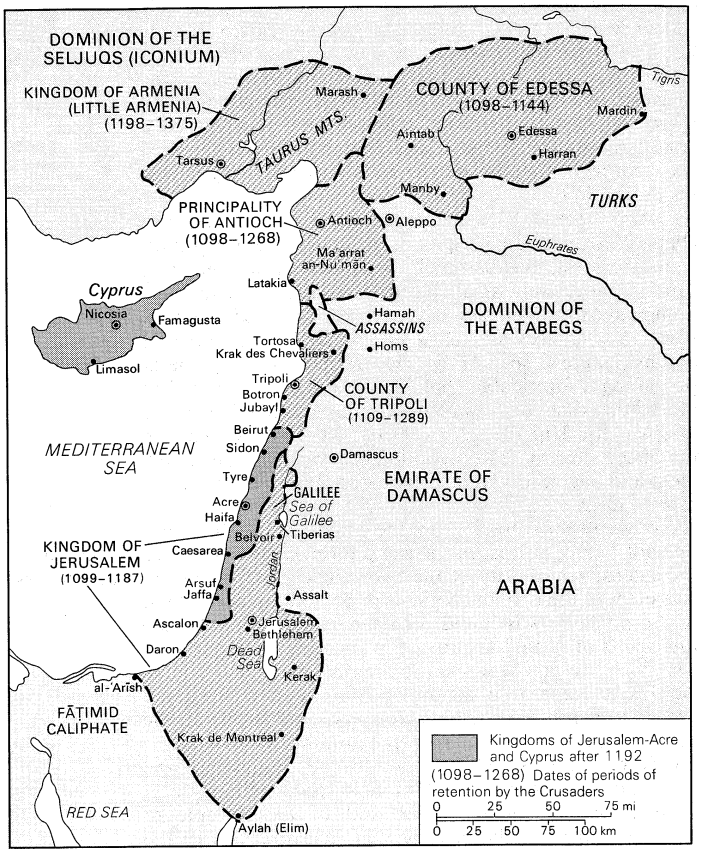
The situation in the East was also different. Manuel Comnenus was favourably disposed toward the West and, because of his interest in Antioch, was concerned over the power of Zangī, the Muslim ruler of Aleppo. He was able to assist the crusaders with guides and supplies but contributed no troops. In anticipation of the arrival of the western armies and more aware than they of the delicate power balance in the Levant, he had made peace with the Seljuqs of Iconium in 1146. Above all, Manuel was alarmed by the possibility of an attack by Roger II, the Norman king of Sicily.

Conrad left in May 1147, accompanied by many German nobles, the Kings of Poland and Bohemia, and Frederick of Swabia, his nephew and heir. On leaving Hungary and entering Byzantine territory, he agreed to an oath of noninjury. Conrad's poorly disciplined troops created tension with the Byzantines in Constantinople, where they arrived in September. Both Conrad and Manuel, however, remained on good terms, and both were apprehensive about the moves of Roger of Sicily, who during these same weeks seized Corfu and attacked the Greek mainland.

Conrad, rejecting Manuel's advice to follow the coastal route around Asia Minor, moved his main force past Nicaea directly into Anatolia. On October 25, at Dorylaeum, not far from the place where the first crusaders won their victory, his army, weary and without adequate provisions, was set upon by the Turks and virtually destroyed. Conrad, with a few survivors, retreated to Nicaea.

Meanwhile, about a month behind the Germans, Louis VII, accompanied by his wife Eleanor of Aquitaine, followed the land route across Europe and arrived at Constantinople on October 4. A few of his more hotheaded followers, on hearing that Manuel had made a truce with the Turks of Iconium and totally misunderstanding the motives, accused the Emperor of treason and urged the French king to join Roger in attacking the Byzantine emperor. Louis preferred the opinion of his less volatile advisers and agreed to restore any imperial possessions he might capture.

In November the French reached Nicaea, where they learned of Conrad's defeat. Louis and Conrad then started along the coastal route, the French now in the van, and reached Ephesus. Conrad became seriously ill and returned to Constantinople to the medical ministrations of Manuel. He eventually reached Acre by ship in April 1148.




The crusader states of the 12th century.
From W. Shepherd, *Historical Atlas*; Barnes & Noble Books, New York City

The French passage from Ephesus to Antioch in midwinter was harrowing in the extreme. Supplies ran short, and the Byzantines were blamed. But Antioch was finally reached in March 1148, and the crusaders were welcomed by Prince Raymond, Queen Eleanor's uncle. Raymond urged an attack on Aleppo, the centre of power of Nureddin, son and successor of Zangī. But King Louis, who resented Eleanor's open espousal of Raymond's project, left abruptly for Jerusalem and forced the Queen to join him.

At Jerusalem, where Conrad had already arrived, a brilliant assembly of French and German notables assembled with Queen Melisend, her son Baldwin III, and the barons of Jerusalem to discuss how best to proceed. Despite the absence of the northern princes and the losses already suffered by the crusaders, it was possible to field an army of nearly 50,000 men, the largest Crusade army so far assembled. After considerable debate, which revealed the conflicting purposes of crusaders and Jerusalem barons, it was decided to attack Damascus.

Attack on Damascus

How the decision was reached is not known. Damascus was undoubtedly a tempting prize, but Unur, the Turkish commander, also fearful of the expanding power of Nureddin and the one Muslim ruler most disposed to cooperate with the Franks, was now forced to seek the aid of his former enemy. And Nureddin was not slow to move toward Damascus. Not only was the campaign mistakenly conceived, it was badly executed. On July 28, after a five-day siege, with Nureddin's forces nearing the city, it became evident that the crusader army was dangerously exposed, and a retreat was ordered. It was a humiliating failure, attributable largely to the conflicting interests of the participants.

Conrad left immediately and stopped at Constantinople, where he agreed to join the Emperor against Roger of Sicily. Louis's reaction was different. His resentment against Manuel, whom he blamed for the failure, was so great that he accepted Roger's offer of ships to take him home and agreed to a plan for a new Crusade against Byzantium. Fortunately the plan was soon abandoned.

The Second Crusade had been promoted with great zeal and had aroused high hopes. Its collapse caused deep dismay, and subsequent expeditions, perhaps wisely, were more limited in their objectives. The Muslims, on the other hand, were enormously encouraged. They had confronted the danger of another major Western expedition and had triumphed over it.

**The crusader states to 1187.** During the 25 years following the Second Crusade the Kingdom of Jerusalem was governed by two of its ablest rulers, Baldwin III and Amalric I. In 1153 King Baldwin captured Ascalon—the last major conquest of the Franks. The victory extended their coastline southward. But its possession was offset in the next year by the occupation of Damascus by Nureddin, one more stage in the encirclement of the crusader states by a single Muslim power.

In 1160–61 the possibility that the Fāṭimid caliphate in Egypt, shaken by palace intrigues and assassinations, might collapse and fall under the influence of Muslim Syria caused anxiety in Jerusalem. Thus, in 1164, when Nureddin sent his lieutenant Shīrkūh to Egypt accompanied by his own nephew, Saladin, King Amalric decided to intervene. After some manoeuvring, both armies withdrew, as they were to do again three years later.

Meanwhile, Amalric, realizing the necessity of Byzantine cooperation, had sent Archbishop William of Tyre as envoy to Constantinople. But in 1168, before the news of the agreement that William of Tyre had arranged reached Jerusalem, the King set out for Egypt. The reasons are not clear, and there had been considerable division among the barons. At any rate, the venture failed, and Shīrkūh entered Cairo. On his death (May 23, 1169), Saladin, then Nureddin's deputy, was left to overcome the remaining opposition and master Egypt.

The accession of Saladin

When the Byzantine fleet and the army finally arrived in 1169, there was some delay, and both armies were forced by inadequate provisions and seasonal rains to retreat once again, each side blaming the other. In 1171, Saladin obeyed Nureddin's order to have the prayers in the mosques mention the Caliph of Baghdad instead of the Caliph of Cairo, who was then in his final illness. Thus ended the Fāṭimid caliphate and the great division in Levantine Islām from which the Latins had profited.

Ominous developments followed the deaths of both Amalric and Nureddin in 1174. In 1176 the Seljuqs of Iconium defeated the armies of Manuel Comnenus at Myriocephalon. It was a shattering blow reminiscent of Manzikert a century earlier. When Manuel died in 1180, all hope of effective Byzantine–Latin cooperation vanished. Three years later Saladin occupied Aleppo, virtually completing the encirclement of the Latin states. In 1185 he agreed to a truce and left for Egypt.

In Jerusalem, Amalric was succeeded by his son Baldwin IV, a boy of 13 suffering from leprosy. Despite the young king's extraordinary fortitude, his precarious health necessitated occasional regencies and created a problem of succession until his sister Sibyl bore a son, the future Baldwin V, to William of Montferrat. Her marriage in 1180 to Guy of Lusignan, a newcomer to the East and brother of Amalric, the constable, accentuated existing rivalries among the barons. A kind of "court party" centring around the queen mother, Agnes of Courtenay, her daughter Sibyl, and Agnes' brother, Joscelin III, and now including the Lusignans was often opposed by another group composed mostly of the so-called native barons—old families, notably the Ibelins, Reginald of Sidon, and Raymond III of Tripoli, who through his wife was also lord of Tiberias. In addition to these internal problems, the kingdom was more isolated than ever. Urgent appeals to the West and the efforts of Pope Alexander III had brought little response.

Baldwin IV died in March 1185, leaving, according to previous agreement, Raymond of Tripoli as regent for the child king Baldwin V. But when Baldwin V died in 1186 the court party outmanoeuvred the other barons and, disregarding succession arrangements that had been formally drawn up, hastily crowned Sibyl. She in turn crowned her husband, Guy of Lusignan.

In the midst of near civil war, Reginald of Châtillon, lord of Kerak and Montréal, broke the truce by attacking

a caravan. Saladin replied by proclaiming *jihād* against the Latin kingdom. In 1187 Saladin left Egypt, crossed the Jordan south of the Sea of Galilee, and took up a position near the riverbank. Near Saffuriyah (modern Zippori) the crusaders mobilized an army of perhaps 20,000 men and including some 1,200 heavily armed cavalry, probably the equal of Saladin's. In a spot well chosen and adequately supplied with water and provisions, they awaited Saladin's first move.

On July 2 Saladin blocked the main road to Tiberias and sent a small force to attack the town, hoping, since Count Raymond's wife was there, to lure the crusaders into the open. Nevertheless, it was Raymond who at first persuaded the King not to fall into the trap. But, late that night, others, accusing the Count of treason, prevailed upon the King to change his mind. It was a fateful decision. For, after an exhausting day's march on July 3, a terrible night spent without water, surrounded and constantly harassed, and a long day of fighting near the Horns of Ḥaṭṭin on July 4 with smoke from grass fires set by the enemy pouring into their faces, the foot soldiers broke and fled, destroying the essential coordination with the cavalry. When Saladin's final charge ended the battle, most of the knights had been slain or captured. Only Raymond of Tripoli, Reginald of Sidon, Balian of Ibelin, and a few others escaped.

Defeat of the crusaders

The King's life was spared, but Saladin killed Reginald of Châtillon and ordered the execution of some 200 Templars and Hospitallers. Other captive knights were treated honourably, and most were later ransomed. Less fortunate were the foot soldiers, most of whom were sold into slavery. Virtually the entire military force of the Kingdom of Jerusalem had been destroyed.

Saladin quickly followed up his victory at Ḥaṭṭin. Pausing only to take Tiberias, he moved toward the coast and seized Acre. By September 1187 he and his lieutenants had occupied most of the major strongholds in the kingdom and all of the ports south of Tripoli, except Tyre, and Jubayl and Botron (al-Batrūn) in the County of Tripoli. On October 2, Jerusalem, then defended by only a handful of men under the command of Balian of Ibelin, capitulated to Saladin, who agreed to allow the inhabitants to leave after paying a ransom. Though Saladin offered to release the poor for a specified sum, several thousand apparently were not redeemed and probably were sold into slavery. In Jerusalem, as in most of the cities captured, those who elected to remain were Syrian or Greek Christians. Somewhat later Saladin permitted a number of Jews to settle in the city.

Meanwhile, Saladin continued his conquests in the north, and by 1189 all the kingdom was in his hands except Belvoir (modern Kokhov ha-Yarden) and Tyre. The County of Tripoli and the Principality of Antioch were each reduced to the capital city and a few outposts. The larger part of the 100-year-old Latin establishment in the Levant had been lost.

**The institutions of the First Kingdom.** The four principalities established by the crusaders—three after the loss of Edessa in 1144—were loosely connected, and such limited suzerainty as the king of Jerusalem exercised over the others became largely nominal after the midcentury. Each of the states was organized into a pattern of feudal lordships by the ruling minority. The institutions of the Kingdom of Jerusalem are best known, partly because its history figures more prominently in both Arab and Christian chronicles but especially because its documents were better preserved. In the 13th century there was compiled a collection of laws, the famous *Assises de Jérusalem.* Though this collection reflects a later situation, certain sections and many individual enactments can be traced back to the 12th century, the period known as the First Kingdom.

In the first half of the 12th century the kingdom presented the appearance of a typical feudal monarchy on the European model, with lordships owing military service and subject to fiscal exactions. There were, however, important differences, not only in the large subject population of diverse ethnic origins but also with respect to the governing minority. No great families with extensive domains emerged in the early years. Further, the typical noble's residence was not, as in Europe, the rural castle or manor house. Castles there were, but they were garrisoned by knights and, increasingly as the century advanced, by the religio-military orders. But most Jerusalem barons lived in the fortified towns. The kings, moreover, possessed a considerable domain and retained extensive judicial rights. As a consequence, the monarchy was a relatively strong institution in early Jerusalem.

Toward the middle of the century this situation began to change. Partly as a consequence of increased immigration from the west, the baronial class grew, and there was formed a relatively small group of magnates with large domains. As individuals, they were less disposed to brook royal interference, and as a class and in the court of barons (Haute Cour, or High Court) they were capable of presenting a formidable challenge to royal authority. The last of the kings of Jerusalem to exercise effective power was Amalric I in the 12th century. In the final years of the First Kingdom, baronial influence was increasingly evident and dissension among the barons, as a consequence, more serious.

Another serious obstacle to the king's jurisdiction and one not found, at least in the same form, in the West was the extensive authority of the two religio-military orders. The older of the two, the Knights of St. John of Jerusalem, or Hospitallers, was founded in the 11th century by the merchants of Amalfi to provide hospital care for pilgrims. The order never abandoned its original purpose, and, in fact, as its superb collection of documents reveals, the order's philanthropic activities expanded. But during the 12th century, in response to the military needs of the kingdom, the Hospitallers followed the example of the Templars and became, in addition, an order of knights.

The military orders

The Poor Knights of Christ and of the Temple of Solomon, so called because of their headquarters in the former temple of Solomon, originated as a monastic-military organization to protect pilgrims on the way to Jerusalem. Its rule, composed by St. Bernard of Clairvaux, was officially sanctioned by the Council of Troyes (1128). Although Templars and Hospitallers took monastic vows, their principal function was soldiering.

The orders grew rapidly, and, as they acquired castles at strategic points in the kingdom and in the northern states and maintained permanent garrisons, they supplemented the otherwise not entirely adequate forces of the barons and king. Moreover, since they were soon established in Europe as well, they became international in character. Virtually independent, sanctioned and constantly supported by the papacy, and exempt from local ecclesiastical jurisdiction, they aroused the jealousy of the clergy and constituted a serious challenge to royal authority.

The crusaders introduced into the conquered lands a Latin ecclesiastical organization and hierarchy. The Greek patriarch of Antioch was, as has been mentioned, removed, and all the subsequent incumbents were Latin, except for one brief period before 1170, when imperial pressure succeeded in installing a Greek. In Jerusalem the Orthodox patriarch had left before the conquest and died soon after. All his successors were Latin.

Under the Latin jurisdiction was the entire Latin population and those natives (Greek in Antioch and Greek or native Syrian [Melchite] in Jerusalem) who had remained Orthodox. Outside was a larger number of Monophysites (Jacobite or Armenian) and some few Nestorians, all adherents of doctrines that had deviated from the decisions of 5th-century ecumenical councils. A number of Maronites of the Lebanon region accepted the Latin obedience late in the 12th century. After some initial confusion, the native hierarchies were able to resume their functions.

As in the West, the church had its own courts and possessed large properties. But each ecclesiastical domain was required to furnish soldiers, and the charitable foundations were considerable. The hierarchy of the Latin states was an integral part of the church of the West. Papal legates regularly visited the East, and bishops from the crusader states attended the Third Lateran Council in 1179. Western monastic orders also appeared in the crusader's states.

In addition to the nobles and their families who had

The bourgeois

settled in the kingdom, a substantially larger number of persons was classified as bourgeois. A small number had arrived with the First Crusade, but most were later immigrants from Europe, representing nearly every nationality, but predominantly from rural southern France. In the East they became town dwellers, but a few were agriculturalists, proprietors of small estates, rarely themselves tillers of the soil, inhabiting the more modest towns. Some immigrants, it appears, perhaps poor pilgrims who remained, failed to obtain a reasonably settled status and could not afford the relatively small ransom offered by Saladin in 1187.

The townsmen of the First Kingdom did not, like their counterparts in Europe, aspire to political autonomy. There were no communal movements in the 12th century. The bourgeois were, therefore, subject to king or seigneur. Some did military service as sergeants; *i.e.,* mounted auxiliaries or foot soldiers. Bourgeois were recognized as a class in the more than 30 "courts of the bourgeois" according to procedures laid down in the *Assises de la cour des bourgeois,* which, unlike other parts of the *Assises,* reflects the traditions of Roman law in southern France.

The Italians, because they supplied indispensable naval aid and shipping essential to regular contact with Europe, had been able to acquire exceptional privileges in the ports. These privileges usually included a quarter that they maintained as a virtually independent enclave. Its status guaranteed by treaty between the kingdom and the mother city (Venice, Genoa, Pisa, etc.), the Italian enclave more nearly resembled the modern overseas colony than any other establishment of the crusaders.

The European settlers in the crusader states constituted only a small minority of the population. If the early crusaders were ruthless, their successors, except for occasional outbursts during campaigns, acquired not only a remarkable tolerance but also a notable flexibility in dealing with the diverse sectors of the native population. Those Muslim town dwellers who had not fled were captured and put to menial tasks. Some, it is true, appeared in Italian slave marts, but royal and ecclesiastical ordinances at least laid down limits to what slave owners might do. Baptism brought with it immediate freedom.

Not all Muslims were slaves. Most of those who remained were peasants who for centuries had constituted a large part of the rural population and who were permitted to retain their holdings, subject to fiscal impositions not unlike those of the European serf, and usually identical with those originally levied by their former proprietors on all non-Muslims. Muslim nomads, or Bedouins, who from time immemorial had moved with their herds as changing seasons provided pasturage, continued to enjoy pasturage rights within the kingdom guaranteed by the king.

Most mosques were appropriated during the conquest, but some were restored, and no attempt was made to restrict Muslim religious observance. Occasionally a *miḥrāb* (prayer niche) was retained for Muslim worshippers in a church that had formerly been a mosque. The tolerance of the Franks, noted by Arab visitors, often surprised and disturbed newcomers from the West.

Legal practices

Native Christians, predominantly townsmen, were governed according to the *Assises de la cour des bourgeois.* Each national group retained its institutions. The Syrians, for example, maintained a *cour du rais,* the *rais* (*raʾīs*) being a chieftain and often of importance under the Frankish regime. An important element in the kingdom's army, the corps of Turcopoles, lightly armed cavalry units, was also composed largely of native Christians, including, apparently, converts from Islām. The principle of personality of law applied to all. The Jew took oath on the Torah, the Samaritan on the Pentateuch, the Muslim on the Qurʾān, and the Christian on the Gospels.

The Jewish community of Palestine, which had already begun to diminish in the 11th century, was drastically reduced by the First Crusade. But gradually, as the Latin kingdom settled into a routine of government, the situation improved. Indeed, there is reason to believe that the later, more stable regime made possible a not inconsiderable Jewish immigration, not, it seems, as in earlier times, from the neighbouring lands of the Middle East, but from Europe.

Thus, by the 1170s the crusader states of Outremer, as the area of Latin settlement came to be called, had developed well-established governments. With allowance made for regional differences (*e.g.,* Antioch in its early years under the Norman dynasty was somewhat more centralized), the institutions of the northern states resembled those of Jerusalem. The governing class of Franks was no longer made up of conquerors from abroad but of local residents who had learned to adjust themselves to a new environment and were concerned with administration. A few—as, for example, Reginald of Sidon and William of Tyre, archbishop and chancellor—were fluent in Arabic. Many others knew enough of the language to deal with natives. Franks adopted native dress, ate native food, employed native physicians, and married Syrian, Armenian, or convert-Muslim women.

The Franks' cultural identity

But the Franks of Outremer, though they sometimes acquired a love of luxury and comfort, did not lose the will or ability to confront danger; nor did they "go native." In fundamentals, they continued to adhere to the traditions of their French forebears. They were Latin Christians. Documents were drawn up in Latin. The *Assises* were in French. William of Tyre, born in the East but educated in Europe, wrote a celebrated *History of Deeds Done Beyond the Sea* in the Latin style of the 12th century.

Artists and architects were influenced by Byzantine and Arab craftsmen, but Oriental motifs are usually limited to details, such as doorway carvings. A psalter for Queen Melisend in the 12th century, for example, shows certain Byzantine characteristics, and the artist may have lived in Constantinople, but the manuscript is in the current tradition of French art. Castles followed Byzantine models, often built on the old foundations, but Western ideas were also incorporated. New churches were built or additions made to existing structures, as for example the Church of the Holy Sepulchre, in the Romanesque style of the homeland.

All in all, the Franks of the First Kingdom were developing a distinctive culture and achieving a sense of identity. Until baronial dissensions weakened the monarchy in later years—a human failing, not a deep cultural maladjustment—the Latin kingdom showed remarkable vitality and ingenuity. It was one of the more sophisticated governmental achievements of the Middle Ages.

**The Third Crusade.** The news of the fall of Jerusalem reached Europe even before the arrival there of Archbishop Josius of Tyre, whom the crusaders had sent with urgent appeals for aid. Pope Urban III soon died, shocked, it was said, by the sad news; his successor, Gregory VIII, issued a Crusade bull and called for fast and penitence.

Before a new Crusade could be organized, a modest recovery had begun in the East. Scarcely two weeks after Ḥaṭṭin, Conrad of Montferrat, Baldwin V's uncle, had landed at Tyre with a small Italian fleet and a number of followers. He immediately established himself sufficiently to stave off an attack by Saladin, who was then intent on reaching Jerusalem without delay. When toward the end of 1188 Saladin released King Guy as he had promised, Conrad refused to submit to the King.

Guy's siege of Acre

In a daring move to establish his authority, Guy suddenly gathered his few followers and besieged Acre. Saladin was taken completely by surprise. When he finally moved his army toward the city, the crusaders in their camp outside had begun to receive reinforcements from the West, many under the banner of Henry of Champagne. Since Saladin could neither enter the city nor dislodge the crusaders, by winter of 1190–91 neither side had made progress, but the crusaders had suffered losses from disease and famine.

Among the victims of disease was Guy's wife, Sibyl, the source of his own claims to the throne. Many of the older barons who had thus far supported him now turned to Conrad. The marriage of Sibyl's sister, Isabel, to Humphrey of Toron was forthwith annulled, and she was constrained to marry Conrad. But Guy refused to abandon his claim to the throne. Such was the situation in May 1191, when ships arrived off Acre bringing welcome supplies and news of the approach of the armies of the Third Crusade.

The first ruler to respond to the papal appeal was William II of Sicily, who immediately abandoned a con-

flict with Byzantium and equipped a fleet that soon left for the East, though William himself died in November 1189. English, Danish, and Flemish ships also departed. Meanwhile, Gregory VIII had sent a legation to the Holy Roman emperor Frederick Barbarossa, now nearly 70 years old and approaching the end of an eventful career, which had included a protracted controversy with the papacy. But Frederick had made peace with the church, had participated in the Second Crusade, and for some time had been genuinely desirous of taking the cross again.

Frederick set out in May of 1189 with the largest Crusade army so far assembled and crossed Hungary into Byzantine territory. His troops caused disturbances in the Balkans, and there were subsequent misunderstandings with Emperor Isaac Angelus. Frederick was eventually persuaded to avoid Constantinople and cross the Dardanelles. In May 1190 he reached Iconium after defeating a Seljuq army. Thence his army crossed into Armenian territory. On June 10 Frederick, who had ridden ahead with his bodyguard, was drowned while attempting to swim a stream. His death broke the morale of the German army, and only a small remnant, under Frederick of Swabia and Leopold of Austria, finally reached Tyre. To Saladin and the Muslims, who had been seriously alarmed by Frederick's approach, Frederick's death seemed an act of God.

Death of Frederick Barbarossa

In Europe, Archbishop Josius had won over Philip II Augustus of France and Henry II of England. Henry died in 1189, leaving the cause to his son and successor, Richard I the Lion-Heart. But pending Anglo-French disputes were not easily settled. The extensive holdings of the English Angevin kings in France and especially Philip's desire to recover Normandy posed problems difficult to lay aside even during a common enterprise. Thus, it was not until July 4, 1190, three years after Ḥaṭṭin, that the two kings met at Vézelay, prepared to move with their armies.

The roles of Richard I and Philip II in the Third Crusade

The two kings who finally led the Third Crusade were very different persons. Richard was in many ways an unstable character. He had opposed his father and was distrustful of his brothers. He could be lavishly generous even to his adversaries but often violent to anyone who stood in his way. His abilities lay not in administration, for which he had no talent, but in war, at which he was a genius. The favourite son of Eleanor of Aquitaine, Richard epitomized the chivalrous crusader and represented the contemporary troubadour view of war with all its aristocratic *courtoisie.* Richard could honour his noble Muslim opponents but be utterly ruthless to lowborn captives.

Philip II Augustus, with his untidy hair and defective vision, was not an attractive person. Unlike Richard, he had been king for 10 years and was a skilled and unscrupulous politician. He had no love for ostentation. Though no warrior himself, he was adept at planning sieges and designing siege engines. But he was a reluctant crusader whose real interests lay in the expansion of his own domains.

According to the original suggestion of King William II, the two kings met at Messina, in Sicily, where they signed an agreement outlining their mutual obligations and rights on the crusade. Philip arrived with the French fleet at Acre on April 20, 1191, and the siege was begun again in earnest.

After a stormy passage, Richard put in at Cyprus, where his sister Joan and his fiancée, Berengaria of Navarre, had been shipwrecked and held by the Byzantine ruler, the rebel prince, Isaac Comnenus. Isaac underestimated Richard's strength and attacked. Not only did Richard defeat and capture him, but he proceeded to conquer Cyprus. He then set sail and arrived on June 8 at Acre, where he gave new vigour to the siege.

A month later, after constant battering at the walls by siege engines and after Saladin's nephew had failed to fight his way into the city, the garrison surrendered in violation of Saladin's orders. Saladin was shocked by the news, but nevertheless he ratified the agreement, which provided for an exchange of prisoners and the restoration of the relic of the True Cross, which he had captured in 1187.

Internal disputes

As the crusaders entered the city, there were disputes over the disposal of areas. Richard offended Leopold of Austria, and Philip, who felt that he had fulfilled his crusader's vow and who was unwell, left for home in August.

Though the English resented Philip's departure, it did leave Richard in control. When the first prisoners were returned by Saladin, Richard found some flaw in the selection of them and refused to free the Saracen captives. Subsequent negotiations failed, and Richard, impatient to move toward Jerusalem, ordered them all killed, along with their wives and children. Thus, the truce arrangements were voided, and Saladin did not return the True Cross.

The English king was never to reach Jerusalem. He won a victory at Arsuf in 1191, reoccupied Jaffa as a base, and took Daron in 1192 but failed to progress further. Messages from home were urging his return. He had been in constant communication with Saladin and his brother al-ʿĀdil, and various peace proposals were made, which included marriage alliances. In fact, there seemed to be warm cordiality and considerable mutual respect. Finally, on September 2, 1192, the two opponents signed a treaty of peace to last five years. The coast from Jaffa north remained in Christian hands, but Ashkelon was to be demolished. Pilgrims were to have free access to the holy places. On October 9 Richard left. He was shipwrecked and finally fell into the hands of Leopold of Austria, who had not forgotten the slight at Acre.

The Third Crusade had failed of its major objective, the capture of Jerusalem. But the possession of Acre and much of the coast permitted the continued existence of a titular kingdom. Before he left, Richard had consented to the request that Guy, who had managed to lose the support of nearly all the barons, be deposed and Conrad immediately be accepted as king. No sooner was this done than Conrad was assassinated. Isabel was persuaded to marry Henry of Champagne, and Guy was given the governorship of Cyprus, where his record was far more successful than his ill-starred career in Jerusalem.

The most lasting achievement of the Third Crusade, however, was Richard's capture of Cyprus. In succeeding decades it gained in importance, not only as an outpost of the coastal possessions but also as a base for future crusades and finally as a kingdom in its own right.

## THE FOURTH CRUSADE AND THE LATIN EMPIRE OF CONSTANTINOPLE

Pope Innocent III, despite manifold problems in the West, was the first pope since Urban II to be both anxious and able to consider the Crusade a major papal concern. In 1198 he broached the subject of a new expedition through legates and encyclical letters. In 1199 a tax was levied on all clerical incomes—later to become a precedent for systematic papal income taxes—and Fulk of Neuilly, a popular orator, was commissioned to preach. At a tournament held by Thibaut III of Champagne, several prominent French nobles took the cross, and others joined later. Among them was Geoffrey of Villehardouin, who was to write one of the principal accounts of the Crusade. Contact was made with Venice to provide transport.

The involvement of Venice proved to be fateful. The republic had acquired considerable trading privileges within the Byzantine Empire, and the growing number of Venetian merchants had long incurred the hostility of the Greeks. In 1171 Manuel Comnenus had ordered the arrest of Venetians, and 11 years later an aroused citizenry massacred a large number of Latins in Constantinople and insulted a papal legate. Further, following the advice of recently returned crusaders, the new Crusade was to be directed against Egypt, now the centre of Muslim power in the Levant but a government with which Venice was closely related commercially. Venetian policy under the aging and blind but ambitious Doge Enrico Dandolo was therefore potentially at variance with that of the Pope and the crusaders. Nevertheless, an agreement was made providing for payments to the Venetians for transportation and an equal division of conquests.

The Venetians' involvement

The crusader army that arrived at Venice in the summer of 1202 was somewhat smaller than had been anticipated, since some of the crusaders were travelling directly from France. Even so, there were not sufficient funds to pay the Venetians. Accordingly, the crusaders accepted the suggestion that in lieu of payment they assist the Vene-

tians in the capture of the Hungarian city of Zara. This was done despite the opposition of many crusaders both to the diversion of the enterprise and to the attack on a Christian city. Innocent was informed of the plan, but his veto was disregarded. Reluctant to jeopardize the Crusade, Innocent gave conditional absolution to the crusaders—not, however, to the Venetians.

The fall of Zara (November 1202) added to the Pope's already considerable misgivings over the transformation of the whole undertaking from a Crusade under papal auspices to one under lay direction. After the death of Thibaut of Champagne, the leadership of the Crusade passed to Boniface of Montferrat, a friend of the Hohenstaufen Philip of Swabia. Both Boniface and Philip had married into the Byzantine imperial family. In fact, Philip's brother-in-law, Alexius, son of the deposed and blinded emperor Isaac Angelus, had appeared in Europe seeking aid and had made contacts with the crusaders.

Innocent was aware of a plan to divert the Crusade to Constantinople in order to gain the throne for Alexius, who promised subsidies to the crusaders. Accordingly, Innocent ordered Boniface of Montferrat to publish immediately his original letter excommunicating the Venetians, which he had refused to do, and forbade any attack on Constantinople. But the papal letter arrived after the fleets had left Zara, and in the summer of 1203 Constantinople fell, Alexius III was deposed, and "Alexius IV" was crowned as coemperor with his father. All Innocent could do was reprimand the leaders and order them to proceed forthwith to the Holy Land. No doubt he hoped that a union of the churches would result and the Crusade thereby promoted. A few crusaders left, but most did not.

The takeover of Constantinople

Following the assassination of the new "emperor" by a resentful Greek population, the Venetians and crusaders themselves took over the city and the government of the empire. It was decided that 12 electors (six Venetians and six crusaders) should choose an emperor who would have one-quarter of the imperial domain. The other three-quarters were to be divided. The clergy of the party not belonging to the emperor elect were to have Hagia Sophia and choose a patriarch. A small amount of property was specifically designated to support the clergy. The rest was to be considered booty and divided. On April 13, 1204, the great city fell and was subjected by the rank and file to pillage and massacre for three days. Many priceless icons, relics, and other objects later turned up in western Europe, a large number in Venice.

However much Innocent III may have hoped that a friendly Constantinople would aid the Crusade and promote the reunion of Eastern and Western churches, he was aghast at the sack of Constantinople and castigated the crusaders and Venetians in no uncertain terms. But the situation was beyond his control, especially after his legate, on his own initiative, had absolved the crusaders from their vow to proceed to the Holy Land.

When order had been restored, the crusaders and the Venetians proceeded to implement their agreement; Baldwin of Flanders was elected emperor and the Venetian Thomas Morosini chosen patriarch. But the lands parcelled out among the leaders did not include all the former Byzantine possessions. The imperial government continued in Nicaea, and an offshoot empire of Trebizond, at the eastern end of the Black Sea, lasted until 1461. There was also established a Byzantine Despotate of Epirus, and the Bulgarians remained hostile. Various Latin-French lordships throughout Greece—in particular, the duchy of Athens and the Principality of the Morea—did provide cultural contacts with western Europe and promoted the study of Greek. There was also a French impact on Greece. A collection of laws, the *Assises de Romanie,* was edited. The *Chronicle of the Morea* appeared in both French and Greek (and later Aragonese) versions. Impressive remains of crusader castles and Gothic churches can still be seen. Nevertheless, the Latin Empire always rested on shaky foundations, and in 1261 a sadly diminished domain was reconquered from the Latins by Michael Paleologus with the aid of the Genoese, the traditional rivals of Venice.

The survival of imperial authority

It is difficult to see in the Fourth Crusade anything but unmitigated disaster to the cause of either Christianity or the Crusade. Not only had a promising Crusade been diverted, but the new Latin principalities proved to be more attractive to Westerners who might otherwise have gone to the Levant and even to a number of former residents of the states of Syria. The rift between the Eastern and Western churches widened, and Greek popular resistance to any schemes of reunion with the empire intensified. The Byzantine Empire, for centuries a bulwark against invasion from the East, was damaged beyond repair.

## The decline of the crusading movement

### THE LATER CRUSADES AND THE DECLINE OF THE LATIN ENCLAVES

**The Latin East after the Third Crusade.** Saladin died on March 3, 1193, not long after the departure of the Third Crusade. One of the greatest of the Muslim leaders and a man devoutly religious and deeply committed to *jihād* against the infidel, he was, yet, respected by his opponents. His death led once again to divisions in the Muslim world, and his Ayyūbid successors were willing to continue a state of truce with the crusaders, which lasted into the early years of the 13th century. The truce was politically and economically advantageous to both sides, and the Italians were quick to make profitable trade connections in Egypt. The Franks were able to adjust themselves to a new situation and to organize what in effect was a new titular kingdom of Jerusalem centring at Acre, generally known as the Second Kingdom.

In 1194 Amalric of Lusignan succeeded his brother Guy as ruler of Cyprus, where he later accepted investiture as king from the chancellor of the emperor Henry VI. In 1197, following the death of Henry of Champagne, Amalric succeeded to the throne of Jerusalem-Acre, and in 1198 he married the thrice-widowed Isabel. He chose, however, to govern his two domains separately, and in Acre he proved to be an excellent administrator. The *Livre au roi* ("Book of the King"), an important section of the *Assises de Jérusalem,* dates from his reign. He also dealt wisely with Saladin's brother, al-ʿĀdil of Egypt. On Amalric's death in 1205 the kingdoms of Cyprus and Jerusalem-Acre were divided, and in 1210 the latter was given to John of Brienne, a French knight nominated by Philip Augustus, who came east and married Conrad's daughter, Mary.

There were also readjustments in the two now-reduced northern states. When Raymond III of Tripoli died (1187), his county passed to a son of Bohemond III of Antioch, thus uniting the two principalities. In general, Antioch-Tripoli continued to pursue the relatively independent course laid down by Bohemond III.

Armenia was more closely involved in Latin politics, partly as a result of marriage alliances with the house of Antioch-Tripoli. King Leo II of Armenia joined the crusaders at Cyprus and Acre. Desirous of a royal crown, he had approached both Pope and Emperor, and in 1198, with papal approval, royal insignia were bestowed by Archbishop Conrad of Mainz, in the name of Henry VI. At the same time, the Armenian Church officially accepted a union with Rome, which, however, was never popular with the lower clergy and general populace.

**The Fifth Crusade.** In Europe after the Fourth Crusade something of the original crusading fervour was still found in certain areas of society. Even children became the victims of mass hysteria; in 1212, in the so-called Children's Crusade, thousands of youngsters from France and Germany set out to free the Holy Land, only to be lost, shipwrecked, or sold into slavery. But there was considerable disillusionment among the nobility, especially when the same religious indulgence was promised in a "Crusade" against heretics in southern France and later against secular opponents of the popes. Innocent III, nevertheless, despite his preoccupation with the kings of England and France, a civil war in Germany, heresy, the advance of Islām in Spain, and his strenuous efforts to promote widespread ecclesiastical reform, renewed his efforts to organize another expedition. Preachers were designated; even troubadours contributed to the propaganda. The final canon promulgated by the Fourth Lateran Council of 1215 was an elaborate Crusade plan repeating earlier pro-

The Children's Crusade

hibitions on the transport of military supplies to Muslims. Levies on clerical incomes were again authorized.

The young emperor Frederick II had promised Innocent to go, but Honorius III permitted him repeated postponements to settle the affairs of Germany. Thus, the first contingents of the Fifth Crusade left without him. Nor did they accomplish anything significant until the arrival of a Frisian fleet with more German crusaders enabled a force to set out for Egypt in May 1218 under the leadership of John of Brienne. An attack on Egypt had been in the planning since the time of the Third Crusade, not as the objective of permanent conquest but as a bargaining point for the recovery of Jerusalem. By August the crusaders had captured a strategic tower at Damietta (Dumyāṭ).

The Siege of Damietta

In September the expedition organized under papal auspices and consisting mainly of French crusaders arrived under Cardinal Pelagius as legate. Since Pelagius regarded the crusaders as being under the jurisdiction of the church, he declined to accept the leadership of John of Brienne. Moreover, he was an imperious person who did not hesitate to interfere in military decisions.

By February 1219 the Muslims were seriously alarmed and offered peace terms that included the cession of Jerusalem. King John and many of the crusaders were eager to accept. But Pelagius, supported by the military orders and the Italians, refused. Damietta was finally taken on November 5, 1219. For over a year no progress was made, though Pelagius remained optimistic, still expecting the arrival of Frederick II and convinced, by rumour, of the imminent approach of a legendary oriental Christian "King David." In July 1221 he ordered an advance toward Cairo, but Nile floods forced him to retreat and to agree to a truce of eight years and an exchange of prisoners, terms far less favourable than those he had previously rejected.

The Fifth Crusade, the last in which the papacy took an active part, was an impressive effort against a divided Muslim opposition. Coming close to success, it failed largely because of divided leadership and the frequently unwise decisions of Pelagius. It might perhaps have succeeded if Frederick II had set out as promised, and it is significant that disillusioned critics blamed Emperor and Pope as well as Pelagius. All in all, it was a dreary episode, relieved only by the presence of Francis of Assisi, to whom Pelagius reluctantly gave permission to cross the lines, where he was courteously received by Sultan al-Kāmil. Francis' visit to the East was a preliminary step in the establishment of a Franciscan province in the Holy Land, a step soon imitated by the Dominicans.

The arrival of missionary orders

**The Sixth Crusade.** The failure of the Fifth Crusade placed a heavy responsibility on Frederick II. His motives as a crusader are difficult to assess. Always a controversial figure, he was to some the archenemy of the papacy, to others the greatest of emperors. His intellectual interests included Islām, and his attitude might seem to be more akin to that of the Eastern barons than the typical crusader from the West. Through his marriage with John of Brienne's daughter Isabella (Yolande), he had a claim first to the kingship and then, on her death in 1228, to the regency of Jerusalem (Acre) for his son Conrad. As emperor he could claim a suzerainty over Cyprus by virtue of the cession by his predecessor Henry VI. Frederick had finally agreed to terms that virtually placed his expedition under papal jurisdiction. Yet his entire Eastern policy was inextricably connected with his European concerns: Sicily, Italy and the papacy, and Germany. Cyprus-Jerusalem became, as a consequence, part of a greater imperial design.

Most of the fleet of the Sixth Crusade had left Italy in the late summer of 1227, but Frederick was delayed by illness. During the delay he received envoys from Sultan al-Kāmil of Egypt, who, threatened by the ambitious designs of his Ayyūbid brothers, was disposed to negotiate. Meanwhile, Pope Gregory IX, less patient than his predecessor, had rejected the Emperor's plea of illness and excommunicated him. Thus, when Frederick departed in the summer of 1228 with the remainder of his forces, he was in the equivocal position of a crusader under the ban of the church. He arrived in Cyprus on July 21.

In Cyprus, John of Ibelin, the leading member of the influential Ibelin family, had been named regent for the young Henry I. Along with most of the other barons, he was willing to recognize the Emperor's rights as suzerain in Cyprus. But in Acre, since news of Isabel's death had arrived, the Emperor could claim only a regency for his infant son. John obeyed the Emperor's summons to meet him in Cyprus but despite intimidation refused to surrender his lordship of Beirut and insisted that his case be brought before the High Court of barons. The matter was set aside, and Frederick left for Acre.

At Acre, Frederick met more opposition. News of his excommunication had arrived, and many refused to support him. Dependent, therefore, on the Teutonic Knights, an organization formed by Germans who remained in the east after an expedition in 1197 and now under the direction of Hermann of Salza, and his own small contingent of German crusaders, he was forced to attempt what he could by diplomacy. Negotiations, accordingly, were reopened with al-Kāmil of Egypt.

The treaty of 1229

The resulting treaty of 1229 is unique in the history of the Crusades. By diplomacy alone and without major military confrontation, Jerusalem, Bethlehem, and a corridor running to the sea were ceded to the Kingdom of Jerusalem. Exception was made for the Temple area, the Dome of the Rock, and the Aqṣā Mosque, which the Muslims retained. The peace was to last for 10 years.

The benefits of the treaty of 1229 were more apparent than real. The areas ceded were not easily defensible, and Jerusalem soon became a prey to disorder. Furthermore, the treaty was denounced by the devout of both faiths. When Frederick, still under excommunication, entered the city, the Patriarch placed it under interdict. No priest was present, and Frederick placed a crown on his own head while one of the Teutonic Knights read the ceremony. Leaving agents in charge, he hastily returned to Europe and at San Germano made his peace with the Pope (July 23, 1230). Thereafter, his legal position was secure, and the Pope ordered the Patriarch to lift the interdict.

What followed in Jerusalem and Cyprus, however, was not orderly government by the Emperor's agents but civil war. For Frederick's imperial concept of government was totally opposed to the now well-established pre-eminence of the Jerusalem baronage. The barons of both Jerusalem and Cyprus, in alliance with the Genoese and a commune formed at Acre that elected John of Ibelin as mayor, resisted the imperial deputies, who had the support of the Pisans, the Teutonic Knights, Bohemond of Antioch, and a few nobles. The clergy, the other military orders, and the Venetians stood aloof.

The barons were able to win out in Cyprus, and in 1233 Henry I was recognized as king. Even after John of Ibelin, the "Old Lord of Beirut," died in 1236, the resistance continued. In 1243 a parliament at Acre refused homage to Frederick's son Conrad, unless he appeared in person, and named Alice, queen dowager of Cyprus, as regent.

Triumph of baronial rule

Thus it was that baronial rule triumphed over imperial administration in the Levant. But the victory of the barons brought to the kingdom not strength but continued division, the more serious as new forces were appearing in the Muslim world. The Khwārezmian Turks, pushed south and west by the Mongols, had upset the power balance and gained the support of Egypt. After the 10 years' peace had expired in 1239, a poorly organized Crusade under Thibaut IV of Champagne and Richard of Cornwall accomplished little. In 1244 an alliance of Jerusalem and Damascus failed to prevent the capture and sack of Jerusalem by Khwārezmians with Egyptian aid. All the diplomatic gains of the preceding years were lost.

**The Crusade of Louis IX of France.** In June 1245, a year following the final loss of Jerusalem, Pope Innocent IV opened a great ecclesiastical council at Lyons. Although urgent appeals for help had come from the East, it is unlikely that the Crusade was uppermost in the Pope's mind for a combination of crises confronted the church: massive complaints of clerical abuses, increasing troubles with the emperor Frederick II in Italy, and the advance of the Mongols into eastern Europe. Nevertheless, when King Louis IX of France announced his intention to lead a new Crusade, the Pope gave it his support and authorized the customary levy on clerical incomes.

Character of Louis IX

As a crusader, the saintly Louis was the antithesis of his predecessor, Frederick. Possessed of a rare combination of religious devotion with firmness as a ruler and bravery as a warrior, he seemed the very ideal of the crusader. He was beloved by his subjects and respected abroad. He ardently believed the Crusade to be God's work, and he was far from sympathetic with the Pope's using Crusade propaganda against the Emperor.

It was three years before Louis was ready to embark. Peace had to be arranged with England, transport had to be provided by Genoa and Marseilles, and funds had to be raised from the King's domain and from the towns. When the King embarked in August 1248 he was accompanied by his queen; his brothers Robert of Artois and Charles of Anjou; many distinguished French nobles, including John of Joinville, author of the *Vie de St. Louis;* and a small contingent of English. His army was a formidable one, numbering perhaps 15,000. France was left in the experienced hands of the queen mother, Blanche of Castile.

The Crusade arrived at Cyprus in September, and it was again decided to attack Egypt. Since a winter campaign was not feasible and Louis rejected the suggestion that he attempt negotiations—though he did exchange envoys with the Mongols—it was not until May 1249 that an expedition of some 120 large and many smaller vessels got under way. Fortune favoured them at first, and Damietta was again in Christian hands by June. Shortly afterward, the army was strengthened by the arrival of Louis's third brother, Alphonse of Poitiers. Sultan aṣ-Ṣāliḥ-Ayyūb's death was followed by confusion in Cairo, which, after some argument, had become the crusaders' objective. In February 1250 Robert of Artois led a surprise attack on the Egyptian camp two miles (three kilometres) from al-Manṣūrah and, rejecting the advice of more experienced campaigners, impetuously moved ahead, only to be trapped within the city. Many knights lost their lives. Louis soon arrived with the main army and won another victory, albeit a costly one, near al-Manṣūrah. It was the last crusader success.

Meanwhile, Tūrān-Shāh, the Sultan's son, arrived and temporarily dominated dissident factions in Cairo. Frankish supply ships from Damietta were intercepted, and before long the crusaders were suffering from famine and disease. Louis, reluctant to abandon a work to which he had dedicated his very kingdom, perhaps delayed too long before ordering a retreat. But he refused the pleas of others to protect himself by moving out ahead. Instead, he led his soldiers and along with many of them was captured as the Muslim forces closed in.

The capture of Louis

The King and nobles were held for ransom, but many non-noble captives were killed. The Queen, who had just given birth to a son sorrowfully named John Tristan, managed with great courage to secure sufficient food and to persuade the Genoese and Pisans not to evacuate Damietta until it could be ceded formally by treaty and the King's ransom arranged. On May 6, 1250, the King was released and Damietta surrendered.

Despite the pleadings of his advisers, Louis did not return home immediately. He felt bound in conscience to negotiate the release of as many prisoners as possible, and he was able before he left in April 1254, four years later, to improve the defenses of the kingdom by strengthening a number of fortifications. Thus, he atoned in some small measure for the failure of the Crusade.

During these same years the southern wing of the Mongols under Hülegü overran Mesopotamia and in 1258 took Baghdad, thus ending the venerable ʿAbbāsid caliphate. But two years later (1260) the Mamlūks of Egypt, a new dynasty that had arisen from the leaders of former slave bodyguards of the sultan, defeated the Mongols at ʿAin Jālūd in Syria and halted their southward advance. The Muslim states of Syria were caught in the middle, and the Latin states were in grave danger. King Hethum (Hayton) of Armenia threw in his lot with the Mongols, and his son-in-law Bohemond VI of Antioch-Tripoli followed suit. But the barons at Acre were still more disposed to deal with the Muslims, whom they knew, than with the terrifying and unknown Mongols.

In 1260, after murdering his predecessor, Baybars became sultan of Egypt. Though this famous Mamlūk sultan did not live to see the final fall of the Latin states, before his death in 1277 he had reduced them to a few coastal outposts. Baybars was ruthless and totally devoid of the generous chivalry that the crusaders had admired in Saladin. Most of his conquests were followed by massacre of the inhabitants, often including the native Christians, especially when they had been in league with the Mongols. In 1265 he took Caesarea, Haifa, and Arsuf. The following year he conquered Galilee and devastated Cilician Armenia. In 1268 Antioch was taken and all the inhabitants slaughtered. The great Hospitaller fortress of Krak des Chevaliers fell three years later.

The accession of Baybars as Mamlūk sultan

These disasters again brought pleas for aid from the West. King Louis, who felt a heavy responsibility for the collapse of his first effort, began to make new preparations in 1267. But the French king's second venture, the Eighth Crusade, never reached the East. For reasons that are not entirely clear but probably owed something to the influence of Louis's brother Charles of Anjou, the expedition went instead to Tunis. Charles had recently been named by the papacy as the successor to the Hohenstaufens in Sicily. In 1268 he defeated Conradin, the last of the Hohenstaufen line, and was soon involved in grandiose Mediterranean projects, which ultimately included even Byzantium.

There was little support in Europe for Louis's Crusade, which embarked from southern France in July 1270. Moreover, soon after the French landed in North Africa, disease struck the troops and claimed the lives of both Louis and his son John Tristan. Charles arrived with the Sicilian fleet in time to bargain for an indemnity to evacuate the remnants of the army. Thus, the Crusade ended in tragedy and brought no help to the East. Moreover, except for the expedition of Prince Edward of England (1271–72), who had arrived in North Africa too late to be of assistance, Louis's Crusade was the last.

**The final loss of the crusader states.** Although Europe was aware of the gravity of the Eastern situation, it was both unwilling and unable to give substantial aid, as Pope Gregory X discovered following his pleas at the Council of Lyons in 1274. The sense of common endeavour that had produced the First Crusade was no longer to be found. Not only had political tensions increased, but the reforms of the Fourth Lateran Council of 1215 had not produced the results hoped for. Criticism of ecclesiastical policies was more outspoken.

Crusading had also become more expensive. The time had passed when a Crusade army was made up of knights serving under a lord and paying their own way. Economic pressures were causing many nobles to seek royal service. Royal armies, therefore, tended to become more professional, and many knights as well as foot soldiers served for pay. King Louis's crusades had necessitated complicated financing, causing his successors serious financial trouble.

Inevitably, Outremer was affected by the general malaise of the West, and the chronic divisions that were a major cause in Outremer's downfall paralleled those of Europe. From the time of Frederick II, the kingdom had been governed by absentee rulers, in theory the Hohenstaufens represented in the East by agents, followed after 1243 by regents of the Jerusalem dynasty chosen by the High Court of barons. Finally, in 1268, on the death of the last Hohenstaufen, the crown was given to Hugh III of Cyprus. But in 1276 he returned to Cyprus, thoroughly frustrated. Then in 1277 Charles of Anjou, with papal approval, bought the rights of the nearest claimant and sent his representative. But Charles's principal interests at the time lay in Byzantium. Finally, after Charles's death in 1285, the barons once again chose a native ruler, Henry II of Cyprus.

Successive regents had failed to dominate the Jerusalem baronage, ultimately resulting in the disintegration of the entire structure of Outremer into separate parts. Antioch-Tripoli before its fall had been increasingly aloof and through intermarriage closely tied to Armenia. In Acre, the seat of government of the kingdom, there was a commune of barons and bourgeois. Immigration had ceased, and the barons were now reduced in numbers as old families had died out. Some resided in Cyprus, and others

were nominal lords in Palestine of fiefs of lands actually under Muslim control. The military orders, habitually in conflict, were virtually distinct entities with extensive connections in Europe. The bourgeois population had also considerably altered in composition during the 13th century. Many criminals and other undesirables had found their way to Acre. More important, the earlier homogeneity of a French character had given way to an Italian predominance. But the Italians of Outremer were as divided as they were in Italy. The Genoese-Venetian rivalry extended to the Levant and occasionally, as in Acre in 1256, resulted in outright war.

Papal concern for Outremer was not confined to efforts to enlist military aid. Its financial support was continuous. Popes took an active interest in diplomacy and exchanged envoys with Eastern rulers, both Muslim and Mongol. Further, the 13th-century patriarchs of Jerusalem, commonly named by the pope, were also papal legates. But neither absentee king nor pope nor patriarch-legate could bring to the Latin East the unity necessary for its survival.

Thus, to the crusaders, divided among themselves and isolated, the death of Baybars in 1277 brought only temporary respite. In 1280 they again failed to join the Mongols, whom Sultan Qalāʾūn defeated in 1281. The ineffectiveness of the Jerusalem administration was becoming apparent even to Easterners, and the Il-Khan Abagha sent his deputy Rabban Sauma to the kings of Europe and the Pope to seek an alliance. The effort was fruitless. Tripoli fell in 1289, and Acre, the last crusader stronghold on the mainland, was besieged in 1291. After a desperate and heroic defense, the city was taken by the Mamlūks, and the inhabitants who survived the massacre were enslaved. Acre and all the castles along the coast were systematically destroyed.

The fall of Tripoli and Acre

Perhaps because they sensed their greater isolation, the Franks of the 13th century seem not to have developed further the distinctive culture of their predecessors. The remarkable palace of the Ibelins at Beirut, built early in the century, did boast Byzantine mosaics. But, no doubt partly because of King Louis's four-year stay in the kingdom, existing remains of churches and castles indicate a close following of contemporary French Gothic, and manuscripts are more strongly French in style. Literary tastes were also distinctly French. At the coronation festivities for Henry II in 1286, in total disregard—or perhaps in chivalrous defiance—of the ruin surrounding them, the nobles amused themselves by acting out the romances of Lancelot and Tristan.

The greatest cultural achievement of the Second Kingdom was the famous collection of legal treatises, the *Assises de Jérusalem.* Those parts that were compiled in the middle years of the century and, therefore, in the atmosphere of the wars against the agents of Frederick II constitute a veritable charter of baronial rights. In fact, two of the authors were members of the Ibelin family, and a third, Philip of Novara, was a close associate. These sections indicate a shift from the earlier *Livre au roi,* which more nearly reflects the attitudes of the 12th century. Nevertheless, the *Assises* belong to medieval Europe's legal renaissance.

In many respects, therefore, the Franks of the last days of Outremer were what in a modern setting might be called colonials, but they were entirely on their own, for there was no mother country to come to their rescue.

**The later Crusades and the Kingdom of Cyprus.** Europe was dismayed by the disaster of 1291 but not surprised and shocked as it had been in 1187, for the end had been foreseen. Pope Nicholas IV had tried to organize aid before 1291, and he and his successors continued to do so afterward, but without success. France, which had always been the main bulwark of the Crusades, was in serious conflict with England, which eventually led to the Hundred Years' War in 1337. Moreover, although it could scarcely have been understood at the time, western Europe around 1300 was experiencing the first impact of a population decline and what proved to be a prolonged economic depression.

In the East, the military orders could no longer offer a standing nucleus of troops. In 1308 the Hospitallers took Rhodes and established their headquarters there. In 1344, with some assistance, they occupied Smyrna, which they held until 1402. Meanwhile, the Teutonic Knights had moved their operations to the Baltic area. The Templars were less fortunate. In 1308 the French Templars were arrested by Philip IV, and in 1312 the order was suppressed by Pope Clement V.

The fate of the military orders

It is not surprising, therefore, that such response as did follow papal urgings was largely in the form of Crusade theories. For some years after 1291 various projects were elaborated, all designed to avoid previous mistakes and explore new tactics. The Franciscan missionary Ramon Llull (died 1315), for example, in his *Liber de fine,* suggested a campaign of informed preaching as well as military force. Pierre Dubois (*c.* 1305–07) submitted a detailed scheme for a Crusade to be directed by Philip IV of France, and in 1321 Marino Sanudo in his *Secreta fidelium crucis* produced an elaborate plan for an economic blockade of Egypt. But none of these or any other such schemes was put into effect.

King Peter I of Cyprus finally organized an expedition that in 1365 succeeded in a temporary occupation of Alexandria. After a horrible sack and massacre, the unruly crusaders returned to Cyprus with immense booty. Peter planned to return, but no European aid was forthcoming, and after his murder in 1369 a treaty of peace was signed. No further crusades set out with Jerusalem as the objective. What followed were not really crusades in the old sense but campaigns such as the crusades of Nicopolis in 1396 and Varna in 1444, whose purpose was to defend Europe against the Ottoman Turks, a new power in the East.

With the failure of all attempts to regain a foothold on the mainland, Cyprus remained the sole crusader outpost, and after 1291 the island kingdom was faced with a serious refugee problem. It was in Cyprus that many of the institutions established by the Franks survived. For, although Jerusalem and Cyprus normally had separate governments, through intermarriage and the exigencies of diplomacy the histories of the two had become interwoven. Regents of one were often chosen from among relatives in the other. It has been noted that many Jerusalem barons resided in Cyprus. With suitable modifications, the *Assises de Jérusalem* applied on the island. As on the mainland, the French character of the Cypriot Latins is evident in the remains of Gothic structures.

In one respect Cyprus did differ from the mainland. Whereas the First Kingdom had established a reasonable modus vivendi with its native population, such was not the case in the island kingdom. Many Greek landholders had fled, and those who remained apparently suffered a loss of status. All Greeks resisted the Latinizing efforts of the early 13th-century popes and their representatives. Innocent IV was more flexible, but tension persisted until the Turkish conquest in the 16th century.

## THE RESULTS OF THE CRUSADES

The entire structure of European society changed during the 12th and 13th centuries, and there was a time when this change was attributed largely to the Crusades. Historians now, however, tend to view the Crusades as only one, albeit a significant, factor in Europe's development. It is probable, for example, that the departure of unruly elements on the First Crusade aided the King of France and the great feudatories in preserving order and extending their authority. It is also likely that the disappearance of old families and the appearance of new ones can be traced in part to the Crusades, but generalizations must be made with caution. It must, moreover, be remembered that, while some crusaders sold or mortgaged their property, usually to ecclesiastical foundations, others made it over to relatives. The loss of life was without doubt considerable; many, however, did return to their homes.

Economic and social impact on Europe

The sectors acquired by burgeoning Italian cities in the crusader states enabled them to extend their trade with the Muslim world and led to the establishment of trade depots beyond the Crusade frontiers, some of which lasted well beyond 1291. The transportation they provided was significant in the development of shipbuilding techniques. Italian banking facilities became indispensable to popes and kings. Catalans and Provençals also profited, and,

indirectly, so did all of Europe. Moreover, returning crusaders brought new tastes or increased the demand for spices, Oriental textiles, and other exotic fare. But such demands can also be attributed to changing life-styles and commercial growth in Europe itself.

Religious effects

The papacy launched the First Crusade, and crusades remained a constant concern. While initial success undoubtedly enhanced papal prestige, later failures had the opposite effect, and the diversion of Crusade propaganda to war against heretics or the emperor aroused criticism. So also did the levies on clerical incomes, though it must be noted that papal crusade financing played an important role in Europe's economic development.

The establishment of the Franciscan and Dominican friars in the East during the 13th century made possible the promotion of missions within the Crusade area and beyond. Papal bulls granted special facilities to missionary friars, and popes sent letters to Oriental rulers soliciting permission for the friars to carry on their work. Often the friars accompanied or followed Italian merchants, and, since the Mongols were generally tolerant of religious propaganda, missions were established in Iran, Inner Asia, and even China. But, since Islāmic law rigidly prohibited propaganda and punished apostasy with death, conversions from Islām were few. The Dominican William of Tripoli had some success, presumably within the crusaders' area; he and his colleague, Ricoldus of Montecroce, both wrote perceptive treatises on Islāmic faith and law. Other missionaries usually failed, and many suffered martyrdom. In the 14th century the Franciscans were finally permitted to reside in Palestine as caretakers for the holy places, but not as missionaries.

The Crusades, especially the Fourth, so embittered the Greeks that any real reunion of the Eastern and Western churches was out of the question. But certain groups of Eastern Christians came to recognize the authority of the pope and were usually permitted to retain their native liturgies. Though most of the missions that grew out of the Crusades collapsed with the advance of the Ottoman Turks in the Middle East in the mid-14th century, some of the contacts which the Western Church had made with its Eastern brethren remained.

The justification of warfare

The association of missions with Crusades posed a moral theological problem that troubled medieval thinkers. Thirteenth-century theologians held that conversion could not be forced, but most agreed that force could legitimately be used to preserve a situation in which peaceful propaganda was possible, and they continued to support the Crusade. Furthermore, Europe's fear of Muslim power was such that the Crusade idea persisted well into the 17th century, and the conviction that, in certain circumstances, war might be just became more deeply enrooted in the conscience of the West. Along with the now generally accepted use of the word crusade to denote any common endeavour in a worthy cause, this is one of the most enduring results of the movement.

Unlike Sicily and Spain, the Latin East did not, it seems, provide an avenue for the transmission of Arabic science and philosophy to the West. But the Crusades did have a marked impact on the development of Western historical literature. From the beginning there was a proliferation of chronicles, eyewitness accounts, and later more ambitious histories, in verse and prose, in the vernacular as well as in Latin.

It must be emphasized that the Crusades should not be viewed too exclusively in terms of cause and effect. If they were sometimes a factor contributing to changes in the West, so also were they affected by those changes. Rather, the Crusades and the Latin Kingdom should be considered as an integral part of the diversified culture of Europe in the Middle Ages. (M.W.B.)

**BIBLIOGRAPHY.** The standard bibliographical work on the Crusades, including sources, secondary studies, and journal articles, is HANS E. MAYER, *Bibliographie zur Geschichte der Kreuzzüge* (1960). A similar study in English, somewhat briefer but well organized, is AZIZ S. ATIYA, *The Crusade: Historiography and Bibliography* (1962). The best full-scale treatments of the Crusades in English are STEVEN RUNCIMAN, *A History of the Crusades,* 3 vol. (1951–54, reissued 1975); KENNETH M. SETTON (ed.), *A History of the Crusades,* 2nd ed., vol. 1–2 (1969), a cooperative work by a number of historians; and CARL ERDMANN, *The Origin of the Idea of Crusade* (1978; originally published in German, 1935), a classic in the history of the Crusades. Briefer studies include: AZIZ S. ATIYA, *Crusade, Commerce, and Culture* (1962), which places the Crusade movement in the context of contemporary developments, Eastern as well as Western; ERNEST BARKER, *The Crusades* (1923, reprinted 1971); and RICHARD A. NEWHALL, *The Crusades,* rev. ed. (1963). Two significant studies are FRANCESCO COGNASSO, *Storia delle crociate* (1967); and HANS E. MAYER, *The Crusades,* rev. trans. (1972; originally published in German, 1965). See also ROBERT PAYNE, *The Dream and the Tomb: A History of the Crusades* (1984).

On the institutions of the Crusader states, see JOSHUA PRAWER, *Histoire du Royaume latin de Jérusalem,* 2 vol., 2nd ed. (1975, originally published in Hebrew, 1963). Still indispensable are JOHN L. LAMONTE, *Feudal Monarchy in the Latin Kingdom of Jerusalem, 1100–1291* (1932, reprinted 1970); and JEAN RICHARD, *The Latin Kingdom of Jerusalem* (1979; originally published in French, 1953). DANA C. MUNRO, *The Kingdom of the Crusaders,* ed. by A.C. KREY (1935, reissued 1966), contains the Lowell lectures given in 1924 by the dean of American Crusade historians.

JAMES A. BRUNDAGE (ed.), *The Crusades: Motives and Achievements* (1964), presents the points of view of different scholars. T.S.R. BOASE, *Kingdoms and Strongholds of the Crusaders* (1971), is a study of Crusade architecture with illustrations. R.C. SMAIL, *Crusading Warfare, 1097–1193* (1956, reissued 1972), is a study of the campaigns of the 12th century. EMMANUEL SIVAN, *L'Islam et la Croisade, idéologie et propagande dans les réactions musulmanes aux Croisades* (1968), is an important analysis of the Muslim reaction to the Crusades. See also ERIC CHRISTIANSEN, *The Northern Crusades: The Baltic and the Catholic Frontier, 1100–1525* (1981), a thorough treatment.

Useful selections of sources in English translation are: JAMES A. BRUNDAGE, *The Crusades: A Documentary Survey* (1962); FRANCESCO GABRIELI (comp.), *Arab Historians of the Crusades* (1969; originally published in Italian, 1957); RÉGINE PERNOUD, *The Crusades* (1963; originally published in French, 1960).

(M.W.B./Ed.)

# Crustaceans

Crustacea is a subphylum of the animal phylum Arthropoda. Crabs, lobsters, shrimps, and wood lice are the best-known crustaceans, but the group also includes an enormous variety of other forms without popular names. Crustaceans are generally aquatic and differ from other arthropods in having two pairs of appendages (antennules and antennae) in front of the mouth and paired appendages near the mouth that function as jaws. Because there are many exceptions to the basic features, a satisfactory inclusive definition of all the Crustacea is extraordinarily hard to frame.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 313, and the *Index.*

This article is divided into the following sections:


Crustaceans: subphylum Crustacea 840
- General features
  - Size range and diversity of structure
  - Distribution and abundance
  - Importance to humans
- Natural history
  - Reproduction and life cycles
  - Ecology
- Form and function
  - General features
  - External features
  - Internal features
- Evolution and paleontology
- Classification
  - Distinguishing taxonomic features
  - Annotated classification
  - Critical appraisal

Major crustacean groups 846
- Branchiopods (fairy shrimps, tadpole shrimps, brine shrimps, and water fleas) 846
  - General features
  - Natural history
  - Form and function
  - Evolution and paleontology
  - Classification
- Cirripedes (barnacles) 849
  - General features
  - Natural history
  - Form and function
  - Evolution and paleontology
  - Classification
- Malacostracans (lobsters, shrimps, crabs, scuds, and pill bugs) 854
  - General features
  - Natural history
  - Form and function
  - Evolution and paleontology
  - Classification

Bibliography 859


## CRUSTACEANS: SUBPHYLUM CRUSTACEA

### GENERAL FEATURES

**Size range and diversity of structure.** The largest crustaceans belong to the Decapoda, an order that includes the American lobster, which can reach a weight of 20 kilograms (44 pounds), and the giant Japanese spider crab, which has legs that can span up to 3.7 metres (12 feet). At the other end of the scale, some of the water fleas (class Branchiopoda), such as *Alonella,* reach lengths of less than 0.25 millimetre (0.009 inch), and many members of the subclass Copepoda are less than one millimetre in length. The range of structure is reflected in the complex classification of the group (Figure 1). Some of the parasitic forms are so modified and specialized as adults that they can only be recognized as crustaceans by features of their life histories.

Aquatic habitats

**Distribution and abundance.** Crustaceans are found mainly in water. As a group they range from fresh water to seawater and even into inland brines, which may have several times the salt concentration of seawater. Various species have occupied almost every conceivable niche within the aquatic environment. An enormous abundance of free-swimming (planktonic) species occupies the open waters of lakes and oceans. Other species live at the bottom of the sea, where they may crawl over the sediment or burrow into it. Different species are found in rocky, sandy, and muddy areas. Some species are so small that they live in the spaces between sand grains. Others tunnel in the fronds of seaweeds or into man-made wooden structures. Some members of the orders Isopoda and Amphipoda extend down to the greatest depths in the sea and have been found in oceanic trenches at depths of up to 10,000 metres. Crustaceans colonize lakes and rivers throughout the world, even high mountain lakes at altitudes of 5,000 metres. They range widely in latitude as well: in the high Arctic some crustaceans use the short summer to develop quickly through a generation, leaving dormant stages to overwinter.

A number of crabs are amphibious, being capable of leaving the water to scavenge on land. Some, like the ghost crabs (*Ocypode*), can run at great speed across tropical beaches. One of the mangrove crabs, *Aratus,* can climb trees. Some crabs spend so much time away from the water that they are known as land crabs; however, these crustaceans must return to the water when their eggs are ready to hatch. The most terrestrial of the Crustacea are the wood lice (order Isopoda, family Oniscoidea), most of which live in damp places, although a few species can survive in deserts. In addition to these well-adapted groups, occasional representatives of other groups have become at least semiterrestrial. Amphipods, members of the subclasses Copepoda and Ostracoda, and the order Anomopoda have been found among damp leaves on forest floors, particularly in the tropics.

Terrestrial species

**Importance to humans.** The crustaceans of most obvious importance to humans are the larger species, chiefly decapods. Fisheries in many parts of the world capture shrimps, prawns, spiny lobsters, and the king crab (*Paralithodes*) of the northern Pacific and its southern counterpart, the centolla, found off the coast of Chile. Many species of true crabs—such as the blue crab, Dungeness crab, and the stone crab, all in North America, and the edible crab of Europe—are valuable sources of food. The most highly prized decapod is probably the true lobster (*Homarus* species), although overfishing since the early 20th century has greatly diminished the catches of both the North American and the European species. Freshwater crustaceans include crayfish and some river prawns and river crabs. Many species have only local market value. It is probable that no crustaceans are poisonous unless they have been feeding on the leaves or fruits of poisonous plants.

Edible crustaceans

The large acorn shell (*Balanus psittacus*), a barnacle (order Cirripedia) measuring up to 27 centimetres (11 inches) in length, is regarded as a delicacy in South America, and a stalked barnacle (*Mitella pollicipes*) is eaten in parts of France and Spain. In Japan, barnacles are allowed to

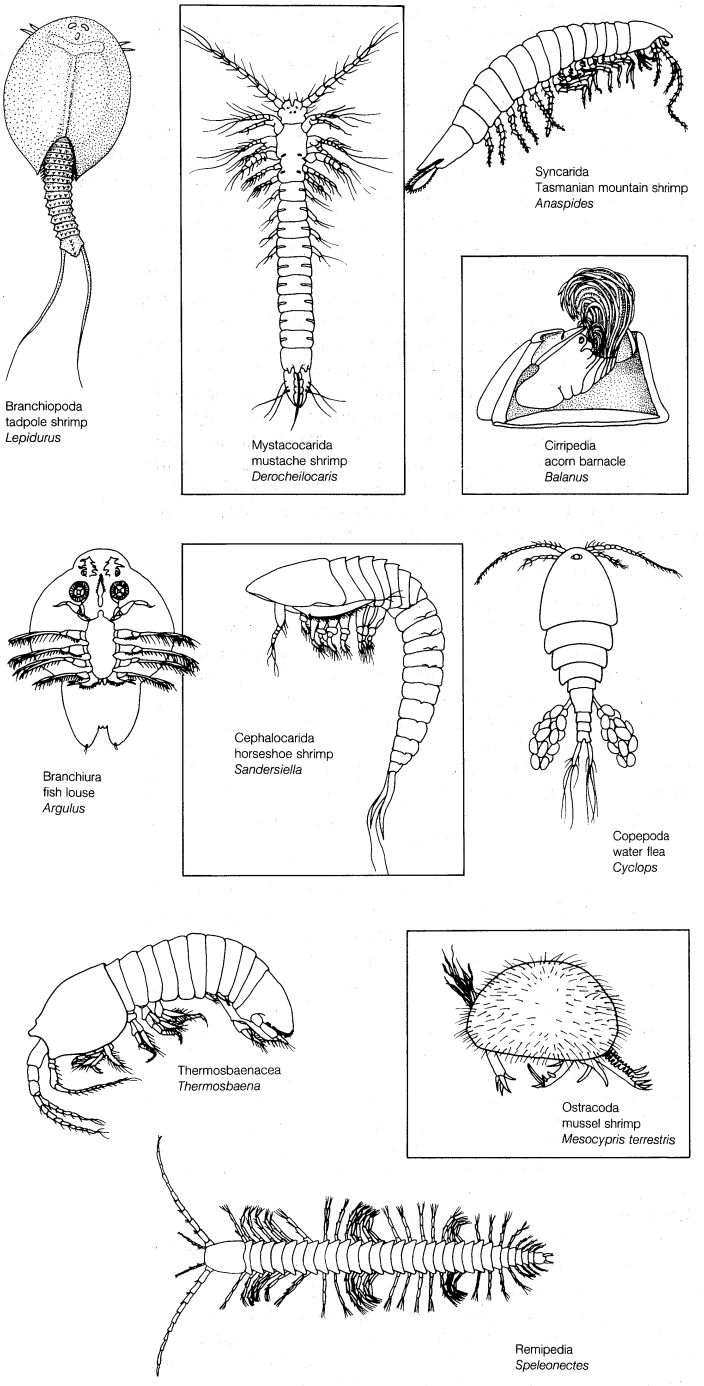



Figure 1: Body plans of representative crustaceans (branchiopod, mystacocarid, syncarid, cirripede, branchiuran, cephalocarid, copepod, thermosbaenid, ostracod, and remipedia).

After (*Anaspides*) Siewing and (*Thermosbaena*) Monod in A. Kaestner, *Invertebrate Zoology* (1970), John Wiley & Sons, Inc.; from (*Mesocypris terrestris*) *Memoires of the Institut Francais d'Afrique Noire* (1966); (*Speleonectes*) J. Yager in L. Botosaneanu (ed.), *Stygofauna Mundi* (1986), E.J. Brill, Leiden; (others except *Lepidurus*) *Invertebrate Zoology* by Paul A. Meglitsch, copyright © 1967 by Oxford University Press, Inc., reprinted by permission

settle and grow on bamboo stakes, later to be scraped off and crushed for use as fertilizer. Planktonic (*i.e.,* drifting) copepods, such as *Calanus* and members of the order Euphausiacea (euphausiids), or krill, may be present in such great numbers that they discolour large areas of the open sea, thus indicating to fishermen where shoals of herring and mackerel are likely to be found. The water flea (*Daphnia magna*) and the brine shrimp (*Artemia salina*) are used as fish food in aquariums and fish ponds, and the larvae of the latter are widely used as food for the larvae of larger crustaceans reared in captivity. Ostracods, of which numerous fossil and subfossil species are known, are of importance to geologists and oil prospectors.

Much damage may be done to rice paddies by burrowing crabs of various species and by the mud-eating, shrimplike *Thalassina* of Malaya. By undermining paddy embankments, they allow water to drain away, thus exposing the roots of the plants to the sun; if near the coast, salt water may thus be allowed to seep into the paddies. Tadpole shrimps (*Triops*) are often numerous in rice fields, where they stir up the fine silt in search of food, killing many of the plants. Land crabs and crayfish may damage tomato and cotton crops.

Harmful and destructive crustaceans

## NATURAL HISTORY

**Reproduction and life cycles.** The sexes are normally, but not always, separate in crustaceans. Most barnacles have both male and female reproductive organs in one individual (hermaphroditism), and in some groups the males, when present, are much smaller than the hermaphrodites. These males attach themselves to the interior of the mantle cavity of the larger individuals and fertilize their eggs. Some of the members of the order Notostraca (tadpole shrimps) are also hermaphrodites; their ovaries contain scattered sperm-producing lobes among the developing eggs. A change of sex during the life of an individual is a regular feature in some shrimps. In *Pandalus montagui,* of the order Decapoda, some individuals begin life as males but change into functional females after about 13 months. Isopods of the genus *Rhyscotoides* show a similar change from male to female as they grow older.

Sexual dimorphism

Characteristic differences in structure or behaviour between the sexes are widespread in the Crustacea and can be extreme; the males of some groups may be so small that they are difficult to find on the much larger female. This is especially true in some of the parasitic copepods. In *Gonophysema gullmarensis* the male is found in a small pouch in the female genital tract. In many of the more advanced decapods, such as crabs and lobsters, however, the males are larger than the females and may have much larger pincers. Another example of sexual dimorphism is the possession by the male of clasping organs, which are used to hold the female during mating. Almost any appendage can be found modified for this purpose. Male appendages also can be modified to aid directly in the transfer of sperm to the female. Frequently the sperm are enclosed in a case, or spermatophore. The first and second abdominal appendages of male decapods are used to transfer spermatophores, as are the highly modified fifth legs of male copepods of the order Calanoida. These copepods can accurately place spermatophores near the openings of the female ducts. The contents of the spermatophores are extruded by a swelling of special sperm, which force out the sperm that actually fertilize the eggs.

Normal sexual reproduction involves the fusion of a sperm with an egg, but some Crustacea are parthenogenetic; that is, they produce eggs that develop without being fertilized by a sperm. Many branchiopods can do this, as can some ostracods and even more advanced crustaceans such as isopods.

Some crustaceans lay their eggs freely in the water—for example, certain copepods, such as *Calanus,* and some members of the malacostracan orders Bathynellacea, Anaspidacea, and Euphausiacea. Some euphausiids and *Nebalia* (of the malacostracan order Leptostraca) carry their eggs between the thoracic limbs. Most decapods carry their eggs attached to the abdominal appendages; special egg-containing setae secrete a cement that flows over the eggs and binds them to the setae. Most of the superorder Peracarida, some isopods, such as *Sphaeroma,* many branchiopods, the Notostraca, and the order Anostraca have a brood pouch on or behind the limbs that is often formed by the carapace. Those free-living copepods that do not cast their eggs freely into the water carry them in one or two thin-walled sacs suspended from the front of the abdomen. Some parasitic copepods produce up to six or eight egg sacs, while others produce the eggs in long strings, which may coil into a tangled mass.

Nauplius and other larvae

The most widespread and typical crustacean larva to emerge from the egg is called a nauplius (Figure 2). The main features of a nauplius are a simple, unsegmented body, three pairs of appendages (antennules, antennae, and

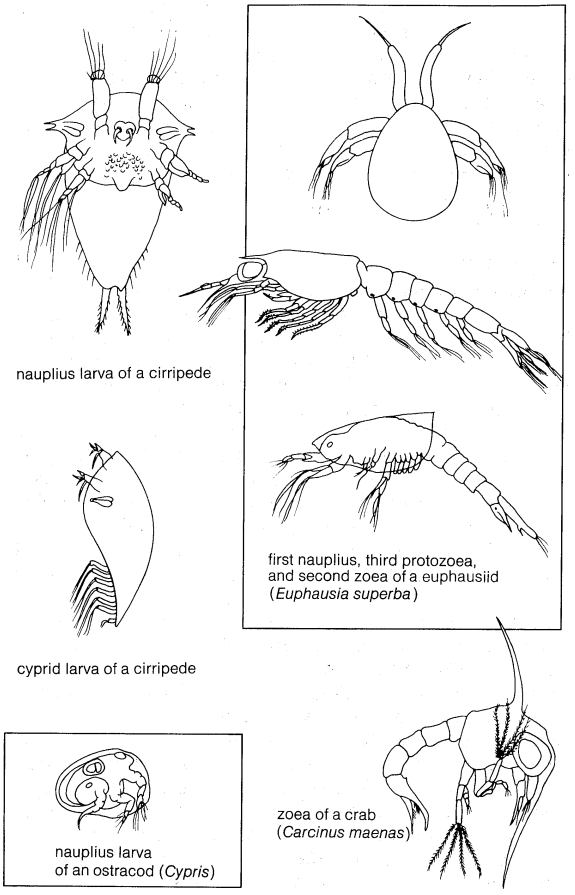



Figure 2: Crustacean larval types.

From (*Euphausia superba*) F.C. Fraser, "On the Development and Distribution of the Young Stages of Krill (Euphasia superba)," *Discovery Reports XIV* (1936), Cambridge University Press; (others) W.T. Calman, *A Treatise on Zoology* (1909), ed. Ray Lankester (copyright A. & C. Black Ltd., London; reprint Asher, Amsterdam, 1964)

mandibles), and a single simple nauplius eye. When the body becomes more elongated and segmented, the nauplius is called a metanauplius. Nauplius larvae are found in the cirripedes, ostracods, branchiopods, copepods, euphausiids, the decapod peneid prawns, and the subclass Thecostraca. Many of the other groups have embryonic stages like the nauplius, or they have larvae with some similarities to the nauplius.

The most primitive type of development from a nauplius is found in the anostracan fairy shrimps, where the young animal gradually adds new segments and appendages as it undergoes a long series of molts. In the free-living copepods, for example, the nauplius goes through five molts and then changes into a copepodid, which resembles the adult except that it does not have a full complement of limbs. These limbs gradually develop over another five molts; when the adult form is reached, the copepod does not molt again. The cirripede barnacle nauplius molts several times and then changes into a cyprid, which has a two-part carapace enclosing six pairs of trunk limbs. The cyprid attaches to a solid object and then metamorphoses into an adult. Larval ostracods are basically nauplii with a bivalved carapace. The euphausiid nauplius is followed by a complex series of shrimplike larvae.

The nauplius of the peneid prawns is followed by a sequence of larval forms characterized by their methods of locomotion: the metanauplius still swims with its antennae, the protozoea also uses its antennae but has developed a small carapace and some thoracic limbs, the zoea uses its thoracic limbs for swimming, and the postlarval stages use the abdominal appendages. Most decapods omit the nauplius stage and hatch as zoeae, which may be heavily ornamented with spines. The zoea of a crab changes into a megalops, which resembles a small crab with its tail extended behind it.

Some crustaceans bypass the larval stages, and the young emerging from the eggs resemble the adults. This is found in the branchiopod order Anomopoda, such as in *Daphnia,* in most isopods and amphipods, and in some decapods, including freshwater crabs and crayfish and some deep-sea and Arctic groups.

**Ecology.** Crustaceans play many roles in aquatic ecosystems. The planktonic forms—such as the copepod *Calanus* and the krill *Euphausia*—graze on the microscopic plants floating in the sea and in turn are eaten by fishes, seabirds, and whales. Benthic (bottom-dwelling) crustaceans are a food source for fish, and some whales feed extensively on benthic amphipods. Crabs are important predators, and the continuing struggle between them and their prey prompts the evolution of newer adaptations: the massive shells of many marine mollusks are thought to be a protective response to the predatory activities of crabs; in turn the crabs develop larger and more powerful pincers.

Elements of food chains

Crustaceans also can be parasites, and the copepods in particular parasitize other aquatic animals ranging from whales to sea anemones. The larger crustaceans are often parasitized by smaller crustaceans; for example, there are parasitic isopods that dwell in the gill chambers of decapod prawns. Freshwater crustaceans can serve as intermediate hosts for the lung fluke, *Paragonimus.*

## FORM AND FUNCTION

**General features.** The crustacean body is difficult to characterize because the orders include a great variety of forms. The anatomy of a decapod is presented in Figure 3. The basic crustacean body consists of a number of segments, or somites. These somites sometimes are fused to form rigid areas and sometimes are free, linked to each other by flexible areas that allow some movement. Each somite has the potential for bearing a pair of appendages, although in various crustacean groups appendages are missing from certain somites. The appendages are also jointed with flexible articulations.

At the front, or anterior end, of the body there is an unsegmented, presegmental region called the acron. In most crustaceans at least four somites fuse with the acron to form the head. At the posterior end of the body there is another unsegmented region, the telson, that may bear two processes, or rami, which together form the furca. These two processes at the tail end of the body vary greatly in form; in many crustaceans they are short, but in some they may be as long as the rest of the body. The Crustacea as a whole shows great variation in the number of somites and the amount of fusion that has taken place. In the class Malacostraca, which includes the decapods, there is a consistent body plan: the trunk (which follows the head) is divided into two distinct regions, an anterior thorax of eight somites and a posterior abdomen of seven somites, although as a rule only six are evident in the adult. The reproductive ducts of a male malacostracan typically open on the last thoracic somite, and the female reproductive ducts open on the sixth thoracic segment.

The carapace is a widespread crustacean feature, arising during development as a fold from the last somite at the back of the head. It may form a broad fold extending toward the rear over the back, or dorsal surface, of the trunk, as in the notostracan tadpole shrimps, but it often

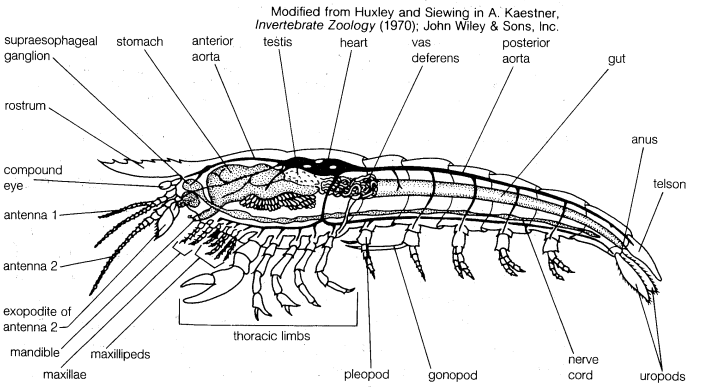



Figure 3: External and internal anatomy of a male decapod crustacean.

encloses the entire trunk, including limbs and gills. In the clam shrimps (orders Spinicaudata and Laevicaudata) and the ostracods, the carapace is split into two parts. In many decapods the carapace projects forward to form a rostrum, which is often sharply pointed and toothed. The carapace is absent from the anostracans, amphipods, isopods, and the superorder Syncarida. The barnacles attach to hard surfaces and use the highly modified carapace to form a mantle.

**External features.** *Appendages.* There is great diversity among the appendages of the Crustacea, but it is thought that all the different types can be derived either from the multibranched (multiramous) limb of the class Cephalocarida or from the double-branched (biramous) limb of the class Remipedia. A biramous limb (Figure 4) typically has a basal part, or protopodite, bearing two branches, an inner endopodite and an outer exopodite. The protopodite can vary greatly in its development and may have additional lobes on both its inner and outer margin, called, respectively, endites and exites. The walking legs of many malacostracans have become uniramous, losing the exopodite.

Appendages and trunk limbs

If one starts at the head of a crustacean and works toward the rear, the following appendages are encountered (some may be missing in certain species): antennae 1, or antennules; antennae 2, or antennae proper; mandibles; maxillae 1, or maxillulae; maxillae 2, or maxillae proper; and a variable number of trunk limbs. The trunk limbs all may be similar, as in the anostracans and the classes Cephalocarida and Remipedia, or they may be differentiated into distinct groups. In the copepods the first pair of trunk limbs is used in the collection of food and is called the maxillipeds. In the decapods there are three sets of paired maxillipeds. In the copepods the maxillipeds are followed by four pairs of swimming legs; a fifth pair is sometimes highly modified for reproductive purposes and is sometimes reduced to a mere vestige. Behind the maxillipeds of the decapods there are five pairs of thoracic limbs, a variable number of which may bear pincers, or chelae. In the crabs there is a single obvious pair of chelae, but in some of the prawns there may be up to three pairs of less conspicuous pincers. The abdomen of a decapod normally bears six pairs of biramous appendages, which are used in swimming in many shrimps and prawns, while in the crabs and crayfish the first two pairs in the male are modified to help in sperm transfer during mating. The last pair of abdominal limbs is frequently different from the others and is called the uropods. In shrimps and lobsters the uropods together with the telson form a tail fan.

Changes over time

The appendages often change both their form and their function during the life cycle of a crustacean. In most adults the antennules and antennae are sensory organs, but in the nauplius larva the antennae often are used for both swimming and feeding. Processes at the base of the antennae can help the mandibles push food into the mouth. The mandibles of a nauplius have two branches with a chewing or compressing lobe at the base; they also may be used for swimming. In the adult the mandible loses one of the branches, sometimes retaining the other as a palp, and the base can develop into a powerful jaw. An alternative development is found in some of the blood-sucking parasites, in which the mandibles form needlelike stylets for piercing their hosts.

Modified from R.R. Hessler, "Cephalocarida, Comparative Skeleton Musculature," *Connecticut Academy of Arts & Sciences Memoirs* (1964)

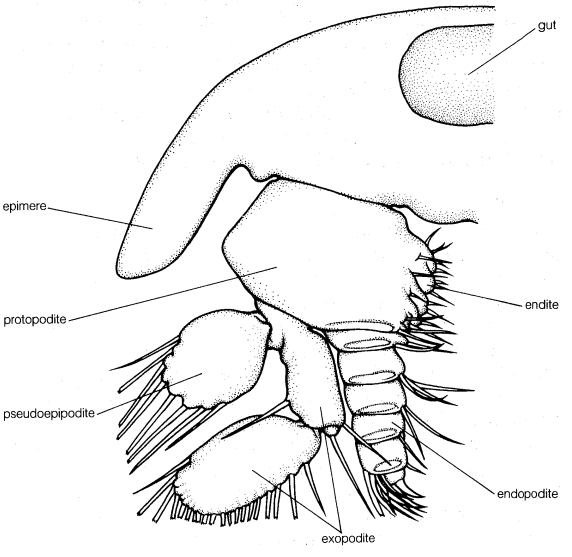



Figure 4: Thoracic leg of the cephalocarid *Hutchinsoniella macracantha*, with cross section of the left half of the trunk somite.

*Exoskeleton.* The outer covering of a crustacean is variously called the integument, cuticle, or exoskeleton. It protects the body and provides attachment sites for muscles. The thickness of the cuticle can vary from a thin, flexible membrane, as in some parasitic copepods, to a massive rigid shell, as in crabs. The cuticle is secreted by a single layer of cells called the epidermis. The outermost layer, or epicuticle, lacks the chitin present in the thicker innermost layers, or procuticle. The procuticle is made up of layers of chitin fibres intermeshed with proteins and, in many species, with calcium salts. Fine tubular extensions (pore canals) from the epidermis pass through the procuticle toward the surface.

Growth and molting

A typical crustacean grows in a series of stages. The hard exoskeleton prevents any increase in size except immediately after molting. The sequence of events during molting can be divided into four main stages: (1) Proecdysis, or premolt, is the period during which calcium is resorbed from the old exoskeleton into the blood. The epidermis separates from the old exoskeleton, new setae form (see below *The nervous system*), and a new exoskeleton is secreted. (2) Ecdysis, or the actual shedding of the old exoskeleton, takes place when the old exoskeleton splits along preformed lines. In the lobster it splits between the carapace and the abdomen, and the body is withdrawn through the hole, leaving the old exoskeleton almost intact. In isopods the exoskeleton is cast in two parts; the front portion may be cast several days after the hind part. Immediately after ecdysis the crustacean swells from a rapid intake of water. (3) Metecdysis, or postmolt, is the stage in which the soft cuticle gradually hardens and becomes calcified. At the end of this stage the cuticle is complete. (4) Intermolt is a period of variable duration, from a few days in small forms to a year or more in some of the large forms. Some crustaceans, after passing through a series of molts, reach a stage where they do not molt again; this is called a terminal anecdysis.

**Internal features.** *The nervous system.* The basic nervous system of a crustacean consists of a brain, or supraesophageal ganglion, connected to a ventral nerve cord of ganglia, or nerve centres. In primitive forms, like the anostracan fairy shrimps, the brain has nerve connections with the eyes and antennules, but the nerves to the antennae come from the connecting ring around the esophagus. In more advanced forms the antennal nerves originate in the brain. The first ventral nerve centre under the esophagus (subesophageal ganglion) is usually formed by the fusion of the ganglia from the mandibular, maxillulary, and maxillary segments, but other ganglia may be incorporated. Often there is a chain of ganglia extending the length of the trunk, but in short-bodied forms, such as barnacles and crabs, all the ventral ganglia may fuse into a single mass.

Compound eyes

The most conspicuous sense organs are the compound eyes. In a typical decapod each eye consists of several hundred tubular units radiating from the end of an optic nerve. Each of these units is a miniature eye, with a central optical tract isolated from the others by two groups of pigment cells. These pigment cells can expand and contract to cover varying amounts of each tubular eye, enabling the eyes to be used over a range of light intensities. The image obtained with such an eye is a mosaic, but there is evidence from the behaviour of the advanced crabs that they perceive a good image and that they can detect small movements. Single median eyes are also found in crustaceans, particularly in the nauplius larvae. Only three or four simple units are usually found in the nauplius eye,

which is innervated by a median nerve from the forebrain. The median eye also may persist through to the adult stage. In the copepods the median eye is the only eye, but in some groups it may persist even when the compound eyes have developed. In the Notostraca, the name *Triops* draws attention to this phenomenon, but it is also known in other groups, including some of the decapods.

Setae

Other stimuli are detected by means of various setae, or hairlike processes, that project from the surface of the exoskeleton and are connected to a nerve supply. Some setae are tactile, detecting contact and movement when deflected. Other setae are used in association with statocysts. Statocysts are paired organs, located at the base of the antennules in decapods or at the base of the uropods in mysids, that enable the crustacean to orient itself with respect to gravity. Each statocyst is a rounded sac containing one or more small granules, called statoliths, that rest on numerous small setae. Any change in orientation causes the statoliths to impinge on the setae at a different angle, and this information is relayed to the brain so that corrective action can be taken. Finally, other setae are chemosensory; they detect a wide range of chemical substances. Such setae are usually tubular and thin-walled, sometimes with a small pore at the top. They are especially abundant on the antennules and mouthparts.

*The digestive system.* The gut is usually direct in its passage through the body and is coiled in only a few water fleas of the order Anomopoda. The foregut shows the greatest range of structure; in some it is a simple tube, while in decapods it reaches great complexity in having a structure called the gastric mill. This consists of a series of calcified plates, or ossicles, that are moved against each other by powerful muscles, making an efficient grinding apparatus. The junction between the mill and the midgut is guarded by a filter of setae, which prevent large particles from passing into the midgut. The structure of the midgut is also variable but generally has one or more diverticula, or pouches, which are involved in various digestive processes. These diverticula may be simple, as in *Daphnia,* or complex and glandular, as in the decapods. The hindgut is usually relatively short and lined with cuticle. The exit is controlled by a muscular anus, which in some forms had dilator muscles that control anal swallowing.

Gastric mill

*The excretory system.* Two different excretory organs are found in the crustaceans: the antennal gland and the maxillary gland. Both have the same basic structure: an end sac and a convoluted duct that may expand into a bladder before opening to the outside. In most adult crustaceans only one or the other gland functions. The functional gland may change during the life cycle.

The main function of the antennal and maxillary glands is the regulation of ionic balance. The total balance of salts and water is also controlled in part by the gut, which can absorb both. The antennal gland also has been shown to reabsorb glucose. Most crustaceans excrete the end product of nitrogen metabolism, in the form of ammonia, through the gills. Some of the more terrestrial forms produce urea or uric acid, which may be stored in special large cells near the bases of the legs.

*The respiratory system.* Many of the smaller crustaceans have no special respiratory organs. Gas exchange takes place through the entire thin integument. The inner wall of the carapace, facing the trunk, is often rich with blood vessels and may in many groups be the only respiratory organ. Gills, when present, are formed by modifications of parts of appendages, most often the epipodites. These thin-walled, lamellate structures are present on some or all of the thoracic appendages in cephalocarids, fairy shrimps, and many malacostracans. In mantis shrimps (order Stomatopoda), for example, gills are found on the exopodites of the pleopods. In euphausiids the single series of branched epipodial gills are fully exposed. In decapods the gills, protected by the overhanging carapace, are arranged in three series at or near the limb bases. As an adaptation to aerial respiration, the branchial chambers are greatly enlarged in certain land crabs and serve as lungs, the inner membrane being richly supplied with blood vessels. In isopods the respiratory function has been taken over by the abdominal appendages; either both rami or the endopodite become thin and flattened. Most sow bugs and pill bugs have, in addition, trachea-like infoldings in some of the exopodites.

*The circulatory system.* As in other arthropods, the blood flows in sinuses, or channels, without definite walls. Cirripedes and many ostracods and copepods have no heart, the blood being kept in motion by either a blood pump or rhythmic movements of the body, gut, or appendages. When present, the heart lies in a blood sinus, or pericardium, with which it communicates by paired valvular openings, or ostia. In the more primitive crustaceans, such as fairy shrimps or stomatopods, the heart is a long tube, with spiral muscles in its wall, and extends almost the entire length of the trunk; there is a pair of ostia in each somite except the last. In more-advanced crustaceans, however, the heart may be shortened, and the number of ostia may be reduced to three pairs or less. The position of the heart depends on that of the respiratory organs; it usually lies in the thorax or cephalothorax but is mainly in the abdomen of isopods. Malacostracans have a well-developed system of elastic-walled arteries, including an anterior and usually a posterior aorta.

The heart

The red respiratory, or oxygen-carrying, pigment hemoglobin has been observed in the blood of the Branchiopoda and in other classes except Malacostraca. Hemocyanin is the respiratory pigment in the malacostracan decapods and stomatopods.

*Hormones.* Hormones are substances produced in one part of the body that act on cells in some other part of the body. The secretory system that produces these substances is known as the endocrine system. Most of the information about crustacean hormones has been obtained from studies on the decapods, but a fair amount is also known about the hormones of the isopods and amphipods.

The X-organ–sinus-gland complex is located in the eyestalk. The X-organ passes its secretions to the sinus gland, which acts as a release centre into the blood. Hormones liberated from the sinus gland have been shown to influence molting, gonad development, water balance, blood glucose, and the expansion and contraction of pigment cells both in the general body and in the retina of the eye. The Y-organs lie in the maxillary segment of decapods and are the source of molting hormones, or ecdysteroids, which promote molting and interact with molt-inhibiting hormones from the X-organ.

X-organ–sinus-gland complex

The brain and thoracic nerve centres produce hormones that promote the development of the sex organs. In addition, certain glands attached to the male reproductive ducts control the development of the male reproductive system; their removal from a young male will cause it to develop into a female. The female ovary also acts as an endocrine organ; its endocrine secretions control the development of the female reproductive system. The brood pouch in both amphipods and isopods also develops under the influence of ovarian secretions. A hormonal system controls the beating of the heart. Nerves from the thoracic centres end in fine secretory fibres in the membrane enclosing the space around the heart (pericardium) and secrete substances that typically produce an increase in both frequency and amplitude of the heartbeat.

## EVOLUTION AND PALEONTOLOGY

There are two approaches to the study of crustacean evolution. The first involves the interpretation of the evidence from comparative anatomy. The second involves a consideration of the fossil record.

Various attempts have been made to construct a hypothetical ancestral crustacean from which it would be possible to derive all the others. The prerequisites for such an ancestor seem to be an elongated body, two pairs of appendages in front of the mouth, a pair of mandibles behind the mouth, and numerous trunk segments with appendages that form a continuous series of similar structure. Before the discovery of the class Cephalocarida, some of the primitive members of the class Branchiopoda, such as the orders Anostraca and Notostraca, were thought to show what such an ancestor might have been like. The Cephalocarida, in having trunk limbs with a jointed inner branch and a platelike outer branch, further showed a pos-

sible original structure from which almost any crustacean limb could have been derived. The discovery of the class Remipedia, with a long series of similar trunk limbs, has reopened the question of the original form of the trunk limb in the ancestral crustacean. The Remipedia are undoubtedly primitive, but they do have some adaptations as cave dwellers. The question is still open as to whether the carapace is a primitive crustacean structure or whether it is a feature that has evolved independently in each group.

The fossil record, although fairly rich, has not solved any of the questions about the early evolution of the Crustacea. The earliest of the definite fossil crustaceans are ostracods, a relatively specialized group. There are also indications from the Burgess shales of the Cambrian period (570 to 505 million years ago) that many features of crustacean organization had already evolved by this time. It is only when the later, more highly evolved class Malacostraca is studied that there is good agreement between comparative anatomy and the fossil record. The decapod *Palaeopalaemon,* a shrimplike form, occurs in the Devonian (408 to 360 million years ago), crayfish occur in the Late Permian (258 to 245 million years ago), and allies of the hermit crabs (Anomura) are found in the Jurassic (208 to 144 million years ago). The true crabs (Brachyura), which represent one of the pinnacles of crustacean evolution, do not occur until the beginning of the Cretaceous (144 to 66.5 million years ago).

## CLASSIFICATION

**Distinguishing taxonomic features.** In classifying the Crustacea, a variety of characters are important: the form and extent of the carapace, if present; the number of trunk somites, or segments, and how many fuse with the head or with the telson; the number and degree of specialization of the trunk limbs; the presence or absence of paired eyes and of a caudal furca—*i.e.,* a forked-tail process; and the position and kind of respiratory organs. The position of the genital openings, the mode of attachment of the eggs to the female, and the stage at which the first larva hatches may also be significant. Parasitic and sedentary forms may differ markedly as adults from free-living species.

(I.G./J.Gre.)

**Annotated classification.** The following classification is based largely on that given in D.E. Bliss, *The Biology of Crustacea,* vol. 1, (1982) but has been modified to take account of advances made since that date. Groups marked with a dagger (†) are extinct and known only from fossils.

### SUBPHYLUM CRUSTACEA

Two pairs of sensory appendages in front of mouth, and 3 pairs of jaws behind mouth; some parasitic and lack all appendages when adult; mostly aquatic; about 39,000 species known.

**Class Cephalocarida** (horseshoe shrimps)

Holocene; primitive; blind; head shield without carapace; maxilla and all trunk limbs alike, with jointed inner branch and leaflike outer branches; abdominal segments without limbs; telson and furca present; length about 3 mm; marine, intertidal down to 300 m; few known species.

**Class Branchiopoda**

Early Devonian to present; limbs usually leaflike; maxillae reduced; eyes sometimes stalked, usually sessile (unstalked), often fused to form a single large median eye; nauplius, but some with direct development; predominantly freshwater, some marine, and some in strong inland brines; about 900 species.

**†Class Enantiopoda**

Carboniferous; single fossil, *Tesnusocaris.*

**Class Remipedia**

Holocene; body elongated; more than 30 segments, each with biramous appendages projecting sideways; antennules biramous; maxillules, maxillae, and maxillipeds uniramous and grasping; marine cave dwellers; few species.

**Class Maxillopoda**

Five pairs of head appendages; single, simple, median eye; antennules uniramous; maxillae usually present; up to 11 trunk segments.

*Subclass Thecostraca*

Bivalved carapace of cypris larva forms an enveloping mantle in the adult; parasitic forms recognizable only by larval stages.

*Order Ascothoracica.* Cretaceous to present; parasites on sea anemones and echinoderms; body typically enclosed in a bivalved carapace; some with segmented abdomen and caudal furca; others distorted by outgrowths of the gut and ovary, giving a bushlike appearance; males dwarfed, living in mantle cavities of females; marine; about 50 species.

*Order Rhizocephala.* Parasites on other crustaceans, mostly decapods; larvae typical nauplii and cyprids; adults ramify inside hosts and produce 1 or more reproductive bodies outside the host; marine; about 230 species.

*Order* (*Subclass*) *Cirripedia* (barnacles). Late Silurian to present; sedentary; 6 pairs of trunk limbs (cirri); larvae free-swimming; sessile adults with carapace developed into a mantle.

*Subclass Tantulocarida*

Holocene; eggs give rise to a tantulus larva with head shield and 6 pairs of thoracic limbs; adult females form large dorsal trunk sac between head shield and trunk, often losing the trunk; males with 6 pairs of trunk limbs; parasites on other crustaceans; marine; about 10 species.

*Subclass Ostracoda* (mussel or seed shrimps)

Cambrian to present; body short; bivalved carapace encloses trunk and limbs; living forms have up to 7 pairs of appendages; most fossils known only from shells (carapaces); marine, freshwater, and some terrestrial; more than 2,000 living species worldwide.

†*Order Bradoriida.* Cambrian to Ordovician.

†*Order Phosphatocopida.* Cambrian; remarkable fossils with up to 9 pairs of well-preserved appendages.

†*Order Leperditicopida.* Cambrian to Devonian.

†*Order Beyrichicopida.* Silurian to Carboniferous.

*Order Myodocopida.* Silurian to present; antennal notch in shell; 5 pairs of postoral appendages; maxilla with a large respiratory plate; eyes usually present; marine.

*Order Halocyprida.* Silurian to present; 5 pairs of postoral appendages; maxilla leglike; no eyes; marine.

*Order Cladocopida.* Silurian to present; only 3 pairs of postoral appendages; marine.

*Order Platycopida.* Ordovician to present; antennae biramous; 4 pairs of postoral limbs; marine.

*Order Podocopida.* Ordovician to present; antennae uniramous; 5 pairs of postoral appendages; marine, freshwater, and terrestrial.

*Subclass Branchiura*

*Order Arguloida* (fish lice). Wide, flat carapace; paired compound eyes; unsegmented abdomen; 4 pairs of trunk limbs; fish parasites; capable of free swimming; mostly freshwater but some marine; about 150 species.

†*Subclass Skaracarida*

Late Cambrian; 12 trunk segments; no thoracic appendages apart from maxillipeds.

*Subclass Copepoda*

Miocene to present; no carapace; no compound eyes; 1 or more trunk segments fused to head; typically 6 pairs of thoracic limbs; no abdominal limbs; larva usually a nauplius; free-living and parasitic; worldwide; marine, freshwater, and some semiterrestrial; at least 10,000 species.

*Order Calanoida.* Antennules long, usually held stiffly at right angles to the length of the body; heart present; thorax articulates with a much narrower abdomen; fifth leg biramous; worldwide; marine and freshwater; mostly planktonic.

*Order Misophrioida.* Carapace-like extension from the head covers the first segment bearing a swimming leg; heart present in some; no eyes; antennule with up to 27 segments; fifth leg biramous; marine.

*Order Mormonilloida.* Antennule with 3 or 4 long segments and long setae; fifth leg absent; marine.

*Order Harpacticoida.* Antennules short; abdomen not markedly narrower than the thorax; articulation between thoracic segments 5 and 6; mostly benthic, some tunnel in the fronds of seaweeds; usually 1 egg sac but some with 2; marine and freshwater, with some semiterrestrial on damp forest floors.

*Order Cyclopoida.* Antennules medium length; thorax wider than abdomen; articulation between thoracic segments 5 and 6; mandibles with biting or chewing processes; eggs normally carried in 2 egg sacs; fifth leg uniramous; marine and freshwater.

*Order Poecilostomatoida.* Parasites and commensals of fish and invertebrates; mouth not tubelike or suckerlike; mandibles reduced; adult segmentation often reduced or lost; mostly marine, few freshwater.

*Order Siphonostomatoida.* Mouth tubelike or forms a sucker with styletlike mandibles; adult segmentation reduced or lost; parasites and commensals on fish and invertebrates; mostly marine, some freshwater.

*Order Monstrilloida.* Parasites on marine worms and mollusks; adults free-swimming; lack mouthparts and gut; biramous swimming legs.

*Subclass Mystacocarida* (mustache shrimps)

Elongated; blind forms living in spaces between sand grains;

antennules uniramous; antennae and mandibles biramous with long branches extending sideways; trunk limbs vestigial but caudal rami well-developed and pincerlike; marine; about 9 species.

**Class Malacostraca**

Cambrian to present; typically with compound eyes, stalked or sessile; 8 thoracic and 6 abdominal segments, each potentially capable of bearing a pair of appendages; about 22,000 species.

*Subclass Phyllocarida*

Early Cambrian to present.

†*Order Archaeostraca.* Devonian to Triassic.

†*Order Hoplostraca.* Carboniferous.

*Order Leptostraca.* Permian to present; bivalved carapace encloses 8 pairs of leaflike limbs; movable rostrum; telson with caudal rami; marine; about 10 species.

*Subclass Hoplocarida*

Carboniferous to present.

*Order Stomatopoda* (mantis shrimps). Jurassic to present; eyes stalked; 2 movable segments in head; carapace leaves 4 thoracic segments uncovered; second thoracic limbs massive; marine; about 350 species.

†*Order Palaeostomatopoda.* Carboniferous.

†*Order Aeschronectida.* Carboniferous.

*Subclass Eumalacostraca*

Late Devonian to Holocene; carapace (when present) not bivalved; rostrum fixed; first antenna 2-branched; thoracic legs with slender, many-segmented outer branch and stout, 7-segmented inner branch, often pincerlike, used in walking or food-gathering; 6 (rarely 7) abdominal segments, with pleopods and terminal uropods.

*Superorder Syncarida.* Carboniferous to present; no carapace.

†*Order Palaeocaridacea.* Carboniferous to Permian; first thoracic segment not fused to head; abdominal pleopods 2-branched, flaplike; 4 families.

*Order Anaspidacea.* Permian to present; with or without eyes; antennules biramous; abdominal appendages well-developed; telson without a furca; South Australia and Tasmania; freshwater; about 8 species.

*Order Stygocaridacea.* Blind, elongated forms with a small rostrum; first thoracic segment fused to head but sixth abdominal segment free; furca present; abdominal appendages reduced or absent; South America and New Zealand; freshwater, in spaces between sand grains; about 5 species.

*Order Bathynellacea.* Blind, elongated forms, without a rostrum; first thoracic segment not fused to head but sixth abdominal segment fused with telson; antennules uniramous; worldwide; freshwater, in spaces between sand grains; about 100 species.

*Superorder Peracarida.* Females with a ventral brood pouch formed by plates at the bases of some of the thoracic limbs; development direct, with offspring resembling adults.

*Order Mysidacea* (opossum shrimps). Triassic to present; carapace well-developed, covering most of thorax; 3–30 mm, with a few much larger; worldwide; mainly marine but some in brackish and fresh water; about 780 species.

*Order Cumacea.* Permian to present; head and carapace much wider than trunk; uropods long and rodlike; 1–35 mm; marine; about 800 species.

*Order Spelaeogriphacea.* Holocene; carapace short, fused to first and covering part of second thoracic segment; 4 pairs of well-developed abdominal appendages; about 8 mm; cave-dwelling; South Africa; freshwater; 1 species.

*Order Mictacea.* Holocene; no functional eyes; carapace forms small lateral folds covering bases of mouthparts and maxillipeds; all trunk segments free; antennules biramous; thoracic limbs with exopods; abdominal appendages reduced, uniramous; 2.7–3.5 mm; deep-sea or in marine caves; 2 species.

*Order Tanaidacea.* Permian to present; carapace short, fused to first 2 thoracic segments; second pair of thoracic limbs usually with pincers; abdomen short, usually with 5 pairs of biramous appendages; 2–25 mm; mainly marine; about 500 species.

*Order Isopoda* (pill bugs, wood lice, sea slaters). Carboniferous to present; eyes sessile; no carapace; abdominal appendages flattened and respiratory; thoracic limbs without exopods; some parasites highly modified as adults; most species 5–30 mm but some up to 270 mm; worldwide; marine, freshwater, and terrestrial; about 4,000 species.

*Order Amphipoda* (beach hoppers, scuds, well shrimps). Eocene to present; eyes sessile; no carapace; thoracic limbs have respiratory plates at base; few parasites; most 5–50 mm but up to 140 mm; worldwide; mainly marine but also numerous in fresh water; about 6,000 species.

*Superorder Eucarida.* Carapace large, fused dorsally to all thoracic segments; eyes stalked; development usually involves larval forms but is sometimes direct.

*Order Euphausiacea* (krill). Holocene; carapace does not cover gills; thoracic limbs with 2 well-developed branches; eggs usually shed freely; first larva a nauplius; 6–81 mm; worldwide; marine; about 85 species.

*Order Amphionidacea.* Holocene; carapace large; mandible and maxillule vestigial; thoracic limbs with small outer branch; ventral brood pouch formed by large forwardly projecting first abdominal appendages; 2–3 cm; worldwide; marine, pelagic; 1 species.

*Order Decapoda* (shrimps, prawns, lobsters, crayfish, crabs). Devonian to present; carapace large, enclosing gills; first 3 pairs of thoracic appendages modified for feeding (maxillipeds); eggs often attached to abdominal appendages; worldwide; mostly marine but also freshwater and a few terrestrial; about 10,000 species.

*Superorder Pancarida*

*Order Thermosbaenacea.* Holocene; eyes reduced or absent; brood pouch formed from dorsal extension of carapace; length about 4 mm; fresh and brackish water, some in warm springs; about 9 species.

**Critical appraisal.** There is no universal agreement on the classification of the Crustacea and even less agreement on the interrelationships among the various groups. Alternative classifications of the classes Branchiopoda and Malacostraca are discussed below. Some authorities, such as the author of the Cirripedes below, rank the cirripedes as a subclass. There is also some disagreement about the limits of the class Maxillopoda. Some would include the class Cephalocarida, others would exclude the class Ostracoda, and yet others do not regard the Maxillopoda as a valid group and would raise the maxilloped subclasses Copepoda and Ostracoda to separate classes. Some of the parasitic forms are sometimes separated and ranked as separate orders.

# MAJOR CRUSTACEAN GROUPS

## Branchiopods (fairy shrimps, tadpole shrimps, brine shrimps, and water fleas)

Branchiopods are generally regarded as primitive crustaceans. Their long fossil record dates back to the Devonian period (408 to 360 million years ago). Although certain members of the group, such as the order Anostraca, are mainly confined to temporary pools, the water flea, order Anomopoda, is so successful that there are few fresh waters in the world without one or more species of anomopod.

### GENERAL FEATURES

**Size range and diversity of structure.** The smallest branchiopods are found among the anomopods, where some species are only 0.25 millimetre (0.01 inch) long. The largest living branchiopod is *Branchinecta gigas,* a fairy shrimp that reaches a length of 10 centimetres (3.9 inches). Some members of the fossil order Kazacharthra also grew to a length of 10 centimetres.

The class Branchiopoda is divided into 10 orders, two of which are extinct and known only through the fossil record. The eight living orders show a great diversity of form. In the Laevicaudata, for example, the number of trunk segments remains constant; there are 12 pairs of trunk limbs in the female and 10 pairs in the male. In the Spinicaudata, however, the number of paired trunk segments varies among its members from 12 up to 32 in some species. A carapace is present in the orders Ctenopoda and Anomopoda, but it encloses only the trunk, leaving the head free. In the orders Onychopoda and Haplopoda the carapace does not enclose the trunk limbs but forms a brood pouch on the dorsal surface. The anostracans (fairy shrimps and brine shrimps) lack a carapace and have stalked eyes, in contrast to the other living group, whose eyes are set into the head.

The two fossil groups as well differ markedly from each other. The order Lipostraca lacked a carapace and had 13 pairs of trunk limbs and a pair of large antennae, which appear to have been used in swimming. The order Kazacharthra had a well-developed carapace and six pairs of large thoracic limbs. The main structural feature linking these diverse forms, both living and fossil, is the flattened, or paddlelike, trunk limb, which often but not always is used in filter feeding. In the orders Onychopoda and Haplopoda even this feature is modified, and the trunk limbs have become specialized for grasping prey.

**Distribution and abundance.** Branchiopods are found worldwide in fresh waters. The anostracans, notostracans, and the orders Laevicaudata and Spinicaudata are particularly characteristic of temporary waters, where they survive dry periods as resting eggs. The anostracan *Branchinecta paludosa* and the notostracan *Lepidurus arcticus* are regularly found in small pools of the Arctic tundra regions. These pools are temporary in the sense that they freeze solid in winter. A few species in these groups are found in permanent waters.

Marine and freshwater habitats

The ctenopods and anomopods play an important role in fresh waters throughout the world, both in the open waters of large lakes and in the shallower plant-rich zones around the margins of ponds and lakes. Their diet consists to a large extent of algae and bacteria, and in turn they are important as food for fish. A few species of the orders Ctenopoda and Onychopoda are found in the sea, and the anostracan brine shrimp *Artemia* is often abundant in inland saline waters. The immature forms of *Artemia* are used as food for the young of various marine fishes commercially reared in artificial conditions.

## NATURAL HISTORY

**Reproduction and life cycles.** A typical branchiopod begins its life cycle as a nauplius larva (Figure 2), which has a simple undivided body and three pairs of appendages: antennules, antennae, and mandibles. The antennae are used for swimming. As the nauplius feeds and grows, it gradually changes into the adult form—the body becomes segmented, or jointed, and additional limbs develop. In adult anostracans and notostracans the antennae lose their swimming function, but in adults of the other six orders they remain large and functional. The spinicaudate *Cyclestheria* lays its eggs in the space between the trunk and the carapace. These eggs develop rapidly into miniatures of the adult, skipping the larval stages. A similar mode of reproduction is found in the ctenopods, anomopods, and onychopods.

Branchiopods mature rapidly. A small cladoceran can lay eggs in the warm water of a temporary desert pool after two days. In temperate latitudes, a cladoceran may mature in less than a week during a warm summer.

The sexual arrangement in some branchiopods is that of separate males and females. Others are modified so that their eggs develop without fertilization (parthenogenesis). Some branchiopods have both male and female reproductive structures in one individual.

Among the branchiopods the anomopods show the greatest variety of reproductive habits. Under favourable conditions the eggs are laid in a brood pouch between the carapace and the trunk. There they develop rapidly and, after about two days, hatch as females, which in turn lay eggs that give rise to more females. No males are necessary for this process. When food is scarce or when there is a sudden temperature change, some of the eggs develop into males, and some of the females begin producing eggs that must be fertilized by sperm from the males. These fertilized eggs are remarkably resistant to unfavourable environmental conditions; even if frozen or dried, they will hatch when returned to favourable conditions. Many anomopods survive the winter as fertilized eggs; species that dwell in temporary pools lay such eggs to survive periods of drought. Certain Arctic or alpine anomopods, such as *Daphnia middendorffiana,* produce resistant eggs that do not require fertilization. The resistant, or dormant, fertilized eggs normally hatch in the following spring, giving rise to the usual miniature adult females. In *Leptodora* the resting egg hatches into a nauplius larva, although the rapidly developing eggs produced in the summer give rise to miniature females.

Resistance of eggs to environmental hazards

Members of the other branchiopod orders also can produce dormant, fertilized eggs. Many desert-pool species produce only resting eggs; they must abbreviate their life cycles to coincide with the brief period of favourable conditions, and their eggs must be capable of remaining dry for long periods, sometimes several years.

**Ecology.** Some species of *Daphnia* in temperate lakes show a remarkable seasonal change in form. In the winter the females have rounded heads, but the females of generations in late spring and summer have pointed heads. High temperature and water turbulence favour the development of a pointed head. The most plausible explanation seems to be related to predation by fish. The feeding activity of plankton-eating fish decreases in winter and increases rapidly in the spring and summer. The fish select the large *Daphnia,* the most conspicuous parts of which are the eye and the carapace with its enclosed limbs and eggs. When the head becomes pointed and enlarged, the size of the carapace is reduced, and the eye is often smaller. Thus, there is an overall decrease in conspicuousness that occurs in the summer forms.

Food gathering

The trunk limbs of all branchiopods are used to gather food. Filters formed by setae, or fine hairs, separate the food particles from the water, and an elaborate mechanism shifts food from the filters to the mouth. The filters enable branchiopods to collect material as small as bacteria for food. The ability to utilize bacteria is important in cleansing water in reservoirs, where *Daphnia* is often abundant.

The notostracans *Triops* and *Lepidurus* can collect small particles, but they can also act as predators. *Lepidurus arcticus* has been observed feeding on another Arctic branchiopod, the anostracan *Branchinecta paludosa,* which often lives in the same tundra pools. Sometimes a species changes its feeding habits with age. The large fairy shrimp *Branchinecta ferox* feeds on small particles when young but becomes a predator when mature.

**Locomotion.** Notostracans and anostracans swim with their trunk limbs, which beat in a rhythm so that jets of water are forced out sideways and backward from the spaces between the limbs to drive the animal ahead. Some anostracans, such as *Chirocephalus,* have a complex system of flaps and muscles in the trunk limbs, and they modify the limb movement in order to hover in one position for long periods. The other six orders swim by means of their antennae, which have two branches bearing featherlike setae that increase the effective area of the antenna. The orders Spinicaudata and Laevicaudata are slow, clumsy swimmers, and they are highly vulnerable to predation by fish; thus, they are most commonly found in temporary pools, where fish are absent. The anomopods, although smaller, are much livelier swimmers.

**Responses to light.** The most notable behavioral responses of branchiopods are in relation to light. The Anostraca are remarkable in showing a ventral light response: when light is directed from above, they turn their ventral surface toward the light. If they are artificially lit from below and not from above, they turn over. In the anomopods the response to light is complex and varies with the colour of the light. In red light, *Daphnia* maintains its position in the water by a hop-and-drop type of swimming. In blue light, it swims more rapidly in a horizontal direction. These two methods of swimming are related to the presence of food. When foods such as small green algae are present in the water, they absorb most of the blue light, and the light that penetrates is mainly red. Stationary swimming in response to this red light is advantageous to *Daphnia,* and it maintains its position. In the absence of food such as green algae, more blue light is present in the water. *Daphnia* is stimulated in response to this blue light to swim horizontally and to search a wider area. If *Daphnia* is starved and kept in red light, however, it eventually swims horizontally; *i.e.,* starvation blocks out the normal response to red light.

## FORM AND FUNCTION

**External features.** The fundamental structure of the Branchiopoda is related to their methods of feeding. In

General features

most species this involves a series of limbs acting together to filter, scrape, or otherwise gather food particles into a ventral food groove and transport them to the mouth. In the elongated forms, such as the anostracans, the segmentation of the trunk is simple and obvious, but in the short-bodied forms, such as the anomopods and onychopods, the trunk is much compressed and the segmentation is obscured. The exoskeleton of the branchiopods is generally thin and flexible, although in the notostracans it can be quite rigid in some parts. The crushing or biting parts of the mandibles are usually the thickest and strongest. The trunk limbs often have a complex intrinsic musculature, which enables the various parts of the limb to be moved relative to each other. Extrinsic muscles, having their origins within the trunk, operate at the bases of the limbs and are responsible for movements of the whole limb. The primitive branchiopod limb can be thought of as a multipurpose flap serving for locomotion, feeding, and respiration.

**Internal features.** *The circulatory system.* The heart of a branchiopod is often visible in the intact animal. In *Daphnia* (Figure 5) the heart is short and almost spherical, with two inlet holes, or ostia, and a single anterior opening. In the more elongated forms, such as the anostracans, the heart is longer, with a pair of ostia in each trunk segment except the first and last. In the notostracans the heart has 11 pairs of ostia, while the spinicaudates have four pairs and the laevicaudates three pairs. In all branchiopods the heart discharges blood into an open body cavity, or hemocoel, without any definite vessels. In spite of this open system the blood follows a fairly definite course around the body, a good proportion passing through the trunk limbs before returning to the heart.

Blood

The blood of branchiopods is unusual among crustaceans in containing the red respiratory pigment hemoglobin dissolved in the plasma. The concentration of hemoglobin in branchiopod blood varies inversely with the oxygen content of the surrounding water: when little oxygen is in the water, the blood contains a large quantity of hemoglobin and is bright red.

*The nervous system.* The nervous system of the branchiopods consists of a cerebral ganglion, or brain, connected to two chains of ventral ganglia, which run along the trunk, underneath the gut. Nerves develop from these ganglia to the various mouthparts and limbs. In the anostracans the two chains are cross-connected in each segment so that the system resembles a ladder. In the short-bodied forms, such as the anomopods and onychopods, the ventral nervous system is condensed into a single mass. The most conspicuous sense organs are the eyes. In the anostracans the eyes are on movable stalks, while in the notostracans the paired eyes lie close together on top of the head. In the other living branchiopods the eyes join together to form a single more or less spherical eye in the middle of the head. All branchiopod eyes are provided with muscles and show rapid trembling movements thought to be part of a scanning process that gives more information about the surroundings than could be gained with a stationary eye. Other sense organs in branchiopods are used mainly as organs of touch (mechanoreception) or taste (chemoreception). These sense organs take the form of bristles connected with nerves at their base, and those concerned with taste are often thin-walled and tubular in form. The notostracans in particular are richly endowed with both sorts of receptors on their trunk limbs; they help in sorting the edible from the inedible as the animal grubs about in the mud at the bottom of a pool.

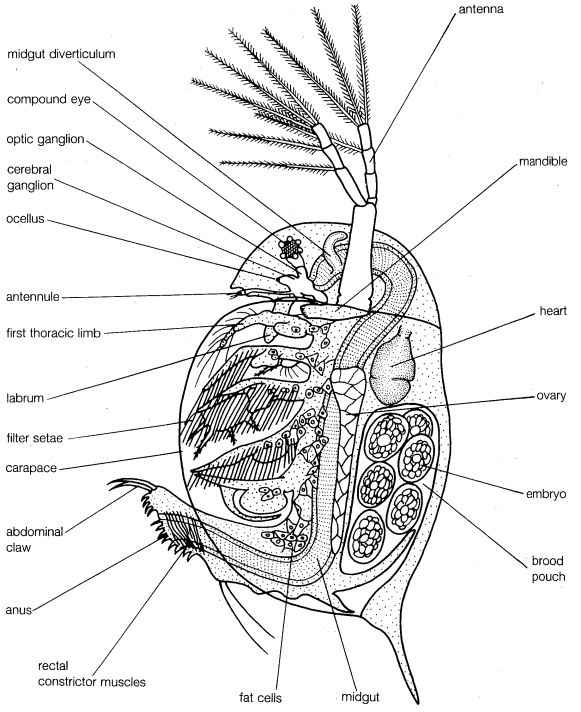



Figure 5: *Daphnia magna*, lateral view of adult female.

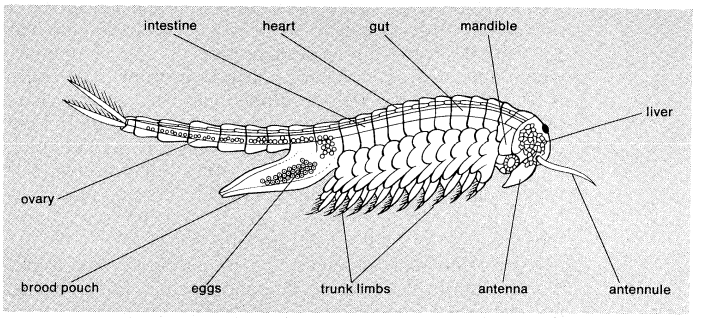



Figure 6: *Chirocephalus diaphanus*, an anostracan branchiopod.
From L. Borradaile and F. Potts, *The Invertebrata*; published by Cambridge University Press

*The digestive system.* The digestive system of the branchiopods shows considerable variation. In most groups the esophagus is narrow and has muscles which can dilate and others which can contract so that food can be pushed rapidly into the midgut. In many branchiopods the midgut is a simple tube with a pair of blind sacs, or diverticula. These diverticula may be simple extensions from the gut, or they may be complexly branched as in the notostracans and the spinicaudates. Some anomopods of the family Chydoridae have coiled midguts and may also have a single posterior diverticulum. One phenomenon shown by many branchiopods is anal swallowing. Water is taken in through the anus and is thought to act like an enema in clearing unwanted material from the hindgut.

The midgut

*The excretory system.* The excretory organ in the branchiopods is the maxillary, or shell, gland, so called because loops of the excretory duct can be seen in the wall of the carapace. In the nauplius larva the excretory function is performed by a gland opening on the antennae, but this degenerates as the animal grows and the maxillary gland takes over. Some excretion also can occur through the wall of the gut, which transfers substances from the blood into the gut lumen, from which it passes to the outside.

Most of the branchiopods have thin cuticles so that a certain amount of respiratory exchange can take place over the general body surface. The trunk limbs of most groups are flattened and leaflike, and on their outer edges they bear thin-walled lobes that can function like gills. The continuous movements of the trunk limbs of an anostracan, for instance, ensure a constant flow of water over these lobes. The lobes on the trunk limbs also play a part in ionic regulation, a process that controls the concentration and composition of the salts in the body fluids.

*Hormones.* There is good evidence of cyclic secretion of substances in the brain, which appears to be related to the control of molting and reproduction.

## EVOLUTION AND PALEONTOLOGY

The Branchiopoda originated in pre-Devonian times, for in the Devonian period two distinct orders are evident: the Lipostraca and the Spinicaudata. The Lipostraca contains only *Lepidocaris rhyniensis,* from the Rhynie cherts of

Scotland. This minute branchiopod is preserved so well that fine details of its limbs can be seen. Its structure is better known than that of any other fossil crustacean. It is even possible to deduce its method of feeding. The first three pairs of trunk limbs could have scraped material from the surfaces of plants or stones, and the food could then be transported forward to the mouth by a series of setae near the bases of the limbs. The trunk limbs lying behind the first three were two-branched and could have been used for swimming. Fossil members of the Spinicaudata are also known from the Devonian period, but their limb structure is not known in the detail available for *Lepidocaris;* many were preserved only as carapaces. The Laevicaudata extends back as far as the Early Cretaceous epoch (144 to 97.5 million years ago).

The Kazacharthra were much larger than *Lepidocaris* and occur later in the fossil record, being found in the Early Jurassic epoch (208 to 187 million years ago). They had elongated bodies with more than 40 body segments, a large carapace, and six pairs of complex flattened limbs.

At various times some of the fossils from the Burgess shales of the Cambrian period have been allocated to the Branchiopoda, but none of these has been generally accepted. Some fossils from the Cambrian period of Sweden, however, show features similar to those of primitive branchiopods, although the preservation is not sufficient to classify them with certainty. The earliest apparent anostracans are found in the Early Cretaceous epoch. They have trunk limbs very similar to those of recent anostracans. They also have stalked eyes and brood pouches.

Notostracan carapaces have been found in the Carboniferous period (360 to 286 million years ago), and the two extant genera, *Triops* and *Lepidurus,* are known from the Triassic period (245 to 208 million years ago). Some have actually been placed in the living species *Triops cancriformis,* indicating that this species has been in existence for more than 200 million years. The Anomopoda occur as fossils in recent deposits. The families Chydoridae and Bosminidae in particular have been used, in conjunction with pollen and diatoms, to interpret climatic and ecological changes during the histories of individual lakes. Older fossils of anomopods are rare, but egg cases, or ephippia, have been found from the Oligocene epoch (36.6 to 23.7 million years ago) and possibly from the Cretaceous period.

### CLASSIFICATION

**Distinguishing taxonomic features.** Branchiopods are free-living forms, the most primitive members of the phylum Crustacea. They have compound eyes and usually a protective plate, or carapace. There are many body segments and four or more pairs of trunk limbs that are usually lobed, broad, and fringed on the inner side. The mouthparts are small and simple, and the nervous system is primitive. Most species occur in fresh water.

**Annotated classification.** The groups indicated by a dagger (†) are extinct and known only from fossils.

#### CLASS BRANCHIOPODA

Distinguishing features include form of trunk limbs and carapace; 8 living and 2 fossil orders.

**†Order Lipostraca**

Known only from the Devonian; contains only the fossil *Lepidocaris rhyniensis;* 18 segments behind the head, plus telson-bearing caudal rami; no carapace; 13 pairs of trunk limbs in female; antennae large and branched, probably used in swimming; first maxillae small in the female but enlarged in the male as claspers; eggs give rise to nauplius larvae; about 3 mm long.

**Order Anostraca**

Elongated forms with paired compound eyes on stalks; no carapace; up to 27 body segments behind head, plus a telson with flattened caudal rami; usually 11, but up to 19, pairs of trunk limbs; eggs carried in a brood pouch behind the last pair of trunk limbs; antennae of female simple, but enlarged in the male to form claspers; worldwide; in fresh water, particularly temporary pools, and inland saline waters.

**Order Spinicaudata**

Large carapace in 2 parts encloses head and trunk; antennae large, branched, and used in swimming; 16 to 32 pairs of trunk limbs, flattened, leaflike, and used in filter feeding; male with first 2 pairs of trunk limbs modified for grasping female during mating; nauplius larvae, except *Cyclestheria;* fossils known from Devonian; recent forms worldwide, except polar regions; in fresh water, usually temporary pools.

**Order Laevicaudata**

Large bivalved carapace encloses the trunk but not the head; antennae large, branched, and used in swimming; first pair of trunk limbs of male modified for grasping the female during mating, other trunk limbs leaflike and used in filter feeding; nauplius larvae; fossils known from Cretaceous; recent forms worldwide in temporary fresh waters but not in polar regions.

**Order Ctenopoda**

Short-bodied forms with 6 pairs of trunk limbs, of which 5 bear filters; bivalved carapace encloses trunk but not head; antennae large, used in swimming, and bearing long swimming setae; all filter feeders; no larval stages, young hatch as miniatures of adult; worldwide in fresh water, except Antarctica; one genus, *Penilia,* is marine.

**Order Anomopoda**

Short-bodied forms with 5 or 6 pairs of trunk limbs; bivalved carapace encloses only the trunk; antennae large, branched, with up to 9 swimming setae; some filter feeders, some scrapers, and 1 genus, *Anchistropus,* parasitic on *Hydra;* no larval stages; resting eggs enclosed in a special case or ephippium; worldwide in fresh water.

**Order Onychopoda**

Short-bodied forms with carapace reduced to a dorsal brood pouch; 4 pairs of trunk limbs, which only grasp prey; single large median eye with many visual elements; antennae large, branched, with 12–15 swimming setae; freshwater and marine, with radiation into endemic species in the Caspian Sea.

**Order Haplopoda**

Contains only 1 genus, *Leptodora,* a plankton feeder; carapace reduced to a dorsal brood pouch; large antennae with more than 20 swimming setae; 6 pairs of grasping trunk limbs; head elongated with small, complex eye; transparent except for eye; young develop into miniatures of adult during summer, but overwintering resting eggs give rise to nauplius larvae; found in the fresh waters of the Holarctic region.

**Order Notostraca**

Large domed carapace covers part of trunk of up to 44 segments; telson with paired, threadlike rami; up to 71 pairs of trunk limbs; elongated first trunk limb functioning as tactile organ; antennae reduced or absent; paired, compound eyes without stalks; eggs, carried in pouches, hatch as nauplius larvae; worldwide except Antarctica; in fresh water, rarely in brackish water, most frequently in temporary pools.

**†Order Kazacharthra**

Early Jurassic; large carapace covers part of trunk; last 32–40 segments lack limbs; 6 pairs of large trunk limbs project beyond carapace; trunk ends in a large flat telson with a pair of long rami; overall length up to 10 cm.

**Critical appraisal.** Some authorities classify the Spinicaudata and Laevicaudata as suborders of a single order, the Conchostraca. Many would also group the Ctenopoda, Anomopoda, Onychopoda, and Haplopoda in another single order, the Cladocera. Some would go further and put the Cladocera and Conchostraca together as the Diplostraca. Other alternatives, which have not found general acceptance, are the inclusion of the Cephalocarida within the Branchiopoda, and the use of the groups Sarsostraca and Calmanostraca, the latter including all the orders except the Anostraca and Lipostraca. (J.Gre.)

## Cirripedes (barnacles)

### GENERAL FEATURES

**Diversity and distribution.** Barnacles and their allies, the parasitic orders Ascothoracica and Rhizocephala, are sedentary marine crustaceans. Barnacles usually have a calcareous shell made up of a number of articulated plates (Figure 7). The subclass Cirripedia is divided into two superorders, Acrothoracica and Thoracica. Members of the Acrothoracica are known as burrowing barnacles because they burrow into calcareous substrates (*e.g.,* limestone, corals, and mollusk shells). The acrothoracicans are recognized as fossils primarily by their burrows, and, while their record extends back into the Devonian period, they are particularly well represented in the Cretaceous period, when they burrowed into a greater variety of shell-bearing invertebrates than do their modern representatives.

The principal superorder, however, is the Thoracica. It comprises the orders Pedunculata and Sessilia and includes

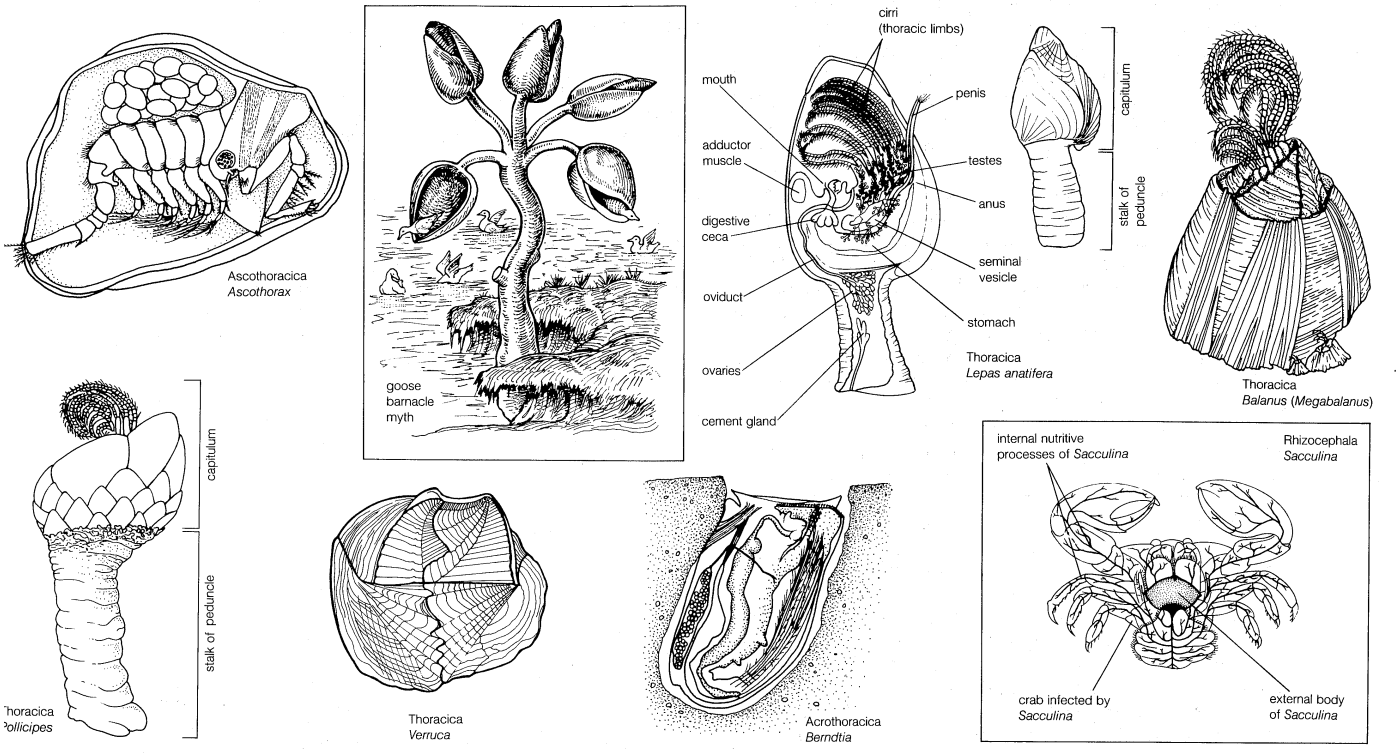



Figure 7: Representative cirripedes.

Modified after (*Ascothorax*) Y.K. Okada, *Les Cirripedes Ascothoraciques* (1938); from (goose barnacle myth) *Barnacles in Nature and Myth* (1928) by E. Heron-Allen, Oxford University Press, Inc., reprinted by permission; (*Lepas anatifera*) L.A. Borradaile and F.A. Potts, *Invertebrata,* 4th ed. (1967), Cambridge University Press; (*Verruca*) P.P.C. Hoek, *The Cirripedia of the Siboga-Expedition: Siboga-Expeditie XXXI;* (*Berndtia*) H. Utinomi, *Studies on the Cirripedia Acrothoracica* (1960), Seto Marine Biological Laboratory; (others) *Invertebrate Zoology* by Paul A. Meglitsch, copyright © 1967 by Oxford University Press, Inc., reprinted by permission

the goose and acorn barnacles. Thoracicans range from one millimetre (0.04 inch) to more than 10 centimetres (four inches) in diameter and from less than one to more than 500 grams (0.04 to 17.6 ounces).

Sym-metrical barnacles

The three suborders of the order Sessilia (sessile barnacles) contain two types of barnacles: symmetrical and asymmetrical. The two symmetrical sessile barnacles are the extinct suborder Brachylepadomorpha (*Brachylepas*) and the suborder Balanomorpha, or acorn barnacles (*e.g., Balanus, Semibalanus,* and *Chthamalus*). An acorn barnacle is a conical, sessile animal whose soft body is contained within a cavity protected by an outer wall. This wall comprises an even number of calcareous plates cemented to the substratum. An opening at the top can be closed by two pairs of plates (an operculum) through which feathery, jointed legs (cirri) can be extended into the water to capture small drifting plants and animals (plankton).

The balanomorphs are now the dominant shallow-water barnacles. Species are found in almost all habitats, from equatorial to polar regions, from estuarine waters and the highest intertidal zones to depths of 2,000 metres (6,560 feet) or more. Several groups of commensal balanomorphs have formed symbiotic associations with a variety of hosts, but only a few are known to have become fully parasitic. The most primitive living genera, such as *Chionelasmus* and *Catophragmus,* appeared in the late Mesozoic era (245 to 66.4 million years ago) and early Tertiary period (66.4 to 1.6 million years ago). Modern representatives are distributed throughout the world in refugial situations, such as abyssal hydrothermal springs.

The third suborder of sessile barnacles, the Verrucomorpha, or wart barnacles, differs from the first two suborders in having the plates of the wall and operculum asymmetrically arranged. With the exception of a primitive genus, *Neoverruca,* found to be associated with abyssal hydrothermal springs at 3,600 metres in the western Pacific, the simple, asymmetrical shell wall and operculum of verrucomorphs are remarkably similar. While the verrucomorphs apparently radiated in relatively shallow-water seas of the Cretaceous period, their modern representatives primarily inhabit the deep sea, where they range to depths of more than 4,000 metres.

Pedunculate barnacles

Members of the second thoracican order, Pedunculata, are similar to the sessile barnacles in having the principal part of the body contained within a protective covering, or wall. They differ from acorn barnacles in that the plates do not form a separate wall and operculum and in having the wall and the cirri it contains elevated above the substratum by a peduncle. The peduncle contains the ovaries and some musculature; it may or may not be armoured by calcareous plates, as in *Pollicipes* and *Lepas,* respectively. Goose barnacles are probably the most commonly observed pedunculate cirripedes.

Pedunculate barnacles occupy a wide variety of substrates. At least half the living species are symbiotic to some degree, and a few have become fully parasitic. In general, however, pedunculate barnacles have not formed as intricate symbiotic relationships as have a variety of balanomorphs. The pedunculate barnacles are fairly well represented in the fossil record, especially in the Mesozoic, but the earliest records date back to the Cambrian period (570 to 505 million years ago) and the Silurian period (438 to 408 million years ago).

Parasitic allies of barnacles

There are two parasitic groups: the order Rhizocephala, found primarily on decapod crustaceans, and the order Ascothoracica, found on echinoderms and cnidarian corals. While they are not included in the Cirripedia in the classification used in this article, they have been so included and therefore will be considered here.

**Importance to humans.** There are about a dozen important species of sessile and pedunculate barnacles that foul ships and submerged portions of marine installations, such as pier pilings, oil platforms, floats, buoys, and mooring cables. This causes drag and increased weight and may also increase corrosion of metals, even stainless steel. Antifouling paints contain toxins (usually heavy metals) and are designed to slough off with age. Low- and high-frequency sound waves may effectively inhibit settling. Ships and shore installations that circulate seawater for various purposes may have problems with lines clogged by fouling organisms. Various methods of prevention or removal, such as back-flushing with heated water or flushing with chemicals or fresh water, have been used with success.

Barnacles are used as food in some countries. In Por-

tugal and Spain a local intertidal pedunculate barnacle, *Pollicipes pollicipes,* is served in gourmet restaurants and occasionally becomes locally depleted. Two related species in the eastern Pacific, *P. polymerus* and *P. elegans,* from the northeastern and tropical eastern Pacific, respectively, are often imported as substitutes. Indians of the American Pacific Northwest consume the large sessile barnacle *Balanus nubilus,* and the inhabitants of Chile eat yet another large balanid species. In Japan barnacles are used as fertilizer.

The cement by which barnacles attach themselves to the substratum sets under water and even sticks to plastics with low surface tensions. It has been investigated because of its unusual properties and possible use in dental applications.

Goose barnacle myth

In the western British Isles during the Middle Ages a prevalent myth involving barnacles purported to explain the annual appearance of certain geese in the fall. Because these geese were arriving from their summer breeding grounds north of the Arctic Circle, they were not observed to breed locally. At the same time, fall gales often blew ashore driftwood fouled by the pedunculate barnacle *Lepas.* The barnacle myth correlated these occurrences; namely, according to the myth, the barnacles, which appeared to grow out of wood steeped in seawater, were actually developing geese, and, indeed, goose feathers (the barnacles' cirri) could be seen within. Further, since these geese were believed to have come from shellfish rather than flesh, they could be eaten on fasting days. The Swedish botanist Carolus Linnaeus was aware of the myth, for he named the genus *Lepas* ("Shellfish") and the local species *L. anatifera* and *L. anserifera* ("duck-bearing" and "goose-bearing," respectively), and these pedunculate barnacles continue to be called goose barnacles.

## NATURAL HISTORY

**Reproduction and life cycles.** Barnacles generally have both male and female reproductive systems in the same individual (hermaphroditism), either or both of which can be active at one time. Although some species are known to self-fertilize if no partners are present, most shallow-water species cross-fertilize. In species in which populations are sparsely distributed, a hermaphrodite may be accompanied by one or more small "complemental" males, or the larger individual may develop into a female whereby a smaller individual attaching to it becomes a "dwarf" male. When the male occupies a fairly exposed position on its partner, it resembles the juvenile and is capable of feeding. When, through coevolution, males have come to be protected by the partner in one way or another, the dwarf male is variously reduced, some to the extent of being little more than a sac containing the testes.

Dwarf males

Adjacent individuals in normally hermaphroditic populations do not simultaneously cross-fertilize; rather, they alternate male and female roles over time. The individual acting as a female lays eggs inside the mantle cavity shortly after molting (Figure 8). Secretions associated with egg laying include a pheromone to which adjacent individuals respond by extending the prosbosciform penis toward the source. Barnacles acting as males are able to inject spermatozoa into the mantle cavity of an individual as far as seven shell diameters away. Hundreds of eggs contained in this mantle cavity are fertilized at one time; usually several batches are laid each year by adults that may live as long as 30 years. The eggs undergo spiral cleavage, and the developing embryos are retained until the first larval stage, called the nauplius. In some species, however, the naupliar stages are passed in the egg, and a cyprid larva is released into the plankton.

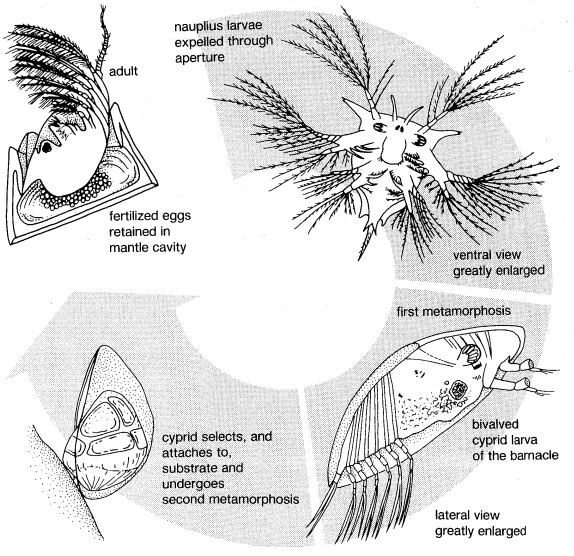



Figure 8: Life cycle of a typical acorn barnacle.

Larval forms

The nauplius larva of crustaceans (Figure 8) has three pairs of cephalic limbs, all of which aid in swimming while the second two are further involved in feeding. Cirripede and rhizocephalan nauplii differ from those of other crustaceans, including the Ascothoracica, in having conspicuous horns on either side near the front on the head. These horns have perforated tips and are provided with large secretory cells, but their function has yet to be determined.

Shortly after they hatch from the egg membranes and are expelled from the mantle cavity, the weak-swimming nauplii molt and begin to feed; the diet, primarily phytoplankton, depends upon the species. The nauplii continue to grow and molt for about two weeks, after which the sixth stage is reached. At this point a profound metamorphosis takes place, resulting in a nonfeeding, relatively strong-swimming cyprid larva (Figure 8). The cyprid must find a suitable surface upon which to settle in a relatively short time, or it will die. Substrate selection is based on light and chemical and tactile stimuli; it is likely that pheromones play a role when sexual selection is involved.

The cyprid swims with six pairs of thoracic limbs (the cirri of the adult). Gregarious forms are attracted by tactile stimuli to established members or to their remains, while commensal and parasitic species, many of which are host-specific, also use chemical stimuli to detect a suitable host. When ready to attach, the cyprid explores using its first antennae, the ends of which stick to the substratum by a temporary cement. When an appropriate place is found, similar glands secrete a permanent cement. The cyprid then undergoes metamorphosis into a juvenile barnacle.

Metamorphosis

Metamorphosis of a cyprid is complicated, some parts being temporarily or permanently lost, others modified and rotated, and still others appearing anew. The first juveniles of pedunculate barnacles are pedunculate, but pedunculate stages have been virtually eliminated from the development of modern sessile barnacles.

The rhizocephalans have an unusual life cycle. A cyprid destined to become a female seeks out a host, such as a crab, and attaches where the cuticle is thin, usually on a gill or at the base of a seta. The cyprid metamorphoses, and all body parts, except certain cells and organ rudiments of the head, are discarded. When this process is completed, a hollow, ventral stylet is, depending upon the species, forced either directly into the host or into the host after passing through one of the cyprid's first antennae. Once in the host's body, the cells and organ rudiments migrate into a central position beneath the gut, where they then send out rootlike absorption processes to all parts of the crab's body. The presence of the parasite not only castrates the host but it also feminizes a male host during subsequent molts both in morphology and behaviour.

Once the parasite has established itself internally, a hollow, mushroomlike reproductive body develops and perforates the ventral cuticle of the host between the thorax and the abdomen. There it enlarges to fill the space where the crab normally broods its eggs, and there the crab cares for the parasite as if it were its own eggs.

If a rhizocephalon cyprid destined to be a male finds the freshly erupted female, it attaches near the brood chamber and undergoes a similar metamorphosis into a minute cell mass surrounded by a thin cuticle. The cell mass migrates into the female's brood chamber, where it finds a special

pocket, or male-cell receptacle. It discards the cuticle as it enters the receptacle and differentiates into spermatozoa. Fertilization occurs when the eggs are laid, and the developing larvae are retained in the cavity until hatching. When the rhizocephalan is ready to release the larvae, the crab starts a ventilating motion with its abdomen, just as it would if it were releasing its own larvae, dispersing the parasite's larvae into the prevailing currents.

**Larval dispersal.** Larval dispersal depends upon the time spent and the behaviour of the various stages, as well as on favourable currents while in the plankton, prior to cyprid settlement. Larvae do not remain in the plankton for more than a few weeks, and larval dispersal is generally limited to less than 1,000 kilometres (620 miles). Species, however, are found on oceanic islands isolated by much greater distances, in part because some benthic barnacles occasionally attach to larger animals such as fish and whales as well as to floating objects such as wood, kelp, and pumice.

Associations

Still other barnacle species develop a symbiotic relationship with an organism, such as a whale, turtle, sea snake, or jellyfish (ectocommensal), and their distributions tend to approximate those of their hosts. In some instances, however, the distribution of the barnacle is only a small portion of that enjoyed by its host, indicating that other factors limit its range.

In the open ocean the larvae of the pedunculate barnacle *Lepas* seek out objects generally large enough to support the weight of the numerous adults (*e.g.,* driftwood). There is one species, however, that selects small objects (*e.g.,* feathers, bits of tar). After metamorphosis the cement glands of this species secrete a multichambered gas-filled float of its own. Floating objects attract other planktonic organisms, such as copepods and small fish, on which the barnacles feed.

## FORM AND FUNCTION

**External features.** It has been said that a barnacle is a shrimplike crustacean that attaches by the top of its head and then kicks food into its mouth with its feet. This likely tongue-in-cheek definition actually distinguishes barnacles from their allies and gives a fair idea of how the animal operates. Furthermore, a sedentary way of life requires protection from many biological and physical situations that can readily be avoided by their motile, free-living counterparts.

A thin, chitinous cuticle covers the appendage-bearing portion of the body, including the cirri, mouthparts, and lining of the mantle cavity. This portion of the exoskeleton is molted periodically, the process being controlled by hormones.

**Internal features.** *Tissues and musculature.* The tissues and organs of the barnacle are bathed by blood, which sometimes contains dissolved hemoglobin. In contrast to most crustaceans, however, the blood circulates in a generally closed system. Blood pressure extends and distends the stalk in pedunculate barnacles; the relatively long cirri, which are curled while at rest; the trunk of the body that supports the cirri; and the prosbosciform penis. The principle pair of plates covering over the mantle opening is provided with a transverse adductor muscle and discrete retractor muscles.

*The nervous system.* The nervous system, ladderlike in some primitive pedunculate barnacles, is condensed in scalpellomorphans and sessile barnacles into a single mass. The second antennae are present in nauplii but lost in cyprids. The first antennae are used by the cyprid in settling, but become buried in the permanent cement following attachment. The lateral, compound eyes of the cyprid are shed with metamorphosis. The nauplius median eye is generally retained in the adult as a photoreceptor. In sessile barnacles the bilateral parts of the median eye separate and migrate laterally to thin places under the anterior pair of opercular plates, where they function better in the shadow reflex. This reflex results in the rapid withdrawal of the cirri, which are otherwise vulnerable to predation, especially by fish.

*The digestive system.* Food gathered by the posterior cirri is collected and passed to the mouthparts by anterior cirri modified to act as maxillipeds. As it is pushed under an upper lip and into the mouth, the food is masticated by the two pairs of spiny maxillae and then by toothed mandibles. Salivary glands generally empty on the second maxillae and provide secretions that stick the food particles together and transport them to the mouth.

As in most crustaceans, a short foregut leads to a spacious midgut, or stomach, usually provided with a pair of digestive pouches, or ceca. The midgut is followed by a relatively long hindgut that passes the length of the trunk to the anus located between the last pair of cirri. The acrothoracicans have a gastric mill, derived from the chitinous lining of the foregut and renewed after being discarded during molting.

*The excretory system.* Maxillary glands are well-developed in adults, where they open just behind the base of the second pair of maxillae. Barnacles are also able to sequester heavy metals and brominated compounds as nodules in the wall of the midgut.

Open circulatory system

*The circulatory system.* As noted above, the circulatory system is a modification of the relatively simple, open system seen in crustaceans. In barnacles, pumping has largely been assumed by the somatic musculature, and in mature and advanced forms a nearly closed system of walled vessels has developed. Gas exchange can take place across any of the thin cuticular surfaces of the body, and many small barnacles lack discrete respiratory organs.

The cirri form excellent respiratory structures when the animals are feeding, and water can be circulated in and out of the mantle cavity by pumping movements of the body when the cirri are withdrawn and the aperture to the exterior is not completely closed. Some pedunculate barnacles have straplike organs, called filamentary appendages, extending from the body wall. On the other hand, most sessile barnacles have a pair of broad, often wrinkled extensions of the mantle wall, called branchiae. These and filamentary appendages are considered respiratory organs.

Ovaries and testes

*The reproductive system.* Ovaries are located in the stalk of pedunculate barnacles and its homologue or in the basal lining of the mantle cavity in sessile barnacles. Paired oviducts pass posteriorly to the bases of the first cirri (the most anterior position for genital openings in any crustacean), where each is joined by an oviductal gland before emptying into the mantle cavity. The oviductal gland secretes a tacky elastic substance that mixes with eggs as they are laid and holds them in one or two discrete masses called ovigerous lamellae.

Testes are situated in the trunk. Paired sperm ducts pass posteriorly below and to the sides of the gut before each expands into a seminal vesicle. An ejaculatory duct enters the base of the prosbosciform penis, situated between the last pair of legs, and runs its length. The penis may be clothed with fine setae, randomly distributed or arranged in discrete rows, or modified into simple or complex spines and hooks.

The nondistensible penis in ascothoracicans is a less-extensive modification of the seventh pair of trunk limbs seen in more primitive crustaceans. It is used to inject spermatozoa into special chitinous seminal receptacles in the bases of the trunk limbs of the female, where they are stored until the eggs are laid. The prosbosciform penis of acrothoracicans and ordinary barnacles injects the spermatozoa into the mantle cavity of the female, or a hermaphrodite acting as a female, at the time the eggs are being laid.

*Hormones.* Although sites of neurosecretory and glandular hormone production have not been identified in barnacles, molting and metamorphosis are controlled by hormones. Two insect molting hormones have been identified in the barnacle.

## EVOLUTION AND PALEONTOLOGY

The Cirripedia belong to the class Maxillopoda, an ancient radiation of relatively small, primarily marine crustaceans presently known by 11 subclasses, five of which (Branchiura, Facetotecta, Tantulocarida, Ascothoracica, and Rhizocephala) are wholly parasitic. Of the six nonparasitic subclasses the Ostenocarida and Skaracarida are

extinct (Cambrian), and the Mystacocarida are generally restricted to a narrow band of the marine interstitial environment. On the other hand, of the primarily nonparasitic groups, the Ostracoda and Cirripedia, ranging from the Cambrian period, and the Copepoda are diverse and occupy a wide variety of aquatic habitats. Of the nonparasitic groups, only the mystacocarids and cirripedes are exclusively marine.

Because of marked similarities in their nauplius and cyprid larval forms, it has generally been considered that the Cirripedia gave rise to the highly modified parasitic Rhizocephala. This view has been appealing because two parasitic pedunculate barnacles draw nutrients from their hosts by a root system, which, if not homologous with that of the Rhizocephala, at least indicates how the rhizocephalans could have evolved from a parasitic barnacle. The mode of host penetration by a stylet near the area of the cyprid mouth and the composition of the injected material, however, indicate that the Rhizocephala evolved from a biting rather than a filter-feeding ancestor and therefore more likely represent a sister group than a derivative of the Cirripedia.

At times the parasitic Ascothoracica also have been included in the Cirripedia because of a similar body plan. They have, however, a nonprobosciform median penis and seminal receptacles, as well as trunk limbs used solely for swimming, that show no indication of ever having been involved in filter feeding. Furthermore, their nauplius larvae lack frontolateral horns, and their cypridlike larvae (often more than one stage) not only are capable of feeding with biting mouthparts but also possess distinctive prehensile first antennae that lack cement glands. Thus, the Ascothoracica are no longer included in the Cirripedia. They share, however, a common nonparasitic ancestor with the parasitic Rhizocephala and Cirripedia, on one hand, and the Facetotecta, on the other. A category that had fallen into disuse, the Thecostraca, is therefore now being used to accommodate these subclasses.

A comparable nonparasitic ancestor was apparently shared by the Tantulocarida since a median nonprobosciform penis occurs on the same segment as in ascothoracids. While the posterior portion of the trunk in the Branchiura is too reduced to be instructive along these lines, the Ostracoda and the extinct Ostenocarida have a pair of penes and a pair of unmodified legs in the same position, respectively, and therefore are apparently nearer the stem of this Cambrian radiation. The Copepoda, Mystacocarida, and extinct Skaracarida, while lacking male genitalia and other features, such as a carapace and lateral eye, apparently also stem from near the base of this radiation in the Cambrian period.

Few Paleozoic barnacles are known. The acrothoracicans, or rather characteristic burrows made by them, appear in the Devonian period, but what they were like before they acquired the ability to burrow is unknown. The lightly armoured pedunculate barnacles *Priscansermarinus* and *Cyprilepas* appear earlier, in the Cambrian and Silurian periods, respectively, and what could pass for contemporary pedunculate barnacles, *Praelepas* and *Illilepas,* appear in the Carboniferous period. The first resembles heteralepadomorphs (genera without a trace of calcareous or primordial chitinous plates), and, other than being uncalcified, the last two resemble lepadomorphans *Lepas* and *Ibla,* respectively.

Heavy calcareous (calcitic) armament first appears in the Scalpellomorpha in the early Mesozoic era (245 to 66.4 million years ago), and by the close of the early Mesozoic the three sessile groups—Brachylepadomorpha, Verrucomorpha, and Balanomorpha—appear in order. The most primitive sessile group, the Brachylepadomorpha, died out by the Miocene epoch (23.7 to 5.3 million years ago), and the asymmetrical sessile Verrucomorpha became pretty much restricted to the deep sea by that time. The Balanomorpha radiated up through the Tertiary, and it is largely on the basis of their remains that Charles Darwin noted that the present epoch could go down in the fossil record as the age of barnacles. It is evident, however, that there was a greater diversity of shallow-water barnacles in the Miocene than there is today.

## CLASSIFICATION

### Annotated classification.

**SUBCLASS (ORDER) CIRRIPEDIA**

Maxillopodans; distinguished from nonparasitic crustaceans in being sedentary as adults; feeding by cirri; attached in burrows in limestone, corals, and shells or on a variety of substrata; usually provided with permanent, commonly calcareous armament; female genital apertures open on first trunk segment; male genital aperture opens on a probosciform median penis on the sixth trunk segment; parasitic isopods (malacostracans); differentiation of the carapace, skeletal armature, and appendages taxonomically significant; approximately 1,000 species known.

**Superorder Acrothoracica** (burrowing barnacles)

Devonian to present; globular in shape; generally without conspicuous calcareous exoskeleton; posterior cirri concentrated at end of trunk; widely distributed in coralline seas, most primitive members in deep sea; approximately 1 mm in length.

**Superorder Thoracica** (barnacles)

Cambrian to present; conspicuous calcareous exoskeleton; posterior cirri evenly distributed along trunk; inhabit virtually all marine environments, several primitive members in the deep sea; 1 mm to 2 cm in length, some larger.

*Order Pedunculata* (stalked or pedunculate barnacles)

Cambrian to present; body generally divided into capitulum and peduncle; capitular armament not differentiated into wall and operculum; includes 6 suborders, 2 extinct (Cyprilepadomorpha and Praelepadomorpha) and 4 extant (Heteralepadomorpha, Iblomorpha, Lepadomorpha, and Scalpellomorpha), the 3 best-known characterized below.

*Suborder Iblomorpha.* Carboniferous? to present; spindle-shaped capitulum/peduncle clothed with stiff chitinous bristles, supporting 2 pairs of sometimes slightly calcified plates; carapace adductor in primitive position, located posterior rather than anterior to esophagus; marine; 2 mm to 2 cm in length.

*Suborder Lepadomorpha* (goose barnacles and allies). Triassic?, Middle Eocene to present; capitulum typically with 5 calcareous plates; peduncle without calcareous armament; 3 mm to 75 cm in length.

*Suborder Scalpellomorpha* (leaf barnacles and allies). Triassic to present; capitulum supporting 6 or more calcareous plates; peduncle usually armoured with calcareous scales or spicules; 2 mm to 10 cm in length.

*Order Sessilia* (operculate or sessile barnacles)

Late Jurassic?, Cretaceous to present; capitulum relatively rigid; cemented directly to the substratum; supporting an operculum of 2 or 3 movable plates, or 2 to 3 pairs of movable plates; transient peduncle, disappearing early in ontogeny, forms the floor of capitulum in adults.

*Suborder Brachylepadomorpha.* Late Jurassic?, Triassic to Miocene; principal capitular plates and undedicated lateral plates; usually supporting 3 pairs of opercular plates; elevated above substratum by several basal whorls of imbricating plates; 1 to 2 cm in height.

*Suborder Verrucomorpha* (wart or asymmetrical sessile barnacles). Cretaceous to present; 2 or 3 plates on the right or left side form movable operculum; of those of the opposite side, 1 lost and 2 immovably incorporated into wall; primitive species with basal whorls of imbricating plates elevating the wall above the substratum; all basal imbricating plates lost in higher forms, the highest being reduced to a 2-plated operculum and a 4-plated wall cemented to the substratum; 2 mm to 2 cm in height.

*Suborder Balanomorpha* (acorn barnacles). Late Cretaceous to present; operculum of 2 pairs of plates supported by a rigid wall, including up to 3 pairs of dedicated lateral plates; principal wall cemented directly to the substratum, even when surrounded by basal whorls of imbricating plates; 2 mm to 10 cm in height and diameter, may reach 23 cm in height.

**Critical appraisal.** Some authorities, such as the author of the overview section above, rank the cirripedes as an order. The Rhizocephala, although closely related by larval similarities (including nauplii with frontolateral horns and a cyprid with prehensile first antennae provided with cement glands), are not included in the Cirripedia by some researchers. Also no longer included in the Cirripedia are the Ascothoracica, which have nauplii that lack horns, cypridlike larvae whose first antennae, while prehensile, lack cement glands, and males that have a nonprobosciform median penis.

Evidence from relatively recently discovered fossil and living taxa indicates that the Maxillopoda quite possibly include two natural lineages stemming from a common

ancestor in the Cambrian. If the definition of the Thecostraca, which presently includes the Facetotecta, Ascothoracica, Rhizocephala, and Cirripedia, were expanded to include all the maxillopodan subclasses also having a carapace and lateral eyes (Tantulocarida, Ostracoda, Branchiura, and Ostenocarida), and if the Copepodoida (originally equivalent to the Maxillopoda) were restricted to include the maxillopodan subclasses without a carapace and lateral eyes (Copepoda, Mystacocarida, and Skaracarida), it would likely better reflect the relationships within the class. (W.A.N.)

## Malacostracans (lobsters, shrimps, crabs, scuds, and pill bugs)

Malacostracans are the most numerous and most successful of the four major classes of Crustacea. Their members constitute more than two-thirds of all living crustacean species. They exhibit the greatest range of size (less than one millimetre, or 0.04 inch, to a limb spread of more than three metres, or 10 feet) and the greatest diversity of body form. Malacostracans are abundant in all permanent waters of the world: in the seas from the tropics to the poles and from the tidal zone to the abyss; in surface and subterranean fresh waters of all continents except Antarctica (where they once existed); and terrestrially on all continental landmasses and all tropical and temperate islands.

Successful radiation

The success of malacostracans can be attributed primarily to their increased body size and to the evolution of more functional body regions and more sophisticated food-gathering appendages than possessed either by their Paleozoic ancestors (570 to 245 million years ago) or by the next largest living crustacean class, the Maxillopoda. This evolutionary thrust has been marked by the development of ambulatory legs and specializations for benthic life and by the brooding of eggs and suppression of free-living larval development. Especially significant has been a shift of food-gathering limbs from head to thorax and of swimming appendages and respiratory organs from head to thorax and finally to the abdomen. This rearward shift freed the antennae for the development of specialized organelles sensitive to odours, sounds, vibrations, and physical contact and added more appendages (maxillipeds) to the mouthpart field. Such changes have enabled malacostracans to utilize efficiently the new food resources that have accompanied the evolution and proliferation of vascular plants from the late Paleozoic to the present.

### GENERAL FEATURES

**Size range and diversity of structure.** Malacostracans are typically large in size. Thus, some decapod crabs have leg spans of more than three metres, and others weigh more than 10 kilograms (22 pounds). Some free-living members of the orders Amphipoda, Isopoda, and Stomatopoda are lobster-sized (25–30 centimetres [0.8 to one inch]); most, however, are medium (one to three centimetres) in size. Paleozoic and primitive extant taxa seldom exceed 10 centimetres in body length, and the adult stages of some parasitic and subterranean groups are very small (less than one millimetre).

From J. Klövekorn, "Das Organsystem der Blutbewegung bei Gammarus," *Zeitschrift für Wissenschaftliche Zoologie*, Bd. 146 (1935)

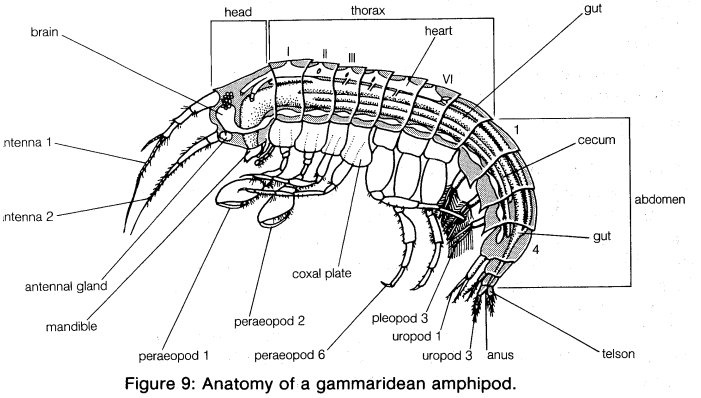



Figure 9: Anatomy of a gammaridean amphipod.

Malacostracans have a fixed body plan of head, thorax, and abdomen (Figure 9). In the adult the head consists of five segments, the thorax of eight, and the abdomen typically of six (or rarely seven) unfused segments. The head supports paired compound eyes, two pairs of antennae, and three pairs of short, chewing mouthparts, each consisting of two branches. The eyes are usually pigmented and borne on movable stalks, but they are sessile on the sides of the head in isopods, amphipods, and the superorder Hemicaridea. The first antennae (antennules) usually have two branches (three in the subclass Hoplocarida). The outer branch of the second antennae (antennal squame), which is usually flat and bladelike for elevation and swimming balance, has two segments in stomatopods and some mysids and one segment in syncarids and eucarids; it may be small or lost entirely in amphipods, isopods, and other bottom-dwelling or subterranean taxa. The first and second maxillae are short, with variable numbers of inner biting plates (endites) and often with outer lobes (epipodites), but the palps are short or lacking.

Carapace

From the hindmost (maxillary) segment projects a head shield, or carapace, which in primitive forms is large and covers the thorax, leg bases, and gill chamber. It may be fused to the dorsum of the thorax, as in the superorder Eucarida, but it is variously reduced and fused only to the anterior thoracic segments in the superorder Hemicaridea and the order Mysidacea or lost altogether in the orders Isopoda and Amphipoda and the superorder Syncarida.

The thoracic legs are typically biramous and eight in number. In free-swimming taxa the legs are more or less alike, and both branches are slender. In bottom-dwelling taxa the inner branch has become a stiff walking limb, and the slender multisegmented outer branch is variously reduced (in hemicarideans) or lost altogether (in amphipods and isopods). In advanced, especially bottom-dwelling, malacostracans, one or more legs are pincerlike.

Limbs

The abdomen bears on each but the last segment a pair of ventral, or ventrolateral, biramous limbs called paraeopods, or pleopods, which are primarily used in swimming. In the males of all eucaridans, hoplocarids, isopods, some hemicarids and syncarids, and rarely some amphipods, the anterior one or two pairs may be specially modified for sperm transfer. In males of most mysidaceans, the fourth and fifth pleopods (and the first and second uropods of some amphipods) may be modified as claspers for holding the female during copulation. The last abdominal segment (of all but the leptostracans) bears a pair of biramous uropods and a median plate, or telson. The uropods are usually setose and paddle-shaped in swimming taxa and form a broad tail fan with the telson for rapid propulsion. In benthic and subterranean taxa the uropods are often slender, elongate, and tactile in function. The telson is bilobed in juvenile syncarids, larval eucaridans, some mysids, and most amphipods but platelike in all other malacostracans.

**Distribution and abundance.** The class Malacostraca contains more than 22,000 living species and represents about two-thirds of all known crustaceans. It is the single largest group not only of marine arthropods but also of all fully aquatic arthropod taxa, including the insects and arachnids. Within the Malacostraca, Decapoda is the largest order, with about 9,000 described species, followed by the orders Amphipoda (6,200 species) and Isopoda (4,600 species). The other major orders have fewer than 1,000 species each.

Most malacostracans are marine. Among the decapods the ancient palinurans, their modern brachyuran (10-legged crab) derivatives, and the dendrobrachiate and stenopodid shrimps dominate in tropical and temperate marine shallows. The decapod caridean shrimps, astacidean lobsters and crayfish, and anomalans (hermits and eight-legged crabs), however, are dominant in cold-water and polar regions, in the deep sea, and in continental fresh waters. The amphipods and isopods are also abundant along cold-water marine shores and in the abyss and have widely penetrated fresh waters. They are also widespread in underground waters and terrestrial environments. Stomatopods are largely confined to tropical marine shallows; tanaids and cumaceans are found mainly in the colder deeps;

and mysids, though mainly marine, are also abundant in relicts of northern glacial lakes.

**Importance to ecology and to humans.** Malacostracans, being large crustaceans, are often predators and scavengers and thus are found at or near the top of the aquatic food pyramid. They are important ecologically in ridding the sea bottom and seashores of decaying animal and plant matter and in serving as middle-level converters of organic food energy to animal protein in a form suitable for fish, sea birds, marine mammals, and ultimately humans. The decapods and the order Euphausiacea (krill) are the only malacostracan groups of major economic value to humans.

## NATURAL HISTORY

**Reproduction and life cycles.** The life cycle of malacostracans typically involves an egg stage; a series of free-swimming, plankton-feeding larval stages; a series of immature (subadult) growth stages; and finally a sexually mature (reproductive) adult stage. Hermaphroditic adults are present in a few isopods. In the primitive swarming type of reproduction the male seeks out the female in the open water, usually in synchrony with lunar periodicity, cycles of temperature, or food availability. Mating (copulation) is very brief, often completed in a few seconds and usually following the reproductive molt of the female, when her exoskeleton is still soft. The eggs are fertilized as they are extruded from the oviductal opening on the sternum of the sixth thoracic segment. The anterior pleopods of the male are typically not modified for the transfer of sperm from the genital opening(s) on the eighth thoracic segment (*e.g.,* in amphipods and primitive decapods). Such males usually do not feed, do not reproduce again, and do not live long after mating. Fertilized eggs may be shed freely in the sea, where they hatch, usually into nauplius larvae. In marine groups that brood the eggs by attaching them to the pleopods, the eggs hatch as late-stage larvae, which are often carnivorous (*e.g.,* zoeae and phyllosoma larvae of decapods, antizoeae and pseudozoeae of stomatopods). These larvae eventually drop to the bottom and pass through one or more stages prior to attaining the immature stage. Where eggs develop within a thoracic brood pouch, the larval stages are suppressed. The eggs hatch as well-developed embryos (in the order Leptostraca) or as immature forms of the adult (*e.g.,* Isopoda, juveniles of the orders Mysidacea and Amphipoda), but parental brooding may be continued for a further few molts. In the deep sea and in fresh waters, whether eggs are laid freely (superorder Syncarida) or brooded on pleopods (decapods) or in a thoracic pouch (isopods and amphipods), the eggs hatch as juveniles or immature adult forms.

Mating

In the more-advanced, especially bottom-dwelling, malacostracans or in those with specialized habits, mating usually takes place on or in the bottom. Males may attend, guard, or carry the female for some time (preamplexus) prior to copulation (amplexus), and mating may be prolonged for several hours; the male usually continues to feed, molt, and mate further (in isopods, creeping decapods, and benthic amphipods). Where the female exoskeleton variously hardens prior to mating, the oviductal opening is often complex, and sperm transfer is assisted by correspondingly modified first and second pleopods of the male ("internal" fertilization of stomatopods, isopods, and the superorder Eucarida). Newly hatched late larvae or juveniles may be initially guarded or carried by the female (in stomatopods and some amphipods and isopods).

**Locomotion.** Malacostracans are primarily swimmers and secondarily walkers, clingers, and burrowers. Swimming is accomplished primitively by coordinated, synchronous beating of the setae on the outer branches of biramous head appendages in early larval stages and thoracic appendages in later larval stages and the adult stages of the order Leptostraca, the order Mysidacea, and the superorders Syncarida and Eucarida. The swimming action characteristic of adult malacostracans is provided by abdominal pleopods. Typically, each of the first five abdominal segments bears, on the ventral (lower) surface, a pair of pedunculate, biramous pleopods. In order to beat in unison, each pair is usually hooked together by spines on the inner margin of the peduncle (retinacula) or the inner ramus ("clothespin spines"). The amphipods are unique in having only three pairs of pleopods, the last two pairs being modified as stiff, thrusting uropods. In primitive forms the pleopod rami are slender and segmented (annulate), as in amphipods and procaridid decapods, all of which are primarily swimmers as adults; however, in all the other malacostracan groups, most of which are crawlers and burrowers, the rami are broad, flaplike, and unsegmented. The pleopods are typically reduced, or even lost, in many burrowers. The swimming crabs use paddle-like fifth thoracic legs for propulsion. Abrupt swimming propulsion is provided by the tail fan. In amphipods the tail fan (consisting of three pairs of uropods and telson) provides a sudden forward thrust. In eucaridans (especially decapods) the tail fan (paired uropods and telson) provides a characteristic "tail-flip" or sudden backward escape reaction.

Swimming

In most benthic malacostracans the hind five to seven pairs of thoracic legs have become essentially uniramous—the inner branch is thickened and stiffened and adapted for walking or crawling. In amphipods the first four pairs are pointed forward and the last three backward, an adaptation for perching, clinging, climbing in "inchworm" fashion, or jumping.

In burrowing malacostracans, especially decapods and stomatopods, the distal segments of some legs attain a semichelate, subchelate, or true chelate (pincerlike) form that facilitates both digging and removal of the soft substratum. In many amphipod burrowers the claws are often reduced, but the adjacent segments are much broadened, strongly spined, and powerfully muscled. Rapid leg movements, often aided by the fanning action of setose antennae and the hydraulic tunneling motion of powerful pleopods, enable these torpedo-shaped crustaceans to swim through loose sandy substrata, feeding as they go.

**Food and feeding.** Malacostracans consume virtually every available kind of organic matter, plant or animal, living or dead. The small- to medium-sized groups primarily consume detritus and plankton, and some are parasites of other aquatic organisms. The larger-sized groups are mainly carnivores and scavengers, preying on a wide range of small invertebrates and fishes or devouring the carcasses of whales, seals, fishes, and large invertebrates. Burrowing and small groundwater malacostracans are filter feeders, consuming microorganisms and bacteria from the sediments. Terrestrial isopods and amphipods consume forest leaf litter and algae at the tide lines.

Malacostracans capture or obtain their food primarily by using their thoracic legs. In early free-swimming larvae and the adults of some filter-feeding or deposit-feeding amphipods, isopods, and hemicarideans and in large carnivorous palinuran decapods, food may be gathered (occasionally killed) by means of the antennae and other head appendages. In carnivorous, or raptorial, species one or more of the thoracic legs are enlarged, and the tips are pincerlike, allowing the animal to capture, kill, and initially shred its prey (Figure 10). In lobsters and crayfish the first walking leg (fourth thoracic) is fully cheliform, and either the left or right claw is massive, with pavementlike teeth, for crushing hard-shelled prey such as snails and clams. In "spearer"-type stomatopods the raptorial claw is toothed and spiny for stabbing soft-bodied prey. "Smashers" have a swollen, hammerlike claw for crushing hard-bodied prey.

Capturing food

Malacostracans (except for leptostracans) typically have one to three pairs of thoracic limbs modified as accessory mouthparts. These maxillipeds (or "jaw legs") pass food to the masticatory, or chewing, mouthparts of the head proper. The thoracic segment of the first pair of maxillipeds is usually fused to the head, forming a cephalon (Figure 9). In stomatopods the first five pairs are called maxillipeds, but only the first pair is functionally so and its body segment is not fused to the head. In amphipods the first two pairs of thoracic legs may also function as food-pushing limbs, but their segments are typically free. In decapods the first two or three pairs serve as maxillipeds, and their segments are fused within the cephalothorax.

The mouthparts generally reflect feeding habits. In flesh eaters and scavengers the mandibular incisors are typically

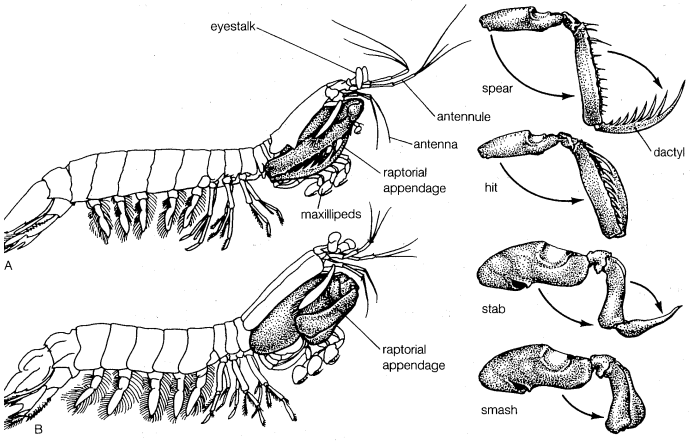



Figure 10: *Raptorial appendages.*
(A) Spearer of the species *Harpiosquilla harpax* has two modes of striking. (Top) In capturing prey, it uses a spearing strike; (bottom) in combat with another stomatopod, spearers usually hit with the dactyls closed. (B) Smasher of the species *Gonodactylus chiragra* uses its raptorial appendages both in combat with other stomatopods and in attacking prey. If the prey is soft-bodied, the smasher will change from a smashing strike to a stabbing one. It also sometimes stabs in intense combat.

Adapted from R.L. Caldwell and H. Dingle, "Stomatopods," copyright © 1975 Scientific American, Inc., all rights reserved

large and the plates and palps of the maxillae and maxillipeds are armed with strong spines and cutting edges, whereas the molar is small or lacking. In those species that consume all organic material and in those that consume only plants, the molar is usually strong, with an inner grinding surface. In filter feeders the plates of the maxillules, the maxillae, or both may be enlarged and equipped with a large number of fine-filtering (plumose) setae. Accessory (baler) plates, for directing feeding currents, are often well-developed (*e.g.,* in cumaceans and haustoriid amphipods).

Relationships

Although malacostracans are typically free-living, several taxa, especially among the amphipods, decapods, and isopods, have formed symbiotic, commensal, and even fully parasitic relationships with other invertebrates, fishes, marine mammals, and reptiles. Many decapods, especially porcellanid and xanthid crabs, live permanently in cavities among sponges, corals, and bryozoans. Leucothoid, sebid, and some lysianassid amphipods live within the respiratory and feeding cavities of sponges, tunicates, and anemones. Lafystiid and some lysianassid amphipods, as well as aegid, cymathoid, and immature gnathiid isopods, are external parasites of fish. Cyamid amphipods occur on whales and some hyalid amphipods in the buccal cavities of marine turtles. Epicaridean isopods are fully parasitic on other crustaceans, especially decapods. The body of the host may be much deformed and the body of the parasitic female very much transformed, quite unlike the small, symmetrically segmented, and otherwise normal male.

## FORM AND FUNCTION

**External features.** The chitinous exoskeleton, or cuticle, covering the body and limbs of malacostracans is divided into segments interconnected by strong, flexible membranes, allowing for articulation at the joints. The cuticle is usually soft and thin in small, wormlike, generally subterranean species, in parasitic species, or in the respiratory surfaces of free-living species, where gas exchange with the environment is vital. In large, heavy, mostly carnivorous forms, the cuticle is highly mineralized or impregnated with calcium salts. Such an exoskeleton provides considerable mechanical leverage and protection to the owner.

Malacostracans, like all arthropods, increase in size by molting. This process may take place very rapidly or require several days for completion (in some hard-shelled bottom dwellers). The animals remain in sheltered locations until the exoskeleton is hardened.

**Internal features.** *The nervous system and sensory organs.* The malacostracan central nervous system consists, in primitive forms, of a ventral nerve cord and ganglia within each body segment. The supraesophageal ganglion innervates the eyes, antennules, and antennae, and the subesophageal ganglion innervates the mouthparts of the head region. In amphipods and anomuran decapods the ganglia of abdominal segments are variously fused. In brachyuran decapods the abdominal and thoracic ganglia are fused into a single central thoracic ganglionic centre.

Eyes

Nearly all surface-dwelling members have pigmented eyes, but these are usually reduced or totally lost in underground and deep-sea species. The eyes of crustaceans are compound and may be composed of thousands of individual facets, or ommatidia. The compound eye of most malacostracans and their advanced larval stages is located on a movable stalk. The overall image is formed by combining the images from many individual ommatidia. Such eyes are called appositional, or mosaic. The compound eye is especially sensitive to movement and has a wide field of vision, often more than 180°, and is of enormous advantage to large, predatory malacostracans.

The eyes of smaller, mainly benthic, nonpredatory malacostracans, such as amphipods, hemicarideans, and isopods, are located on small lobes or flat on the sides of the head. Except in predatory or nocturnal amphipods, the eyes are small and consist of relatively few facets. Light that may strike a large patch of facets is concentrated on one ommatidium. Such eyes provide poor visual acuity. Compound eyes can discriminate colour, initiating changes within skin cells to match the colour of the substratum.

Olfactory hairs, or esthetascs, are used to locate food and recognize other crustaceans and their sexual states. Tactile setae occur generally over the external surfaces and appendages, especially of the antennae, food-gathering limbs, and mouthparts. Tactile hairs are present in the statocysts (organelles of balance) located, for example, in the first peduncular segment of the antennules in amphipods and the superorder Eucarida.

Some decapods and amphipods are sensitive to pressure change. Minute pit sensory organs of the general body surface are suspected receptors. Many decapods and amphipods produce sound by striking (percussion) or rasping (stridulating) or by internal mechanisms. Organs of sound reception include, in brachyurans, the chordotonal organs on the hinges of walking legs. Highly specialized sound and vibration receptors include the antennal calceoli of amphipods, the individual microstructure of which consists of receiving elements arranged serially and attached to the antennal segment by a slender stalk. In more-advanced groups the basal elements are expanded into a cuplike receptacle, and the stalk is distally expanded into a bulla, or resonator. The mechanism of transmission to the brain is unknown since nerve connections have not been discovered. In highly advanced predatory amphipods two types of calceoli are found: one type is used to detect mates (found in males only), and the other is used to detect prey (found in both sexes).

Digestive tract

*Digestion and nutrition.* The digestive tract of malacostracans consists of an esophagus; a two-chambered foregut; a midgut with outpocketings called digestive glands, or hepatopancreas; and a hindgut, or rectum. The large anterior foregut, or cardiac stomach, occupies much of the posterior aspect of the head and the anterior thoracic body cavity. A constriction separates it from the smaller, more ventral, pyloric stomach that lies in the posterior part of the thorax. Lining the inside of the greatly folded and muscular stomach walls, especially the pyloric portion, are groups or rows of stiff bristles, teeth, and filtering setae known as the gastric mill. The mill is strongly and complexly developed in large decapods, which ingest food quickly and in coarse chunks. The filtering setae are prominent in malacostracans that ingest fine materials or masticate their food throughly with the mouthparts. The macerated and partly digested food works its way through the filtering system of the pyloric stomach into the ceca, or pouches, of the hepatopancreas. There enzyme production and the storage and absorption of food takes place. The digestive secretions depend on the species and diet and include cellulase and chitinase. In stomatopods the

cardiac stomach is large enough to hold the remains of large prey; it opens directly from the mouth without an intervening esophagus. The midgut, or main intestine, may either extend throughout the abdomen, as in lobsters, or be very short, as in crabs. Fecal material is voided through the anus from the short rectum.

*Excretion.* Malacostracans excrete waste fluids mainly through the ducts of the nephridial glands, which are present in the body segments of the second antennae and the maxillae. The ducts open on the basal segments of those head appendages. Antennal nephridial glands are present in the adult stages of eucaridans, mysidaceans, and amphipods and in the larval stages of stomatopods and hemicarideans. The antennal glands of amphipods are enlarged in freshwater forms but are small in terrestrial species. Maxillary nephridial glands are typical of adult stomatopods, syncaridans, hemicaridans, and isopods. Adult leptostracans have both types of glands. Nephrocytes are present at the bases of thoracic legs and elsewhere in the body of mainly primitive groups. Bathynellaceans have a unique uropodal gland. The sternal gills of amphipods are osmoregulators.

Gills

*Respiration.* Most large malacostracans respire through gills, which develop as vascularized outgrowths of the first segment of the thoracic legs (epipodal gills). The gills of decapods are in a branchial chamber beneath the carapace, and oxygenated water is funneled through them. The lining of the chamber itself may be soft and vascularized for respiration, as in mysids, thermosbaenaceans, hemicarideans, and peneid shrimps. Land crabs have larger and more vascularized branchial chambers than do aquatic crabs.

The epipodal gills in syncarids and euphausiids are unprotected, since a carapace is either lacking or does not cover the leg bases. In amphipods the gills are usually simple sacs or plates, which in the course of evolution have migrated to the inner side of the legs. The gills are fanned and oxygenated by the pleopods in the ventral tunnel formed by the coxal plates. In stomatopods and isopods gill-like outgrowths of the pleopods or invaginated pseudotracheae (in terrestrial isopods) are the main organs of respiration.

Gas moves across the respiratory surface by diffusion rather than by active transport. Since the chitinous material of the body wall is relatively impermeable, special mechanisms have evolved to boost oxygen uptake. These include increased surface area (dendritic, foliate, pleated, or "double" gills), rich vascularization of respiratory surfaces, ventilating mechanisms (current-directing exopods and baler plates of the maxillae and maxillipeds), and presence in the blood of special respiratory pigments such as hemocyanin.

*Blood vascular system.* Malacostracans have a more complex open circulatory system than do other crustaceans. The single-chambered heart is enclosed in a pericardial sinus and is located dorsally, above the gut. It is elongate and tubular with several ostia for return flow in primitive forms (orders Leptostraca and Stomatopoda and the superorder Syncarida), but it is short and boxlike with one to two ostia and located in the thorax in advanced forms (decapods). The blood, or hemolymph, is pumped to the head through an aorta and to the gills and locomotor appendages through lateral and ventral arteries. Veins are lacking, and the blood returns to the heart via a series of sinuses.

*Endocrine system and hormones.* The major neuroendocrine control centre of malacostracans is the X-organ–sinus-gland complex, which lies in the eyestalk or in an equivalent part of the head in which the eyes are sessile. This complex regulates maturation, dispersal of pigments and colour change, and some metabolic processes. The female's ovaries, the male's reproductive glands, the pericardial organs, and the maxillary Y-organs of decapods also produce hormones that function in the molt and reproductive cycles.

*Defense and aggression.* Malacostracans must defend and fight for food, shelter, space, or mates. Hermit crabs fight over shells to occupy, stomatopods and alpheid shrimps fight over shelters, and terrestrial crabs and tube-building amphipods contest burrows and domiciles. Males may enlarge and embellish some of their appendages at maturity in order to enhance status and indicate sexual intent. Fights to determine status range from highly ritualized displays to death struggles. In decapods the most aggressive fighters are aquatic species, which are well armed, meet infrequently, and compete only occasionally over patchy, ephemeral resources. Terrestrial species, which are more prone to injury, more social, and less limited by availability of resources, exhibit more complex, formalized interactions. Male fiddler crabs attract females by waving the enlarged claw and sending sound signals. The signals establish the identity and intent of the sender. Male ghost crabs build sand pyramids to attract females. Numerous shrimps and some amphipods snap the movable finger of the enlarged claw against the hand as part of threat displays and courtship signals. Many stomatopods have a colour-coded, species-specific eyespot on the claws, which is displayed during posturing. More aggressive species have brighter eyespots. Stomatopods that fight with the same or closely related species reduce the force of their blows or engage in ritualized combat. Relatively docile species are more aggressive when facing more bellicose neighbours. An elaborate set of courtship signals is needed by the male stomatopod to prevent the female from attacking him.

Behaviours

## EVOLUTION AND PALEONTOLOGY

The fossil record of the Malacostraca extends from the early Paleozoic era (Early Ordovician epoch, 505 to 478 million years ago) to the present (Figure 11). The early phyllocarids (order Archaeostraca) had a body form which resembled the aquatic branchiocarid arthropods that were diverse in Cambrian seas, 570 to 505 million years ago. Those primitive forms (*e.g.,* Canadaspida) were not directly ancestral, however, since they lacked gnathobasic (chewing) head appendages (*e.g.,* mandibles, maxillae) and other major characteristics of the true Crustacea. Malacostracans share a number of advanced characteristics with the enigmatic class Remipedia, including biramous antennules, a first trunk segment fused to the head, limbs modified as maxillipeds, and paired swimming appendages on all trunk segments posterior to the genital openings.

The first eucaridan malacostracans appeared in the middle Paleozoic (Late Devonian epoch, 374 to 360 million years ago). These were burrowing, lobsterlike, protoglyphaeids with primitive, somewhat pincerlike walking legs and a tail fan with uropods. During the late Paleozoic (Early Carboniferous epoch through Permian period, 360 to 245 million years ago) malacostracans evolved rapidly, apparently in step with the proliferation of coastal vascular plants that formed a major new aquatic food resource. At least 16 new orders arose during that time, some members of moderate size, with both subcheliform and true pincerlike walking legs (*e.g.,* Hoplocarida, Astacidea). In other, mostly smaller, bottom dwellers in brackish to fresh lagoons and estuaries (*e.g.,* Hemicaridea, Syncarida, Mysidacea, Isopoda) the carapace and thoracic respiratory chamber were reduced or lost altogether, the eggs developed directly, within a thoracic brood pouch, and respiration and swimming propulsion became increasingly abdominal. At least eight primitive and unspecialized orders died out by the close of the Permian (*e.g.,* aeschronectid stomatopods, Pygocephalomorpha, Belotelsonidea). During the Mesozoic heyday of the malacostrans, 245 to 66 million years ago, however, an equal number of new orders arose. With the evolution of the anomurans and true crabs during this era, the decapods diversified and grew to large sizes. All major amphipod suborders and infraorders are believed to have evolved by the Jurassic and Cretaceous periods. The isopods had diversified into their nine existing suborders, including those fully parasitic on other crustaceans and fishes. All major continental fresh waters had been widely penetrated via estuaries and coastal groundwaters. Moist lands, then becoming forested with angiosperms, were being occupied by terrestrial isopods and amphipods.

With the subsequent cooling of coastal seas in the Tertiary period, several malacostracan groups (*e.g.,* asellote isopods, lysianassid amphipods, and anomuran decapods)

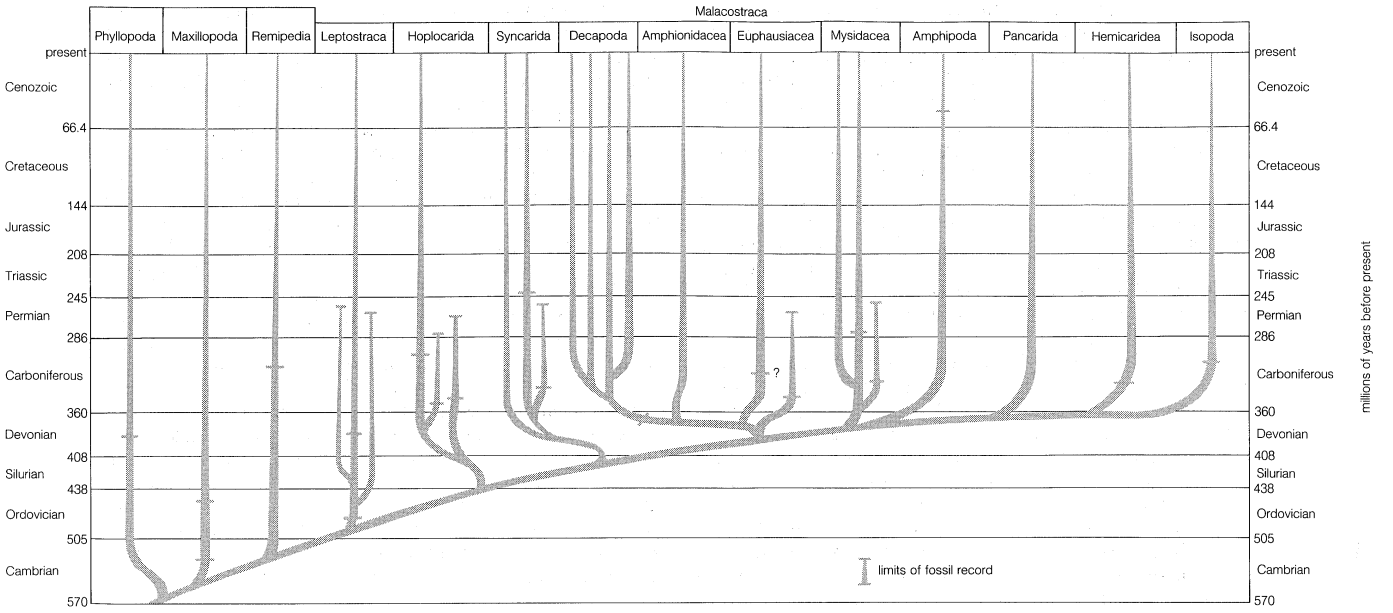



Figure 11: Phylogenetic tree of the Malacostraca.

proliferated in cold-water regions and in the deep sea. The amphipods became associated with mammals and tortoises, which first moved into their ancestral shallows, and coevolved with them to their specialized status as epiparasites of whales and marine turtles. Several malacostracan groups that had proliferated in the warm shallows of late Paleozoic and Mesozoic seas either disappeared or were reduced to a few relict species in deep or anoxic marine habitats (*e.g.,* Lophogastrida, glyphaeid decapods, Leptostraca, Mictacea) or in continental groundwaters (*e.g.,* Syncarida, Spelaeogriphacea, Thermosbaenacea). The isopods, decapods, and amphipods now make up 90 percent of all living malacostracans, with new species still in the process of evolution.

## CLASSIFICATION

**Diagnostic classification.** Malacostracan characters used in diagnosis and classification include type of eye (stalked or sessile), type of antennule (one-, two-, or three-branched, with or without sensory structures), type of antenna (with or without accessory branch, sensory structures), mouthpart structure (including presence or absence of palps, plates, and spines that reflect feeding preferences), carapace (presence or absence, type), anterior segments (degree of fusion with head), anterior limb pairs (degree of modification as maxillipeds, gnathopods), posterior limb pairs (whether single- or double-branched, simple or pincerlike, bearing gills or not), male and female sex ducts (type and position of openings), segmentation (degree of fusion of segments), pleopods (whether annulate or flaplike, sexually modified or not, gill-bearing or not), uropods (present or not, single- or double-branched), and telson (bilobate or platelike, with or without furcae).

Schram (1986) has revised much of the earlier classifications of Calman and others and is generally followed here. A dagger (†) indicates extinct groups.

**Annotated classification.**

### CLASS MALACOSTRACA

Double- or triple-branched antennules; single-branched ambulatory (walking) limbs often equipped with pincers; thoracic and abdominal respiration; terminal body segment with uropods; carapace, variously reduced or lacking, does not cover thoracic limbs; larval development usually of an advanced free-swimming type (*e.g.,* zoea) or often completed within the egg, in which case the first stage is an immature form of the adult; nauplius larva, when present, an advanced maxillopodan type lacking primitive frontal filaments but possessing specialized median eye.

**Subclass Phyllocarida**

Carapace large, appearing bivalved; thoracic legs with leaflike outer branch; abdomen 7-segmented, lacking uropods; anterior segments with pleopods (swimming legs).

†*Order Archaeostraca.* Early Ordovician to Permian; carapace bivalved and hinged; both antennae with two branches; 6 families.

†*Order Hoplostraca.* Late Carboniferous; carapace short; antennae 2 one-branched, raptorial in form; 1 family.

*Order Leptostraca* (visored shrimp). Permian to Holocene; carapace large, not hinged; antennae 2 one-branched, slender; terminal abdominal segment with pair of large paddlelike branches; eggs brooded under carapace; marine; on muddy bottoms low in oxygen; intertidal to the deeps; about 15 species in 3 families.

**Subclass Hoplocarida**

Late Devonian to Holocene; carapace large, not bivalved; rostrum hinged; antennules 3-branched; forward thoracic legs subchelate (clawlike); hind thoracic legs ambulatory (walking) or burrowing; abdomen large; pleopods bearing gills; terminal segment with large tail fan; 3 orders.

†*Order Aeschronectida.* Carboniferous; carapace covers entire thorax; thoracic legs without pincers; terminal body segment elongate; 3 families.

†*Order Palaeostomatopoda.* Late Devonian to Early Carboniferous; carapace covers thorax; anterior thoracic legs with claws; terminal body segment normal; 1 family.

*Order Stomatopoda* (mantis shrimps). Early Carboniferous, Mesozoic to Holocene; carapace short, exposing thoracic segments 5–8; first 5 pairs of legs clawlike, hind 3 stiltlike; terminal body segment normal; telson unbranched, simple; live in burrows or dens from which they dart forth to smash or spear prey with large clawlike second legs; mainly in tropical marine shallows; 4 superfamilies and 15 families.

**Subclass Eumalacostraca**

Late Devonian to Holocene; carapace (when present) not bivalved; rostrum fixed; first antenna 2-branched; thoracic legs with slender, many-segmented outer branch and stout, 7-segmented inner branch, often pincerlike, used in walking or food-gathering; 6 (rarely 7) abdominal segments, with pleopods and terminal uropods.

*Superorder Syncarida*

Late Devonian to Holocene; carapace lacking; thorax and abdomen weakly separated; thoracic legs biramous, bearing gills but without pincer claws; abdomen 6-segmented.

†*Order Palaeocaridacea.* Carboniferous to Permian; first thoracic segment not fused to head; abdominal pleopods 2-branched, flaplike; 4 families.

*Order Anaspidacea* (torrent shrimps). Triassic to Holocene; first thoracic segment fused to head; pleopods 1-branched, slender, multisegmented; fresh waters of Australia, New Zealand, and South America; 2 suborders and 4 families.

*Order Bathynellacea.* Permian to Holocene; body minute; wormlike; blind; thorax of 8 segments; legs short, weak; abdomen nearly lacking pleopods; in groundwaters of all continents except Antarctica; 2 families.

*Superorder Eucarida*

Carapace fused to thorax; thoracic legs usually with gills at bases; eggs usually hatch as free-swimming larvae.

†*Order Belotelsonidea.* Carboniferous; carapace large; thoracic legs 1-branched, simple, without pincers; pleopods flaplike; telson with furcae; 1 family.

*Order Euphausiacea* (krill). Carboniferous? to Holocene; carapace not covering leg bases; 8 thoracic legs biramous, unspecialized, bearing tuffy gills; telson with furcae; long series of larval stages; marine, pelagic; 2 families.

*Order Amphionidacea.* Holocene; carapace large; thoracic legs 1-branched; in female, first pleopod expanded under carapace to enclose a brood pouch; deep-swimming, tropical marine.

*Order Decapoda* (shrimps, lobsters, hermit crabs, crabs). Carapace large, enclosing thorax and gill chamber; inner branch of thoracic legs strong, often pincerlike; first 2–3 pairs smaller, modified as accessory feeding limbs (maxillipeds); uropods and telson usually forming broad tail fan; marine, some freshwater, a few terrestrial; about 9,000 species.

*Superorder Pancarida*

*Order Thermosbaenacea* (hot-springs shrimps). Holocene; minute; blind; wormlike; carapace short, forming dorsal brood pounch in female; thoracic legs weak, mostly 2-branched; lacking gills; pleopods weak or lacking; subterranean; about 15 species in 5 families.

*Superorder Peracarida*

Carapace shortened, attached anteriorly to thorax, or lacking; eggs develop in ventral thoracic brood pouch and hatch as miniature adults.

†*Order Pygocephalomorpha.* Carboniferous to Permian; carapace large, unridged, covering thorax; ventral plates of thorax widening behind; walking legs 6-segmented; abdomen 6-segmented; coastal marine; 4 families.

*Order Lophogastrida.* Late Carboniferous to Holocene; carapace large, ridged, covering thorax; ventral plates of thorax evenly widened; thoracic legs 7-segmented, weakly modified for grasping prey; abdomen basically 7-segmented; pleopods slender, branches segmented; deep-sea, free swimming; 3 families.

*Order Mysidacea.* Jurassic to Holocene; carapace short, exposing hind segments; thoracic legs simple, 7-segmented; abdomen 6-segmented; pleopods usually reduced in female, hind pairs modified as claspers in male; brood plates on posterior legs only; marine, freshwater; about 800 species in 4 families.

*Order Amphipoda* (well shrimps, night shrimps). Eocene to Holocene; carapace lacking; eyes flat on head, not stalked; 7 pairs of 1-branched thoracic legs, each covered basally by a coxal plate; last 5 or 6 pairs bearing gill on inner side; first 2 pairs usually subcheliform (pincerlike); abdomen 6-segmented, with 3 forward pairs of slender, segmented swimmerets and 3 hind pairs of stiff uropods; telson basically bilobed; thoracic brood pouch; eggs hatch as miniature adults; marine, freshwater, one family terrestrial; about 6,200 described species in 4 suborders, 31 superfamilies, and 137 families.

*Order Cumacea* (tadpole shrimps). Carboniferous to Holocene; head and thorax short, deep; carapace enclosing functional respiratory chamber; abdomen slender; pleopods lacking in female; marine, burrowing in sediments; about 800 species in 9 families.

*Order Mictacea.* Holocene; body elongate; carapace lacking; respiratory chamber vestigial; pleopods very reduced; marine; 2 families.

*Order Spelaeogriphacea.* Early Carboniferous to Holocene; body elongate; carapace short; thoracic legs slender, 2–4 with small outer branch; uropods broad; marine and freshwater; 2 families.

*Order Tanaidacea* (tanaid shrimps). Early Carboniferous to Holocene; body small, cylindrical; eyes on small lobes; carapace short; second thoracic legs large and pincerlike in male; 5 pairs of pleopods; marine, brackish, rarely freshwater; about 350 species in 4 suborders and 21 families.

*Order Isopoda* (pill bugs, sow bugs, sea slaters). Body flattened dorsoventrally or cylindrical (greatly modified in parasitic members); carapace and respiratory chamber lacking; eyes sessile; 7 pairs of uniramous thoracic legs (some may be pincerlike), lacking gills; leg segment 3 elongate; pleopods broad, often with gills; marine, freshwater, and terrestrial; about 4,500 species in 9 suborders and 100 families.

**Critical appraisal.** Schram (1986) has divided the Crustacea into four main classes: Remipedia, Malacostraca, Phyllopoda, and Maxillopoda. The Malacostraca are grouped with the last two (higher) crustacean classes in having a clearly demarcated head, thorax, and abdomen, the thorax usually with eight or fewer segments; it is distinct from them but more like the Remipedia in having paired appendages on all body segments.

The relationships of various groups within malacostracan subclasses are constantly undergoing revision in the light of new faunal discoveries and new taxonomic methodology. Schram has revised much of the earlier format (of Calman and others). He eliminated the Calmanian concept of "Peracarida," since it unites groups that are superficially similar rather than naturally closely related, a move with which the classification used here agrees. His format elevates to superorder the Mysidacea (containing orders Mysida, Lophogastrida, and Pygocephalomorpha), the Amphipoda, and the Isopoda and creates a new superorder Hemicaridea (containing orders Cumacea, Spelaeogriphacea, and Tanaidacea, to which this classification also adds the Mictacea). Thus, Schram's elevation of Hoplocarida to subclass level is accepted here, but not his reduction of superorder Syncarida to an order. The classification of the Amphipoda is updated from that of Schram. The mainly fossil subclass Phyllocarida (with two noncrustacean groups removed) is here retained in the Malacostraca (after Dahl). (E.L.B./K.E.C.)

**BIBLIOGRAPHY**

*Crustaceans:* A major reference on all aspects of the class is DOROTHY E. BLISS (ed.), *The Biology of Crustacea,* 10 vol. (1982–85). ALFRED KAESTNER, *Invertebrate Zoology,* vol. 3 (1970; originally published in German, 2nd ed., 1967), gives an excellent survey of morphology, physiology, embryology, and ecology. Other overviews are provided by the section "Crustacea" in SYBIL P. PARKER (ed.), *Synopsis and Classification of Living Organisms,* vol. 2 (1982), pp. 173–326; RAYMOND C. MOORE (ed.), *Treatise on Invertebrate Palaeontology,* pt. Q, *Arthropoda 3* (1961), and pt. R, *Arthropoda 4,* 2 vol. (1969); and ROBERT D. BARNES, *Invertebrate Zoology,* 5th ed. (1987). ROBERT H. GORE and KENNETH L. HECK (eds.), *Crustacean Biogeography* (1986), is an important discussion on distribution. PATSY A. MCLAUGHLIN, *Comparative Morphology of Recent Crustacea* (1980), gives clear diagrams of external and internal anatomy. FREDERICK R. SCHRAM, *Crustacea* (1986); and FREDERICK R. SCHRAM (ed.), *Crustacean Phylogeny* (1983), include recent theories of crustacean evolutionary relationships.

*Branchiopods:* H.G. CANNON and F.M.C. LEAK, "The Feeding Mechanism of the Branchiopoda," *Philosophical Transactions of the Royal Society of London,* Series B, 222:267–352 (1933), provides a classical, beautifully illustrated account. A.R. LONGHURST, "A Review of the Notostraca," *Bulletin of the British Museum (Natural History),* 3:1–57 (1955), is a revision of the systematics of the Notostraca and a standard work on the subject. See also G. FRYER, "A New Classification of the Branchiopod Crustacea," *Zoological Journal of the Linnean Society,* 91(4):357–383 (1987).

*Cirripedes:* CHARLES DARWIN, *A Monograph on the Subclass "Cirripedia,"* 2 vol. (1851–54, reissued as vol. 11–13 of *The Works of Charles Darwin,* 1988), is a comprehensive monograph on the Thoracica and still an important source. PAUL KRÜGER, *Cirripedia,* in *Bronns Klassen und Ordnungen des Teirreichs,* vol. 5, div. 1, fasc. 3, pt. 3 (1940), pp. 1–560, provides an authoritative compilation of knowledge. ALAN J. SOUTHWARD (ed.), *Barnacle Biology* (1987), includes chapters on evolution and genetics, physiology and function, larval biology and settlement, and pollution and fouling. H.G. STUBBINGS, *Balanus Balanoides* (1975), is a comprehensive monograph on the best-known barnacle in the world. Useful journal articles include D.T. ANDERSON, "The Larval Musculature of the Barnacle *Ibla quadrivalvis* Cuvier (Cirripedia, Lepadomorpha)," *Proceedings of the Royal Society of London,* Series B, 231(1264):313–338 (1987), an in-depth comparison of the larval musculature of one of the most primitive pedunculate barnacles with that of a rhizocephalan and balanomorphan; and W. KLEPAL, "*Ibla cumingi* (Crustacea Cirripedia): A Gonochoristic Species (Anatomy, Dwarfing, and Systematic Implications)," *Marine Ecology,* 6(1):47–119 (1985), a study of the gross and fine structure.

*Malacostracans:* In addition to the general books on crustaceans cited above, see E.L. BOUSFIELD, *Shallow-water Gammaridean Amphipoda of New England* (1973), describing and illustrating the range, ecology, life history, and behaviour of 125 species. (J.Gre./W.A.N./E.L.B./K.E.C.)

# Cryptology

Cryptology (from the Greek *kryptós,* "hidden," and *lógos,* "word") is the science of secure (generally secret) communications. This security obtains from legitimate users, the transmitter and the receiver, being able to transform information into a cipher by virtue of a key—*i.e.,* a piece of information known only to them. Although the cipher is inscrutable and often unforgeable to anyone without this secret key, the authorized receiver can either decrypt the cipher to recover the hidden information or verify that it was sent in all likelihood by someone possessing the key. Cryptography (from the Greek *kryptós* and *gráphein,* "to write") is the study of the principles and techniques by which information can be concealed in ciphers and later revealed by legitimate users employing the secret key, but in which it is either impossible or computationally infeasible for an unauthorized person to do so. Cryptanalysis (from the Greek *kryptós* and *analýein,* "to loosen" or "to untie") is the science (and art) of recovering information from ciphers without knowledge of the key. Cryptology is often—and mistakenly—considered a synonym for cryptography and occasionally for cryptanalysis, as in the popular solution of cryptograms or ciphers, but specialists in the field have for years adopted the convention that cryptology is the more inclusive term encompassing both cryptography and cryptanalysis.

Cryptography was concerned initially with providing secrecy for written messages. Its principles apply equally well, however, to securing data flow between computers, to digitized speech, and to encrypting facsimile and television signals. Most communications satellites, for example, routinely encrypt the data flow to and from ground stations to provide both privacy and security for their subscribers. Because of this broadened interpretation of cryptography, the field of cryptanalysis has also been enlarged to include the recovery of information from ciphers concealing any form of data.

This article discusses the basic elements of cryptology, delineating the principal systems and techniques of cryptography as well as the general types and procedures of cryptanalysis. It also provides a concise historical survey of the development of cryptosystems and cryptodevices. For additional information on the encoding and encryption of facsimile and television signals and of computer data, see TELECOMMUNICATIONS SYSTEMS and INFORMATION PROCESSING AND INFORMATION SYSTEMS.

The article is divided into the following sections:


General considerations 860
  The fundamentals of codes, ciphers, and authentication 860
  Applications of cryptology in private and commercial life 861
Cryptography 862
  Transposition ciphers 862
  Substitution ciphers 862
  Product ciphers 864
  The Data Encryption Standard 866
  Key distribution problem 866
  Two-key cryptography 867
  Block and stream ciphers 868
Cryptanalysis 869
  Basic aspects 869
  Types of cryptanalysis 869
History 870
  Early cryptographic systems and applications 870
  Developments during World Wars I and II 870
  The impact of modern electronics 872
Bibliography 873


## GENERAL CONSIDERATIONS

Because much of the terminology of cryptology dates to a time when written messages were the only things being secured, the source information, even if it is an apparently incomprehensible binary stream of 1s and 0s, as in computer output, is referred to as the plaintext. As noted above, the secret information known only to the legitimate communicants is the key and the transformation of the plaintext under the control of the key into cipher (also called ciphertext or, in older sources, cryptogram) is referred to as encryption or encipherment. The inverse operation, by which a legitimate receiver recovers the concealed information from the cipher using the key, is known as decryption or decipherment.

**The fundamentals of codes, ciphers, and authentication.** The most frequently confused, and misused, terms in the lexicon of cryptology are *code* and *cipher.* Even experts occasionally employ these terms as though they were synonymous.

A code is simply an unvarying rule for replacing a piece of information (*e.g.,* letter, word, or phrase) with another object, but not necessarily of the same sort. Probably the most widely known code in use today is the American Standard Code for Information Interchange (ASCII). Employed in all personal computers and terminals, it represents 128 characters (and operations such as back space and carriage return) in the form of seven-bit binary numbers—*i.e.,* as a string of seven 1s and 0s. In ASCII a lowercase *a* is always 1100001, an uppercase *A* always 1000001, and so on. In the first quarter of this century, before the telegraph was supplanted by radio communications, elaborate commercial codes that encoded complete phrases into single words (five-letter groups) were developed, so that telegraphers became conversant with such "words" as BYOXO (Are you trying to crawl out of it?), LIOUY (Why do you not answer my question[s]?), AYYLU (not clearly coded, repeat more clearly), or AZQUL (recheck coding and verify). Acronyms are also widely known and used codes, as, for example, RSVP and WASP. Occasionally a code word achieves an independent existence (and meaning) while the original equivalent phrase is forgotten or at least no longer has the precise meaning attributed to the code word—*e.g.,* SNAFU.

ASCII

Difference between ciphers and codes

Ciphers, as in the case of codes, also replace a piece of information (an element of the plaintext that may consist of a letter or word or string of symbols) with another object. The difference is that the replacement is made according to a rule defined by a secret key known only to the transmitter and legitimate receiver(s) in the expectation that an outsider, ignorant of the key, will not be able to invert the replacement to decrypt the cipher. In the past, the blurring of the distinction between codes and ciphers was relatively unimportant. In contemporary communications, however, information is frequently both encoded and encrypted so that it is important to understand the difference. A satellite communications link, for example, may encode information in ASCII characters if it is textual, or pulse-code modulate and digitize it in binary-coded decimal form if it is an analog signal such as speech. It then encrypts the resulting coded data into ciphers by using the Data Encryption Standard (DES; see below). Finally, the cipher stream itself is encoded again, using error-correcting codes for transmission from the ground station to the orbiting satellite. These operations are undone, in reverse order, by the intended receiver to recover the original information.

In the simplest possible example of a true cipher, A wishes to send one of two equally likely messages to B, say, to buy or sell a particular stock. The communication must take place over a party line on which eavesdroppers are listening. It is vital to A's and B's interests that others not be privy to the content of their communication. In order to foil the eavesdroppers, A and B agree in advance as to whether A will tell the truth or lie in what he says to B. Because this decision on their part needs to be unpredictable, they decide by flipping a coin. If heads comes up, A will say "buy" when he wants B to buy and "sell" when he wants B to sell. If tails comes up, however, he will say "buy" when he wants B to sell, and so forth.

| | | plaintext | |
|---|---|---|---|
| | | Buy | Sell |
| key | H | Buy | Sell |
| | T | Sell | Buy |

Using this encryption/decryption protocol, it should be obvious that the eavesdropper will know no more about the actual (concealed) instruction A sent to B as a result of listening to their telephone communication than he would if he did not listen at all. Such a cryptosystem is defined as *perfect*. The key in this simple example is the knowledge (shared by A and B) of whether A is telling the truth or not. Encryption in this case is the act by A of either lying or not lying as determined by the key, while decryption is the interpretation by B of what A actually meant, not necessarily of what he said.

This example can be extended to illustrate the second function of cryptography: providing a means for B to assure himself that an instruction actually came from A and not from someone impersonating A and that it correctly informs him of A's intentions—*i.e.,* a means of authenticating the message. If an eavesdropper, C, for cryptanalyst, pretends to be B and listens to A's instructions, he could—even though he cannot interpret the cipher to learn A's intentions—cause B to act contrary to A's intention by passing along to B the opposite of what A said. Similarly, C could simply impersonate A and tell B to buy or sell; however, he would not know which action B would take as a result, since he does not know whether A planned to lie or tell the truth. In either event, C would be certain of deceiving B into doing something that A had not requested.

To protect against this sort of deception by outsiders, A and B could use the following encryption/decryption protocol. They secretly flip a coin twice to choose one of four equally likely keys, labeled HH, HT, TH, and TT, with both of them knowing which key has been chosen:

| | | plaintext | |
|---|---|---|---|
| | | Buy | Sell |
| key | HH | Buy-Buy | Sell-Buy |
| | HT | Buy-Sell | Buy-Buy |
| | TH | Sell-Buy | Sell-Sell |
| | TT | Sell-Sell | Buy-Sell |

In this case the ciphers will consist of A's giving two instructions to B, such as Buy-Sell or Sell-Sell, which B can easily interpret to recover A's intended message because he knows the key (row) A used to encrypt the message. However, an outsider attempting to impersonate A will, with probability 1/2, choose a pair of messages that do not occur in the row corresponding to the key A and B are using; hence, the attempted deception will be detected by B, with probability 1/2, and the fraudulent message ignored. If C waits and eavesdrops when A communicates a cipher to B, no matter which cipher is used, he will be faced with a choice between two equally likely keys that A and B could be using and with a different cipher that he would have to substitute to have it be accepted by B in either case. Consequently, C's chances of deceiving B are still 1/2; namely, eavesdropping on A and B's conversation has not improved C's chances of deceiving B.

Clearly in either example, secrecy or authentication with secrecy, the same key cannot be reused. If C learned the cipher by eavesdropping and observed B's response to the cipher, he could deduce the key and thereafter impersonate A with certainty of success. If, however, A and B had chosen as many random keys as they had messages to exchange, the security of the information would have remained the same for all exchanges. When used in this manner, these examples illustrate the vital concept of a onetime key, which is the basis for the only cryptosystems that can be mathematically proved to be cryptosecure (see below). The hot line that links the White House and the Kremlin depends on a onetime key to achieve its cryptosecurity.

Onetime key

**Applications of cryptology in private and commercial life.** Although people may doubt that they have any personal involvement with cryptology, most adults depend on it to protect their interests or rights in several areas. For example, the personal identity number (PIN) that must be entered into an automated teller machine (ATM), along with a bankcard to corroborate that the card is being used by an authorized bearer, may either be stored in the bank's computers in an encrypted form (as a cipher) or be encrypted on the card itself. The transformation used in this type of cryptography is called one-way; *i.e.,* it is easy to compute a cipher given the bank's key and the customer's PIN but computationally infeasible to compute the plaintext PIN from the cipher, even though the key is known. This is to protect the cardholder from being impersonated by someone who has access to the bank's computer files. Similarly, the communications between the ATM and the bank's central computer are encrypted to prevent a would-be thief from tapping into the phone lines and recording the signals sent to the ATM to authorize the dispensing of cash in response to a legitimate user request and then later feeding the same signals to the ATM repeatedly to deceive it into dispensing money illegitimately.

Another example is the means used to prevent forgers from counterfeiting winning lottery tickets. Unlike currency, which typically involves state-of-the-art engraving as well as specially compounded inks and watermarked or tagged papers that make counterfeiting difficult, lottery tickets are simply printed on pasteboard much like the admission tickets used by movie theatres and hence are easily counterfeited if one knows what to print on the ticket. Each ticket, however, has two numbers printed on it—one being the identifying number that will be announced when a winner is selected and the other being an encrypted version of this number. Thus, when the winning number is made known, the would-be counterfeiter is unable to print an acceptable forgery unless he also has successfully cryptanalyzed the lottery's cryptosystem.

The two preceding examples involve only the use of the authentication feature of a cryptosystem, although secrecy is incidental to the communications between the ATM and the bank's central computer. A novel application that involves all aspects of cryptography are "smart" credit cards, which have a microprocessor built into the card itself. Smart credit cards first saw general use in France in 1984 and promise to supplant in large part the simple plastic cards currently being used. Cryptology is essential to the functioning of these cards in several ways. The user must corroborate his identity to the card each time a transaction is made in much the same way that a PIN is used with an ATM. The card and the card reader execute a sequence of encrypted sign-/countersign-like exchanges to verify that each is dealing with a legitimate counterpart. Once this has been established, the transaction itself is carried out in encrypted form to prevent anyone, including the cardholder or the merchant whose card reader is involved, from eavesdropping on the exchange and then later impersonating either party to defraud the system. This elaborate protocol is carried out in such a way that it is transparent to the user, except for the necessity of entering a PIN to initiate the transaction.

There are many other novel areas in which cryptography

plays a role in everyday life. In electronic mail, which has had a pilot test in the United States and was already in operation in France by 1984, the only way to provide an "envelope" for the messages is by some form of encryption. Increasingly, the data bases in which personal tax, income, credit-rating information, and other related data are compiled have become shared resources, remotely accessible to read from or to write into, so that cryptographic protection is vital for safeguarding the rights of the individual. Recognizing the threat to national security posed by breaches in computer-system security and attempts to eavesdrop on telecommunications, the U.S. government has taken steps to counteract the problem. Executive approval in 1984 of National Security Directive 145 led to the establishment of the National Telecommunications and Information Systems Security Committee, whose objective is to provide telephone and computer security for the federal government and its contractors. Clearly, information security—and this generally means cryptographically protected information—is one of the major problems faced by postindustrial society, and as such touches almost every aspect of private and commercial life.

Data security

## CRYPTOGRAPHY

Classification of cryptographic systems

Cryptographic systems are generically classified (1) by the mathematical operations through which the plaintext information is concealed using the encryption key—namely, transposition, substitution, or product ciphers in which two such operations are cascaded; (2) according to whether the transmitter and receiver use the same key (symmetric cryptosystem) or different keys (asymmetric [two-key or public-key] cryptosystem); and (3) by whether they produce block or stream ciphers.

The easiest way to describe the techniques on which cryptography depends is to first examine some simple cipher systems and abstract from these examples features that apply to more complex systems. There are two basic kinds of mathematical operations used in cipher systems: transpositions and substitutions. Transpositions rearrange the symbols in the plaintext without changing the symbols themselves. Substitutions replace plaintext elements (symbols, pairs of symbols, etc.) with other symbols or groups of symbols without changing the sequence in which they occur.

**Transposition ciphers.** In manual systems transpositions are generally carried out with the aid of an easily remembered mnemonic. For example, a popular schoolboy cipher is the "rail fence" in which the plaintext is staggered between rows and the rows are then read sequentially to give the cipher. In a depth two rail fence (two rows) the message WE ARE DISCOVERED SAVE YOURSELF becomes

W A E I C V R D A E O R E F
E R D S O E E S V Y U S L

or

WAEICVRDAEOREFERDSOEESVYUSL.

Simple frequency counts on the ciphertext would reveal to the cryptanalyst that letters occur with precisely the same frequency in the cipher and in the plaintext and, hence, that a simple transposition or rearrangement of the letters is involved.

Route ciphers

The rail fence is the simplest example of a class of transposition ciphers known as route ciphers, which enjoyed considerable popularity in the early history of cryptology. In general, the elements of the plaintext (usually single letters) are written in a prearranged order (route) into a geometric array (matrix) agreed upon in advance by the transmitter and receiver—typically a rectangle—and then read off by following another prescribed route through the matrix to produce the cipher. The depth two rail fence is a two-row by $n$-column matrix in which the plaintext is entered sequentially by columns; the encryption route is to read the top row first and then the lower:

W A E I C V . . .
E R D S O . . .

Clearly both the matrix and the route can be much more complex than those in this example. One form of transposition (permutation) that has been widely used depends on an easily remembered key word for identifying the order (route) in which the columns of a rectangular matrix are to be read. For example, using the key word AUTHOR and ordering the columns by the lexicographic order of the letters in the key word

| A | U | T | H | O | R |
|---|---|---|---|---|---|
| 1 | 6 | 5 | 2 | 3 | 4 |
| W | E | A | R | E | D |
| I | S | C | O | V | E |
| R | E | D | S | A | V |
| E | Y | O | U | R | S |
| E | L | F | A | B | C |

yields the cipher

WIREEROSUAEVARBDEVSCACDOFESEYL.

A significant improvement in cryptosecurity can be achieved by reencrypting the cipher obtained from one transposition with another transposition. Because the result (product) of two transpositions is also a transposition, the effect of multiple transpositions is to define a complex route in the matrix, which in itself would be difficult to describe by any simply remembered mnemonic device.

In decrypting a route cipher, the receiver enters the ciphertext symbols into the agreed-upon matrix according to the encryption route and then reads the plaintext according to the original order of entry. The matrix may take the form of a rectangle, trapezoid, hexagon, triangle, or other geometric figure; however, transposition systems in which the keys consist solely in keeping the matrices, starting points, and routes secret are not often employed because of limited security and because manual systems have largely been replaced by automated cipher systems. In the same class also fall systems that make use of perforated cardboard matrices called grilles; descriptions of such systems can be found in most of the older books on cryptography. In contemporary cryptography transpositions serve principally as one of several encryption steps in forming a compound or product cipher.

**Substitution ciphers.** In substitution ciphers, units of the plaintext (generally single letters or pairs of letters) are replaced with other symbols or groups of symbols, which need not be the same as those used in the plaintext. In Sir Arthur Conan Doyle's "Adventure of the Dancing Men," Sherlock Holmes solved a monoalphabetic substitution cipher in which the ciphertext symbols were stick figures of a human in various dancelike poses.

Caesar cipher

The simplest of all substitution ciphers are those in which the cipher alphabet is merely a cyclical shift of the plaintext alphabet. Of these, the best known is the Caesar cipher, used by Julius Caesar, in which A is encrypted as D, B as E, and so forth. As many a schoolchild has discovered to his sorrow, cyclical-shift substitution ciphers are not secure. As is pointed out in the section on cryptanalysis (see below), neither is any other monoalphabetic substitution cipher in which a given plaintext symbol is always encrypted into the same ciphertext symbol. Because of the redundancy of the English language, only about 25 symbols of ciphertext suffice to permit the cryptanalysis of monoalphabetic substitution ciphers. The explanation for this weakness is that the frequency distributions of symbols in the plaintext and in the ciphertext are identical, only the symbols having been relabeled. In fact, any given structure or pattern in the plaintext is always preserved intact in the ciphertext, so that the cryptanalyst's task is an easy one.

There are two main approaches that have been employed with substitution ciphers to lessen the extent to which structure in the plaintext—primarily single-letter frequencies—survives in the ciphertext. One approach, which culminates in the algebraic cryptosystems of Lester S. Hill, is to encrypt elements of plaintext consisting of two or more symbols; *e.g.*, digraphs and trigraphs. The other is to use several cipher alphabets. When this approach of polyalphabetic substitution is carried to its limit, it results in

onetime keys, or pads, which are the only cryptosystems that can be proved to be cryptosecure.

In manual cryptosystems for encrypting units of plaintext made up of more than a single letter, only digraphs were ever used. By treating digraphs in the plaintext as units rather than as single letters, the extent to which the raw frequency distribution survives the encryption process can be lessened but not eliminated, as letter pairs are themselves highly correlated. The best known digraph substitution cipher is the Playfair invented by Sir Charles Wheatstone but championed at the British Foreign Office by Lyon Playfair, the first Baron Playfair of St. Andrews, whose name the cipher bears. Below is an example of a Playfair cipher, solved by Lord Peter Wimsey in Dorothy L. Sayers' *Have His Carcase.* Here, the mnemonic aid used to carry out the encryption is a $5 \times 5$-square matrix containing the letters of the alphabet (I and J are treated as the same letter). A key word, MONARCHY in this example, is filled in first, and the remaining unused letters of the alphabet are entered in their lexicographic order:

Playfair cipher

| | | | | |
|---|---|---|---|---|
| M | O | N | A | R |
| C | H | Y | B | D |
| E | F | G | I/J | K |
| L | P | Q | S | T |
| U | V | W | X | Z |

Plaintext digraphs are encrypted with the matrix by first locating the two plaintext letters in the matrix. They are (1) in different rows and columns; (2) in the same row; (3) in the same column; or (4) alike. The corresponding encryption (replacement) rules are the following:

1. If the pair of letters are in different rows and columns, each is replaced by the letter that is in the same row but in the other column; *i.e.,* to encrypt WE, W is replaced by U and E by G.
2. A and R are in the same row. A is encrypted as R and R (reading the row cyclically) as M.
3. I and S are in the same column. I is encrypted as S and S as X.
4. If a double letter occurs, a spurious symbol, say Q, is introduced so that the MM in SUMMER would encrypt into NL for MQ and CL for ME.
5. An X is appended to the end of the plaintext if necessary to cause the plaintext to have an even number of letters.

Encrypting the familiar plaintext example using Sayer's Playfair array yields:

Plaintext: WE ARE DISCOVERED SAVE YOURSELFX
Cipher: UG RMK CSXHMUFMKB TOXG CMVATLUIV

Figure 1 plots the frequency distribution for the roughly 70,300 alphabetic characters in the present article and, hence, the frequency distribution for any simple substitution cipher of this text and the frequency distribution that results when the text is encrypted using the Playfair key of the example above. If the frequency distribution information were totally concealed in the encryption process, the ciphertext plot of letter frequencies would be flat. The deviation from this ideal is a measure of the tendency of some letter pairs to occur more frequently than others and of the Playfair's row and column correlation of symbols in the ciphertext—the essential structure exploited by a cryptanalyst in solving a Playfair cipher. The loss of a significant part of the plaintext frequency distribution, however, makes a Playfair cipher much harder to cryptanalyze than a monoalphabetic cipher.

The other approach to concealing plaintext structure in the ciphertext involves using several different monoalphabetic substitution ciphers rather than just one; the key specifies which particular substitution is to be employed for encrypting each plaintext symbol. The resulting ciphers, known generically as polyalphabetics, have a long

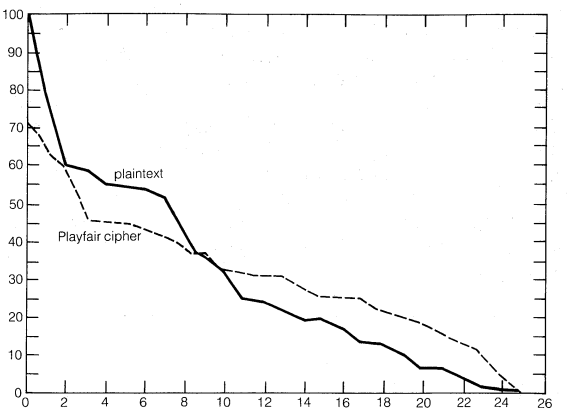



Figure 1: Relative frequency of occurrence of letters in this section and hence in any simple substitution cipher of the text and in a Playfair cipher.

history of usage. The systems differ mainly in the way in which the key is used to choose among the collection of monoalphabetic substitution rules.

The best known polyalphabetics are the simple Vigenère ciphers, named for the 16th-century French cryptographer Blaise de Vigenère. For many years this type of cipher was thought to be impregnable and was known as *le chiffre indéchiffrable.* The procedure for encrypting and decrypting Vigenère ciphers is illustrated in Figure 2.

Vigenère cipher

In the simplest systems of the Vigenère type, the key is a word or phrase that is repeated as many times as required to encipher a message. If the key is DECEPTIVE and the message is WE ARE DISCOVERED SAVE YOURSELF, then the resulting cipher would be

Message: WE ARE DISCOVERED SAVE YOURSELF
Key: DE CEP TIVEDECEPT IVED ECEPTIVE
Cipher: ZI CVT WQNGRZGVTW AVZH CQYGLMGJ.

The cryptanalysis, by Friedrich W. Kasiski, of repeated-key Vigenère ciphers is based on the fact that identical pairings of message and key symbols generate the same cipher symbols. Cryptanalysts look for precisely such repetitions. In the example given above, the group VTW appears twice, separated by six letters, suggesting that the key (*i.e.,* word) length is either three or nine. Consequently, the cryptanalyst would separate the cipher symbols into three and nine monoalphabets and attempt to solve each




Figure 2: *The Vigenère table.* In encrypting plaintext, the cipher letter is found at the intersection of the column headed by the plaintext letter and the row indexed by the key letter. To decrypt ciphertext, the plaintext letter is found at the head of the column determined by the intersection of the diagonal containing the cipher letter and the row containing the key letter.

of these as a simple substitution cipher. With sufficient ciphertext, it would be easy to solve for the unknown key word. Figure 3 shows the extent to which the raw frequency of occurrence pattern is obscured by encrypting the text of this article using the repeating key DECEPTIVE.

The periodicity of a repeating key exploited by Kasiski can be eliminated by means of the running-key Vigenère cipher. Such a cipher is produced when a nonrepeating text is used for the key. Vigenère actually proposed concatenating the plaintext itself to follow a secret key word in order to provide a running key in what is known as an autokey.

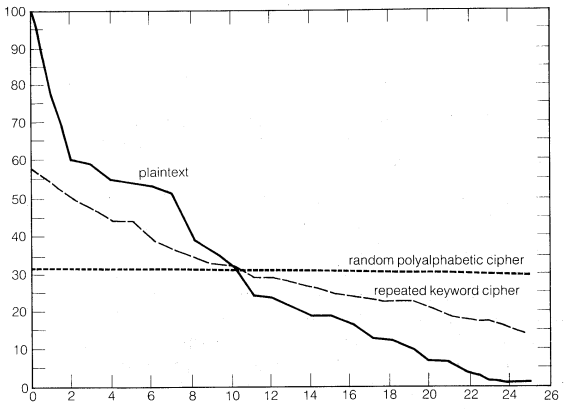



Figure 3: Relative frequency of occurrence of letters in a repeated-key Vigenère cipher—the key word is DECEPTIVE—and in a random polyalphabetic cipher.

Even though running-key or autokey ciphers eliminate periodicity, it is still possible to cryptanalyze them by means of two methods. In one, the cryptanalyst proceeds under the assumption that both the ciphertext and the key share the same frequency distribution of symbols and applies statistical analysis. For example, E enciphered by E would occur in English text with a frequency of ~ 0.0169, and T by T would occur only half as often. The cryptanalyst would, of course, need a much larger segment of ciphertext to solve a running-key Vigenère cipher, but the basic principle is essentially the same as before—*i.e.,* the recurrence of like events yields identical effects in the ciphertext. The second method of solving running-key ciphers is commonly known as the probable-word method. In this approach, words that are thought most likely to occur in the text are subtracted from the cipher. This process is continued until fragments of meaningful key have been obtained. These fragments can then be expanded with either of the two techniques described above. If provided with enough ciphertext, the cryptanalyst can ultimately decrypt the cipher. What is important to bear in mind here is that the redundancy of the English language is high enough that the amount of information conveyed by every ciphertext component is greater than the rate at which equivocation (*i.e.,* the uncertainty about the plaintext that the cryptanalyst must resolve to cryptanalyze the cipher) is introduced by the running key.

In 1918 Gilbert S. Vernam, an engineer for the American Telephone & Telegraph Company (AT&T), introduced the most important of key variants to the Vigenère system. At that time all messages transmitted over AT&T's teleprinter system were encoded in the Baudot Code, a binary code in which a combination of marks and spaces represents a letter, number, or other symbol. Vernam suggested a means of introducing equivocation at the same rate at which it was reduced by redundancy among symbols of the message, thereby safeguarding communications against cryptanalytic attack. He saw that periodicity (as well as frequency information and intersymbol correlation), on which earlier methods of decryption of different Vigenère systems had relied, could be eliminated if a random series of marks and spaces were added to the message being transmitted using a rule that the sum of a pair of like symbols was a mark and that the sum of a pair of unlike symbols was a space (logical sum). The validity of Vernam's idea was confirmed some 30 years later by another AT&T researcher, Claude Shannon, the father of modern information theory.

There was one serious flaw in Vernam's system, however. The system required one symbol of key for each message symbol, which meant that communicants would have to exchange impractically large amounts of the key in advance. The key itself consisted of a punched paper tape that could be read automatically while symbols were typed at the teletypewriter keyboard and encrypted for transmission. This operation was performed in reverse using a copy of the paper tape at the receiving teletypewriter to decrypt the cipher. Vernam initially believed that a short random key could safely be reused many times, but this resulted in a periodicity of the key vulnerable to attack by methods of the type devised by Kasiski. Vernam offered an alternative solution: a key generated by forming the logical sum of two shorter key tapes of $m$ and $n$ binary digits, or bits, where $m$ and $n$ share no common factor other than 1. A bit stream so computed does not repeat until $mn$ bits of key have been produced. This version of the Vernam system was adopted and employed by the U.S. Army until Maj. Joseph O. Mauborgne of the Army Signal Corps demonstrated that a cipher constructed from a key produced by linearly combining two or more short tapes could be decrypted by methods of the sort employed to cryptanalyze running-key ciphers. Mauborgne's work led to the realization that neither the repeating single-key nor the two-tape Vernam–Vigenère system was cryptosecure. Of far greater consequence to modern cryptology—in fact, an idea that remains its cornerstone—was the conclusion drawn by Mauborgne and William F. Friedman that the only type of cryptosystem that is unconditionally secure is a random onetime key. In such a stream cipher, the key is incoherent—*i.e.,* the uncertainty that the cryptanalyst has about each successive key symbol must be no less than the average information content of a message symbol. The dotted curve in Figure 3 indicates that the raw frequency of occurrence pattern is lost when the draft text of this article is encrypted with a random onetime key. The same would be true if digraph or trigraph frequencies were plotted for a sufficiently long ciphertext. In other words, the system is unconditionally secure, not because of any failure on the part of the cryptanalyst to find the right cryptanalytic technique but rather because he is faced with an irresolvable number of choices for the key or plaintext message.

Random onetime key

**Product ciphers.** In the discussion of transposition ciphers it was pointed out that by cascading (forming the product of) two or more simple transpositions, one could realize an equivalent transposition much harder to specify than the factor transpositions. In the days of manual cryptography this was a useful device for the cryptographer and, in fact, double transposition ciphers on keyword-based rectangular matrices were widely used. There was also some use of a class of product ciphers known as fractionation systems, wherein a substitution was first made from symbols in the plaintext to multiple symbols (usually pairs, in which case the cipher is called a biliteral cipher) in the ciphertext, which was then superencrypted by a transposition. One of the most famous field ciphers of all time was a fractionation system, the ADFGVX cipher employed by the German Army during World War I. This system used a $6 \times 6$ matrix to substitution-encrypt the 26 letters and 10 digits into pairs of the symbols A, D, F, G, V, and X. The resulting biliteral cipher was then written into a rectangular array and route encrypted by reading the columns in the order indicated by a key word as illustrated in Figure 4.

Fractionation systems

In spite of the great French cryptanalyst Georges J. Painvin's success in cryptanalyzing critical ADFGVX ciphers in 1918, with devastating effect for the German army in the battle for Paris, the U.S. cryptanalyst Stephen M. Matyas presented, as recently as 1984, new research on the cryptanalysis of ADFGVX ciphers.

Of much greater significance in the history of cryptology is a class of cipher machines generically known as rotor systems. Pneumatic and optical systems have been built, but only the electrical variety has been of value. Typically,

Rotor systems

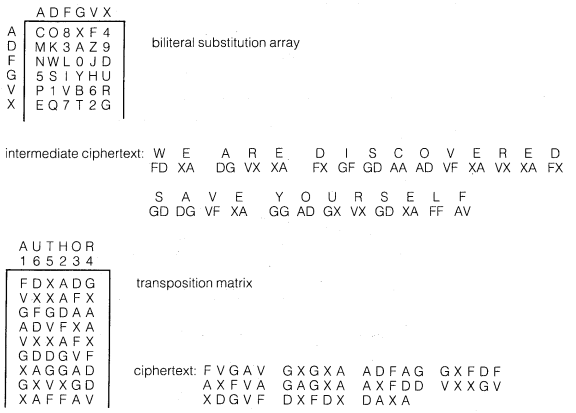



Figure 4: German ADFGVX cipher.

these electrical cipher machines employ wired rotors as a means of generating a multiplicity of substitutions. Such a rotor generally consists of a small disk. On each of the two sides of the rotor are 26 electrical contacts arranged in a circle near its edge. (The normal A–Z alphabet sequence is often engraved on the periphery of the rotor, providing a convenient means of aligning it at specific positions against a benchmark.) The contacts on one side of the rotor are connected by wires in mixed order to the contacts on the opposite side. In this way, an arbitrary set of one-to-one connections (*i.e.*, simple monoalphabetic substitutions) is realized between the two opposite sides of the rotor. A set of these rotors is usually arranged in a stack called a basket; the rotation of each of the rotors in the stack causes the next one to rotate. In some systems, each rotor advances one step in a regular sequence much as the wheels in an odometer advance $^1/_{10}$ of a revolution for every full revolution of its driving wheel. In operation, the rotors in the stack provide an electrical path from contact to contact through all of the rotors. In a straight-through rotor system (Figure 5), closing the key contact on a typewriter-like keyboard sends a current to one of the contacts on the end rotor. The current then passes through the maze of interconnections defined by the remaining rotors in the stack and their relative rotational positions to a point on the output end plate, where it is connected to either a printer or an indicator, thereby producing the cipher equivalent of a plaintext letter.

During World War II an important variation was introduced: the output end plate was a reflector to which 13 pairs of electrical contacts on the end rotor were connected. In this type of arrangement, an electrical current flows through the rotor stack and is then turned back to pass through it a second time. The output is taken from

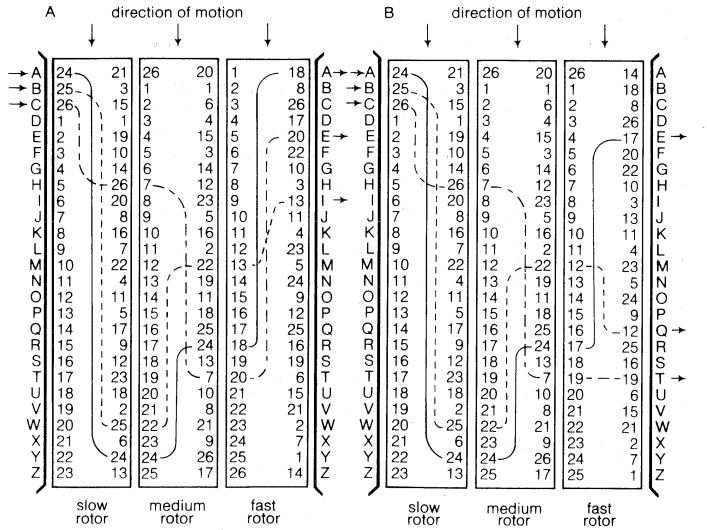



Figure 5: Three-rotor machine with wiring represented by numbered contacts. (A) Position as enciphering is begun. (B) Position after fast rotor has moved one position.

a contact in the same set to which the input was made. Accordingly, if A encrypted to W, then conversely W encrypted to A, for a particular set of rotors and positions. The advantage of this scheme is that when a pair of rotor machines is set to the same starting configuration, plaintext input to one machine generates ciphertext, which when input to the other reproduces the plaintext. The reflector also ensures that the ciphertext symbol is different from the plaintext symbol. The German Enigma cipher machine of World War II made use of this innovation.

Another type of rotor machine is much more like the Vernam encryption system (see above). Such devices are pin-and-lug machines, and they typically consist of a collection of rotors having a prime number of labeled positions on each rotor. At each position a small pin can be set to an active or inactive position. In operation, all of the rotors advance one position at each step. Therefore, if the active pin settings are chosen appropriately, the machine will not recycle to its initial pin configuration until it has been advanced a number of steps equal to the product of the number of positions in each one of the rotors. Figure 6

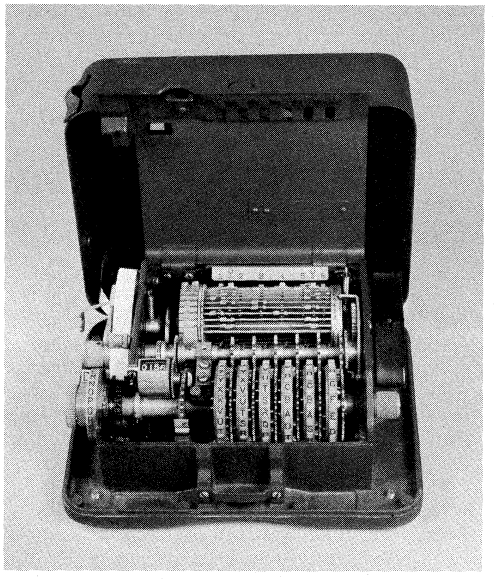


Figure 6: Hagelin design M-209 U.S. cipher machine used for tactical communications during World War II.

shows a machine of this type, the Hagelin M-209 (named for the Swedish engineer Boris Hagelin), which was used extensively by the U.S. military for tactical field communications during World War II. In the M-209 the rotors have 26, 25, 23, 21, 19, and 17 positions, respectively, so that the key period length is 101,405,850. The relationship to the Vernam encryption system is not only through the way in which a lengthy binary sequence of active pin settings in the rotors is achieved by forming the product of six much shorter ones but also in the way a symbol of plaintext is encrypted using the resulting key stream. Just behind the rotors is a squirrel cage consisting of 27 bars on each of which is a pair of movable lugs. Either or both of the lugs can be set in a position to be engaged and moved to the left on each step by a diverter actuated by the presence of an active pin on the corresponding rotor. The result is an effective gear wheel in which the number of teeth is determined by both the active pin settings and the movable lug settings. The number of teeth set determines the cyclical shift between one direct alphabet (plaintext) ABC . . . and a reverse standard alphabet ZYX . . . Thus if no tooth were present, A would encrypt to Z, B to Y, and so forth, while one tooth present would cause A to encrypt to Y, B to Z, etc. This is strictly a Vernam-type encryption—*i.e.*, encryption by subtraction modulo 26 of the key symbol from the plaintext symbol. To decrypt, the ciphertext is processed with the same pin settings that were used to encrypt it but with the cyclical shift set to occur in the opposite direction.

The significance of the above historical remarks about

product ciphers and related matters is that they point the way to the most widely adopted and used cipher in the history of cryptography—the Data Encryption Standard (DES).

**The Data Encryption Standard.** The rapid development of electronic data processing during the 1960s led to the creation of large computer data bases that were routinely communicated or made remotely accessible via resource-sharing networks. Public and private sectors alike soon recognized the need to safeguard both the privacy and the integrity of the stored data. In the United States, for example, the Privacy Act of 1974 and subsequent statutes imposed legal requirements on government agencies and certain government contractors to ensure the confidentiality of records in their possession, and, as banking became more automated and electronic fund transfers mounted into billions of dollars, the potential liability from information misuse in commercial situations created an urgent need for some way of providing information security. As a result, the National Bureau of Standards—after a lengthy evaluation in consultation with the National Security Agency, the cryptologic service of the U.S. government—announced in July 1977 that a Data Encryption Standard was to be implemented in special-purpose electronic devices and to be used in unclassified U.S. government applications for the protection of binary-coded data during transmission and storage in a computer system or network. As a consequence of the certification of the DES by the Bureau of Standards and its continuing support of the cipher in the certification of large-scale integrated (LSI) and very large-scale integrated (VLSI) electronic chips, the DES has become a de facto standard for business and commercial data security as well.

roduct
lock
ipher

The DES is a product block cipher in which 16 iterations, or rounds, of the substitution and transposition (permutation) process are cascaded. The block size is 64 bits, so that a 64-bit block of data (plaintext) can be encrypted into a 64-bit cipher at any one time. (A 64-bit block cipher can be decrypted by the DES as well.) The key, which controls the transformation, also consists of 64 bits. Only 56 of these, however, are at the user's disposal; the remaining eight bits are employed for checking parity (the state of being odd or even used as a basis for detecting errors in binary-coded data). Figure 7 is a functional schematic of the sequence of events that occurs in the DES encryption (or decryption) transformation. Subsets of the key bits are designated $K_1$, $K_2$, etc., with the subscript indicating the number of the round. The cipher function (substitution and transposition) that is used with the key bits in each round is labeled $f$. At each intermediate stage of the transformation process, the cipher output from the preceding stage is partitioned into the 32 leftmost bits, $L_i$, and the 32 rightmost bits, $R_i$. $R_i$ is transposed to become the left-hand part of the next higher intermediate cipher, $L_{i+1}$. The right-hand half of the next cipher, $R_{i+1}$, however, is a complex function of the key and of the entire preceding intermediate cipher. The essential feature to the security of the DES is that $f$ involves a very special nonlinear substitution—*i.e.*, $f(A) + f(B) \neq f(A + B)$—specified by the Bureau of Standards in tabulated functions known as $S$ boxes. This operation results in a 32-bit number, which is logically added to $R_i$ to produce the left-hand half of the new intermediate cipher. This process is repeated, 16 times in all. To decrypt a cipher, the process is carried out in reverse order, with the 16th round being first. The DES process lends itself well to integrated-chip implementation. By 1984 the Bureau of Standards had certified over 35 LSI- and VLSI-chip implementations of the DES, most on single 40-pin chips, some of which operate at speeds of several million bits per second.

The use of the DES algorithm has been made mandatory for all financial transactions of the U.S. government involving electronic fund transfer, including of course those conducted by member banks of the Federal Reserve System. Such transactions must be authenticated in conformance with the American National Standards Standard X9.9, which is written around the DES algorithm.

In a sense the DES is the logical culmination of a long history of development of single-key cryptographic algorithms, and it is this aspect that has been emphasized in the discussion thus far. In another sense, however, the DES is quite different from anything that preceded it. Cryptology has traditionally been a secretive science, so much so that it is only in the last few years that the principles on which the cryptanalysis of the Japanese and German cipher machines of World War II were based have been declassified and released. What is different about the DES is that it is a totally public cryptographic algorithm. Every detail of its operations—enough to permit anyone who wishes to to program it on a microcomputer—is available in published form in Federal Information Processing Standards Publication 46. The paradoxical result is that what is generally conceded to be one of the best cryptographic systems in the history of cryptology is also the least secret.

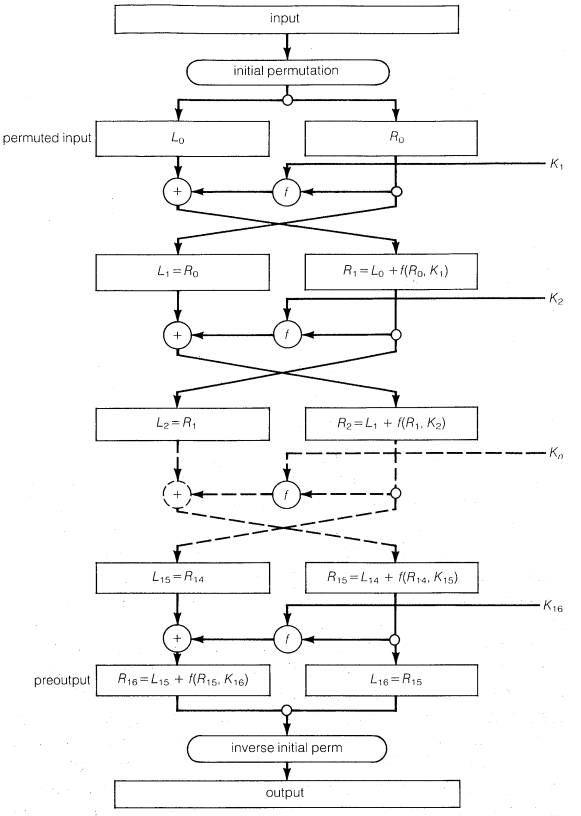



Figure 7: Flow diagram for the Data Encryption Standard operation.
From *Data Encryption Standard*, FIPS Publ. no. 46, National Bureau of Standards, 1977

**Key distribution problem.** A driving force in the practical use of cryptography is the key distribution problem inherent in single-key cryptosystems. This problem can best be described by means of a simple analogue to encryption/decryption. A and B are to be physically apart and will be able to communicate only by way of intermediaries who will pass along messages, which they may either read or alter. In order to communicate in secret, A and B obtain a strongbox with a double-acting dead bolt on it, so that a key must be used to lock or unlock the box. Locking a message in the box is analogous to encryption and unlocking to decryption. The locked box, containing a secret message, is the analogue of a cipher. If A and B each have a copy of the key and no one else does, then whenever either of them receives a locked box containing a message that person can be confident the message actually came from the other party because only one or the other of them could have locked the box (authentication). By the same token, the sender of the message can be certain that only the intended recipient can open the box (secrecy). This, of course, requires that A and B either meet in person at the time the system is set up to share the secret key or exchange the key via a trusted courier, which is how actual cryptographic keys have traditionally been disseminated. A and B must thereafter protect the key in their possession, because if C, the cryptanalyst, can

gain access to either's key to make a copy he could then lock messages in the box to impersonate A or B or else open the box sent by either party and read the enclosed message. If A or B allows his key to be compromised, however, neither authentication nor secrecy can be relied on thereafter.

Consider a system with only 1,000 subscribers, any two of whom may wish to communicate in secret or be able to verify the source of a message. In this case, A would need to exchange secret keys with 999 other subscribers and would have to protect all 999 keys in his possession in precisely the same way he did the single key he shared with B before. One of these keys would still be shared with B, but B would also need to have 998 other keys to enable him to communicate securely with persons other than A. In short, 499,500 different keys are needed if every pair of subscribers among 1,000 is to be able to communicate securely. Each subscriber would have to exchange secret keys with every one of the 999 other subscribers and then protect the integrity of the 999 keys that he is responsible for. The key distribution problem is that the number of keys required increases proportionate to half the square of the number of users. Furthermore, since the same key is held by both users of any one of the secure data links, say between A and B, A must trust B to protect his key so that A's messages to B are secret and the messages (locked boxes) A receives—purportedly sent by B—are necessarily from B (authentic). In most military and diplomatic communications this latter condition of mutual trust is satisfied; *e.g.,* one embassy, A, can rely on another embassy, B, to protect the cryptographic keys entrusted to its care. In other words, A is not concerned that B may send messages to him and later disavow them as forgeries created by A, nor claim to have received messages from A that he fraudulently created himself. In commercial and business applications, however, these are precisely the major concerns of the users—namely, the customer does not want to give the merchant the ability to create undetectable forgeries in his stead, nor would the merchant accept a system in which the customer could disavow a legitimate draft.

The key distribution problem of single-key cryptography is that the number of secret keys needed increases proportionate to the square of the number of users of the system and that each user must be responsible for the protection of as many keys as there are users (less one). The related problem arising from the symmetry of the cryptographic system is that each user must trust the other users to protect their copies of the keys that he depends on. This is a characteristic of single-key cryptographic systems and not itself part of the key distribution problem.

**Two-key cryptography.** Continuing with the locked strongbox analogy of the preceding section, assume that the double-acting dead bolt can only be locked with a key that will not unlock it once it has been locked. Assume, too, that it can be unlocked with another key that cannot lock the box. Most importantly, assume that neither of the keys can be re-created from the other. Now consider the 1,000-user system discussed earlier. A makes 999 copies of a locking key to a strongbox for which he keeps and protects the only copy of the unlocking key. He could then distribute the locking keys publicly by passing them out to everyone. Anyone wishing to send A a secret message would be able to lock it in a strongbox with confidence that only A could open the box. In this scheme, for the 1,000-user example, each user would need to have access to 1,000 keys, one of which is his personal decryption (unlocking) key that he must protect (in his own interest) and 999 of which are the other users' publicly exposed locking keys that he can use but need not protect. The total number of keys is only 2,000, in contrast to the 499,500 keys required for the single-key cryptosystem, with everyone having a single secret key and with 1,000 public keys in all. A system of this sort, variously called a two-key, public-key, or asymmetric cryptosystem, largely solves the key distribution problem.

There is, however, an obvious flaw in such a system. When A receives a locked box and opens it with his private unlocking key, he can be certain that only he and the person locking the box are privy to the contents. Yet, he cannot be sure who sent him the locked box, since any user could have locked it by using A's public locking key. The system has preserved secrecy at the expense of forfeiting any capability of authentication. If, on the other hand, A made public the unlocking (decryption) key and kept secret the locking key, any of the users could open a box that A locked, with confidence that only A could have locked it. A would have no way of restricting who opens the box, however. In this case, the ability to authenticate messages has been preserved at the expense of completely giving up secrecy or privacy. The analogy can be profitably extended one step further. If each user had distinct private locking and unlocking keys with publicly exposed matching keys, then A could first lock a message intended for B in a box using A's private locking key and then lock that box inside another box using B's public locking key. Now only B could open the outer box, thus ensuring secrecy. Because the inner box could only have been locked by A, B would be confident of the authenticity of the message. It is this protocol involving iterated encryptions with public and private keys that makes public-key cryptography so attractive for commercial and private applications.

In single-key cryptosystems, the secret encryption and decryption keys $K_E$ and $K_D$ used by the transmitter and receiver, respectively, were either the same or could be easily computed from a knowledge of the other key. In two-key cryptography this is not true. Whereas single-key cryptosystems have been in use for centuries, two-key cryptography is a recent development. In 1976, Whitfield Diffie and Martin E. Hellman proposed a conceptual scheme for this kind of cryptosystem, which they called a public-key cryptosystem, because users could avoid the key distribution problem by simply publishing their encryption keys in a public directory. This was the first discussion of two-key cryptography in the open literature. However, Adm. Bobby Inman, while director of the U.S. National Security Agency, pointed out that two-key cryptography had been discovered at the agency a decade earlier.

Avoiding the key distribution problem

For two-key cryptography to be possible, cryptosystems must be devised with the following properties:

1. It must be easy for the cryptographer to calculate a matched pair of keys $K_E$ and $K_D$ but computationally infeasible (virtually impossible) for a cryptanalyst to recover either key no matter how much text is available.
2. The encryption and decryption operations should be (computationally) easy for legitimate users to carry out.
3. At least one of the keys must be computationally infeasible for the cryptanalyst to recover even when he knows the other key and arbitrarily many matching plaintext and ciphertext pairs.

No cryptosystem so far devised has satisfied all of these conditions, at least not in a provable way. As a consequence, cryptographers have devised cryptographic functions of this sort by starting with a "hard" mathematical problem, such as the knapsack problem or factoring a product of large prime integers, and attempting to make the cryptanalysis of the scheme be equivalent to solving the hard problem. If this can be done, the cryptosecurity of the scheme is at least as good as the underlying mathematical problem is hard to solve.

Knapsack problem

Ralph C. Merkle and Martin E. Hellman have put forth one of the best known proposals for a public-key cryptosystem. They suggest using the knapsack, or subset-sum, problem as the basis for the system. This problem entails determining whether a number can be realized as the sum of some subset of a given collection of weights and, even more importantly, which subset has the desired sum. For example, using the ten weights 14, 28, 56, 82, 90, 132, 197, 284, 341, 455, the figure 515 is realizable in three different ways as a sum while 516 cannot be realized at all. Merkle and Hellman proposed starting with a collection of weights in which each was greater than the sum of all smaller weights. In this case, it is easy to decide whether a given number $S$ can be represented as a sum of a subset of a set of $n$ weights. $S$ is first compared to the largest weight in the set, and if $S$ is larger the weight is in the subset.

If $S$ is smaller, the weight is not in the subset. On each successive step, $S$ less the sum of the weights that have been kept in the subset is compared to the next weight in order of size. The weight is kept in the subset if and only if it is no larger than the difference of $S$ and the sum of the weights already retained. Clearly this finds the subset that sums to $S$, if any, in only $n$ comparisons.

The cryptographer uses a secret (invertible) transformation to turn such an easy knapsack into a hard one. The legitimate users, however, knowing the secret transformation, could easily invert the hard knapsack back to the easy knapsack. In the Merkle–Hellman scheme, ciphers were to be sums of subsets chosen from the hard knapsack. Decryption was equivalent to finding the identity of the weights used in forming the sum, which was thought to be a very hard problem if one only knew the hard knapsack, but was an easy problem for anyone knowing the secret transformation because then the problem could be solved using the easy knapsack. Unfortunately for the cryptographer, knapsack-based cryptosystems have been shown not to be secure.

Other proposals for basing the cryptosecurity of two-key cryptoalgorithms on the difficulty of taking discrete logarithms or of inverting large linear codes, etc., also have been shown to be cryptanalytically weak—at least for those cases most attractive to cryptography.

As a result, the main contenders for two-key cryptosystems in the mid-1980s were those whose cryptosecurity derives from the difficulty of factoring large composite integers. The best known of these is the Rivest–Shamir–Adleman (RSA) cryptoalgorithm. In this system a user chooses a pair of prime numbers $p$ and $q$ so large that factoring the product $n = pq$ is beyond all projected computing capabilities. Testing for primality is easy while factoring is very difficult, so that this is easy to do. As of 1984, the consensus was that $p$ and $q$ need to be about 75 decimal digits in size, and so $n$ is roughly a 150-digit number. Since the largest hard numbers that can currently be factored are 80 digits or less in size and the difficulty of factoring grows exponentially with the size of the number, 150 digits appear cryptosecure for the foreseeable future. Having chosen $p$ and $q$ and hence the modulus $n = pq$, the user selects an arbitrary integer $K_E$ less than $n$ and relatively prime to $p-1$ and $q-1$—*i.e.*, so that 1 is the only factor in common between $K_E$ and $(p-1)(q-1)$. This ensures that there is another number $K_D$, so that the product $K_E \times K_D$ will leave a remainder of 1 when divided by the least common multiple of $(p-1)$ and $(q-1)$. $K_D$ can be easily calculated using the Euclidean algorithm if one knows $p$ and $q$. If one does not know $p$ and $q$, it is equally as difficult to find $K_D$ as it is to factor $n$, which is the basis for the cryptosecurity of the RSA algorithm. $K_E$ and $K_D$, when chosen in this way, exhibit the following remarkable behaviour. If $m$ is any number less than $n$, then

RSA cryptogorithm

$$m^{K_E} \equiv c \pmod{n} \qquad (1)$$

and

$$c^{K_D} \equiv m \pmod{n} \qquad (2)$$

where the mathematical notation is a shorthand for the statement that $c$ is the remainder (cipher) left when $m$ is raised to the $K_E$ power and divided by the modulus $n$, and conversely that $m$ is the remainder obtained when $c$ is raised to the $K_D$ power and divided by $n$. This is true for any $m$ less than $n$ where $m$ is the encoded—not encrypted—message. $K_E$ is the encryption key and $K_D$ is the decryption key. There are some constraints on the choice of $p$ and $q$ and of $K_E$ to maximize cryptosecurity, but these are unimportant to understanding the RSA two-key cryptosystem. For example, if $p = 5$ and $q = 11$ so that the $n = 55$, then the least common multiple of $p-1$ and $q-1$ is $20 = 2^2 \cdot 5$. In this case, any key $K_E$ not divisible by 2 or 5 will have a matching key $K_D$. In particular, let $K_E = 7$ for which $K_D = 3$. If $m = 2$, then $2^7 = 128$, which leaves the remainder (cipher) 18 when divided by 55. Similarly, when the cipher, 18, is raised to the power $K_D$, $18^3 = 5{,}832$, which leaves the remainder (plaintext) 2 when divided by 55. Similar behaviour would have been observed for any other choice of a number less than 55 or for any other pair of matched exponents (keys), such as $K_E = K_D = 11$.

To implement a secrecy channel using the RSA cryptosystem, user A would publish $K_E$ and $n$ in the public directory but keep secret $K_D$ and $p$ and $q$ (he need not record nor remember $p$ and $q$). Anyone wishing to send a private message to A would encode it into numbers less than $n$ and calculate $c$ using formula (1). A can calculate $m$ using formula (2), but the presumption—and evidence thus far—is that for almost all ciphers no one else can find $m$, given $c$, unless he can also factor $n$.

Similarly, to implement an authentication channel, A would publish $K_D$ and $n$ and keep $K_E$ (and $p$ and $q$) secret. To sign a message, A simply appends some agreed-upon numbers to the encoded message and then encrypts the resulting number using the secret $K_E$ to obtain $c$. Anyone can decrypt $c$ using the public $K_D$ and must conclude, if the agreed-upon numbers are appended, that A originated the cipher, since only he knew $K_E$ and hence could have calculated a cipher that would decrypt to a message with these properties.

Thus far, all proposed two-key cryptosystems exact a high price for this separation of the privacy or secrecy channel from the authentication or signature channel. The greatly increased amount of computation involved in the encryption/decryption process significantly cuts the channel capacity (bits per second of message information communicated). While there exist several single-chip implementations of the DES single-key algorithm that process information at several million bits per second, the throughput of a comparably secure RSA chip—one using roughly a 150-digit modulus—is only a few thousand bits per second. As a result, the main application of two-key cryptography is in hybrid systems. In such a system a two-key algorithm is used necessarily at low speed to avoid the key-distribution problem by exchanging a randomly generated session key to be used with a single-key algorithm at high speed for communication. At the end of the session, this key is discarded and others are generated as the need arises.

**Block and stream ciphers.** In general, cipher systems transform fixed-size pieces of plaintext into ciphertext. In the older manual systems these pieces were usually single letters or characters or occasionally, as in the Playfair cipher, digraphs, since this was as large a unit as could feasibly be encrypted and decrypted by hand. Systems that operated on trigrams or larger groups of letters were proposed and understood to be potentially securer, but they were never implemented because of the difficulty in manual encryption and/or decryption. In modern single-key cryptography, using the DES, for example, the units of information are often as large as 64 bits or about 13½ alphabetic characters, whereas two-key cryptography based on the RSA algorithm appears to have settled on 512 bits, or about 109 alphabetic characters, as the unit of encryption.

Characteristics of block and stream ciphers

A block cipher breaks the plaintext into blocks of the same size for encryption using a common key: the block size for a Playfair cipher is two letters and for the DES used in electronic codebook mode, 64 bits of binary-encoded plaintext. While a block could consist of a single symbol, normally it is larger.

A stream cipher also breaks the plaintext into units, normally a single character, and then encrypts the $i^{th}$ unit of the plaintext with the $i^{th}$ unit of a key stream. Vernam encryption with a onetime key is an example of such a system, as are rotor cipher machines and the DES used in the output feedback mode to generate a key stream. Stream ciphers depend on the receiver's using precisely the same part of the key stream to decrypt the cipher that was employed to encrypt the plaintext. They thus require that the transmitter's and receiver's key-stream generators be synchronized. This means either that they must be synchronized initially and stay in sync thereafter or that if synchronization is lost the cipher will be decrypted into a garbled form until it can be reestablished. This latter property of self-synchronizing cipher systems results in what is

| Letter | Number of occurrences | Frequency | Letter | Number of occurrences | Frequency | Letter | Number of occurrences | Frequency |
|---|---|---|---|---|---|---|---|---|
| E | 8,915 | .127 | C | 3,188 | .045 | G | 1,113 | .016 |
| T | 6,828 | .097 | L | 2,810 | .040 | W | 914 | .013 |
| I | 5,260 | .075 | D | 2,161 | .031 | V | 597 | .008 |
| A | 5,161 | .073 | P | 2,082 | .030 | K | 548 | .008 |
| O | 4,814 | .068 | Y | 1,891 | .027 | X | 330 | .005 |
| N | 4,774 | .067 | U | 1,684 | .024 | Q | 132 | .002 |
| S | 4,700 | .067 | M | 1,675 | .024 | Z | 65 | .001 |
| R | 4,517 | .064 | F | 1,488 | .021 | J | 56 | .001 |
| H | 3,452 | .049 | B | 1,173 | .017 | | | |

known as error propagation, an important parameter in any stream-cipher system.

## CRYPTANALYSIS

History abounds with examples of the seriousness of the cryptographer's failure and the cryptanalyst's success. In World War II the Battle of Midway, which marked the turning point of the naval war in the Pacific, was won by the United States largely because cryptanalysis had provided Adm. Chester W. Nimitz with information about the Japanese diversionary attack on the Aleutian Islands and of the Japanese order of attack for Midway. Another famous example of the consequences of a cryptanalytic success was the British cryptanalysis during World War I of a telegram from the German foreign minister, Arthur Zimmermann, to the German minister in Mexico City, Heinrich von Eckardt, laying out a plan to reward Mexico for entering the war as an ally of Germany. This breakthrough caused Pres. Woodrow Wilson of the United States to reverse his earlier opposition to U.S. entry into the war on the side of the Allies, thereby causing that momentous action much sooner than it would have occurred otherwise.

**Basic aspects.** While cryptography is clearly a science with well-established analytical and synthetic principles, cryptanalysis is as much an art as it is a science. The reason is that success in cryptanalyzing a cipher is as often as not a product of flashes of inspiration, gamelike intuition, and, most importantly, recognition by the cryptanalyst of pattern or structure, at almost the subliminal level, in the cipher. The great U.S. cryptanalyst Herbert O. Yardley described the crucial step in breaking the Japanese ciphers soon after World War I: "Finally one night I awakened at midnight, for I had retired early, and out of the darkness came the conviction that a certain series of two-letter codewords absolutely must equal AIRURANDO (Ireland). The other words danced before me in rapid succession: DOKURITSU (independence), DOITSU (Germany), OWARI (stop)." It is easy to state and demonstrate the principles on which the scientific part of cryptanalysis depends but nearly impossible to convey an appreciation of the art with which the principles are applied.

Cryptanalyzing single-key cryptosystems

Cryptanalysis of single-key cryptosystems depends on one simple fact—namely, that traces of structure or pattern in the plaintext may survive encryption and be discernible in the ciphertext. Take for example the following: in a monoalphabetic substitution cipher in which each letter is simply replaced by another letter, the frequency distribution with which letters occur in the plaintext alphabet and in the ciphertext alphabet is identical. The cryptanalyst can use this fact in two ways: first, to recognize that he is faced with a monoalphabetic substitution cipher and, second, to aid him in selecting the likeliest equivalences of letters to be tried. The Table shows the number of occurrences of each letter in the text of this article, which approximates the raw frequency distribution for most technical material. The following cipher is the encryption of the first sentence of this paragraph using a monoalphabetic substitution:

UFMDHQAQTMGRG BX GRAZTW PWM
UFMDHBGMGHWOG VWDWAVG BA BAW
GRODTW XQUH AQOWTM HCQH HFQUWG
BX GHFIUHIFW BF DQHHWFA RA HCW
DTQRAHWLH OQM GIFJRJW WAUFMDHRBA
QAV SW VRGUWFARSTW RA HCW
URDCWFHWLH.

*W* occurs 20 times in the cipher, *H* occurs 16, etc. Even the most unskilled cryptanalyst using the frequency data in the Table should have no difficulty in recovering the plaintext and all but four symbols of the key in this case.

It is possible to conceal information about raw frequency of occurrence by providing multiple cipher symbols for each plaintext letter in proportion to the relative frequency of occurrence of the letter; *i.e.,* twice as many symbols for E as for S, etc. The collection of cipher symbols representing a given plaintext letter are called homophones. If the homophones are chosen randomly and with uniform probability when used, the cipher symbols will all occur (on average) equally often in the ciphertext. No less a mathematician than Carl Friedrich Gauss believed that he had devised an unbreakable cipher by introducing homophones. Unfortunately for Gauss and other cryptographers such is not the case, since there are many other persistent patterns in the plaintext that may partially or wholly survive encryption. Digraphs, for example, show a strong frequency distribution: TH occurring most often, about 20 times as frequently as HT, and so forth. With the use of tables of digraph frequencies that partially survive even homophonic substitution, it is still an easy matter to cryptanalyze a random substitution cipher, though the amount of ciphertext needed grows to a few hundred instead of a few tens of letters.

Homophones

If the cipher preserves the breaks between words as they existed in the plaintext, the frequency distribution for starting and ending letters and digrams in words that are all distinct and different from the raw frequency of letter distributions can be used to advantage. Patterns of letters in words such as XYXX where X is the same cipher symbol can only fit a small selection of words such as BIBB, EPEE, LOLL, LULL, and SASS. In the heyday of manual cryptanalysis, volumes of word patterns were compiled. These are only some of the most obvious and easily described patterns whose persistence may provide a clue to the cryptanalyst. In English there are useful correlations between symbols up to eight or nine positions displaced in a word and of course context dependencies over entire sentences, all of which are of potential use to the cryptanalyst.

**Types of cryptanalysis.** There are three generic types of cryptanalysis characterized by what the cryptanalyst knows: (1) ciphertext only; (2) known ciphertext/plaintext pairs; and (3) chosen plaintext or chosen ciphertext. In the discussion of the preceding paragraphs, the cryptanalyst knew only the ciphertext and general structural information about the plaintext. Often the cryptanalyst either will know some of the plaintext or will be able to guess at, and exploit, a likely element of the text, such as a letter beginning with "Dear Sir" or a computer session starting with "LOG IN." The last category represents the most favourable situation for the cryptanalyst in which he can cause either the transmitter to encrypt a plaintext of his choice or the receiver to decrypt a ciphertext that he chose. Of course, for single-key cryptography there is no distinction between chosen plaintext and chosen ciphertext, but in two-key cryptography it is possible for one of the encryption or decryption functions to be secure against chosen input while the other is vulnerable.

Because of its reliance on "hard" mathematical problems as a basis for cryptoalgorithms and because one of the keys is publicly exposed, two-key cryptography has led to a new type of cryptanalysis that is virtually indistinguishable from research in any other area of computational

mathematics. Unlike the ciphertext attacks or ciphertext/plaintext pairs attacks in single-key cryptosystems, this sort of cryptanalysis is aimed at breaking the cryptosystem by analysis that can be carried out based only on a knowledge of the system itself. Obviously there is no counterpart to this kind of cryptanalytic attack in single-key systems. One of the most attractive schemes for exchanging session keys in a hybrid cryptosystem depended on the ease with which a number (primitive root) could be raised to a power (in a finite field), as opposed to the difficulty of calculating the discrete logarithm. A special-purpose chip to implement this algorithm was produced by a U.S. company, and several others designed secure electronic mail and computer-networking schemes based on the algorithm. In 1983 Donald Coppersmith found a computationally feasible way to take discrete logarithms in precisely those finite fields that had been of greatest cryptographic interest and thereby gave to the cryptanalyst a tool with which to break those cryptosystems. Similarly, the RSA cryptoalgorithm is no securer than the modulus is difficult to factor, so that a breakthrough in factoring would also be a cryptanalytic breakthrough. In 1980 one could factor a difficult 50-digit number at an expense of 1,000,000,000,000 elementary computer operations (add, subtract, shift, and so forth). By 1984 the state of the art in factoring algorithms had advanced to a point where a 75-digit number could be factored in 1,000,000,000,000 operations. If a mathematical advance made it feasible to factor 150 or more digit numbers in the same number of operations, this would make it possible for the cryptanalyst to break several commercial RSA schemes. In other words, the security of two-key cryptography depends on well-defined mathematical questions in a way that single-key cryptography generally did not and conversely equates cryptanalysis to mathematical research in an atypical way.

Factoring algorithms

## HISTORY

**Early cryptographic systems and applications.** People have probably tried to conceal information in written form from the time that writing developed. Examples survive in stone inscriptions, cuneiform tablets, and papyruses showing that the ancient Egyptians, Hebrews, Babylonians, and Assyrians all devised protocryptographic systems both to deny information to the uninitiated and to enhance its significance when it was revealed. The first recorded use of cryptography for correspondence, however, was by the Spartans, who as early as 400 BC employed a cipher device called the scytale for secret communications between military commanders. The scytale consisted of a tapered baton, around which was spirally wrapped a strip of parchment or leather on which the message was written. When unwrapped, the letters were scrambled in order and formed the cipher; however, when the strip was wrapped around another baton of identical proportions to the original, the plaintext reappeared. Thus, the Greeks were the inventors of the first transposition cipher. During the 4th century BC Aeneas Tacticus wrote a work entitled *On the Defense of Fortifications,* one chapter of which was devoted to cryptography, making it the earliest treatise on the subject. Another Greek, Polybius, devised a means of encoding letters into pairs of symbols by a device called the Polybius checkerboard, which is a true biliteral substitution and presages many elements of later cryptographic systems. Similar examples of primitive substitution or transposition ciphers abound in the history of other civilizations. The Romans used monoalphabetic substitution with a simple cyclic displacement of the alphabet. Julius Caesar employed a shift of three positions so that plaintext A was encrypted as D, while Augustus Caesar used a shift of one position so that plaintext A was enciphered as B.

The scytale

The first people to clearly understand the principles of cryptography and to elucidate the beginnings of cryptanalysis were the Arabs. They devised and used both substitution and transposition ciphers and discovered the use of both letter frequency distributions and probable plaintext in cryptanalysis. As a result, by about 1412, al-Kalkashandī could include a respectable, if elementary, treatment of several cryptographic systems in his encyclopaedia *Ṣubḥ al-aʿshā* and give explicit instructions on how to cryptanalyze ciphertext using letter frequency counts complete with lengthy examples to illustrate the technique.

European cryptology dates from the Middle Ages, during which it was developed by the Papal States and the Italian city-states. The earliest ciphers involved only vowel substitution (leaving consonants unchanged). The first European manual on cryptography (*c.* 1379) was a compilation of ciphers by Gabriele de Lavinde of Parma, who served Pope Clement VII. This manual, now in the Vatican archives, contains a set of keys for 24 correspondents and embraces symbols for letters, nulls, and several two-character code equivalents for words and names. The first brief code vocabularies, called nomenclators, were gradually expanded and became the mainstay for several centuries for diplomatic communications of nearly all European governments. In 1470 Leon Battista Alberti published *Trattati in cifra,* in which he described the first cipher disk; he prescribed that the setting of the disk should be changed after enciphering three or four words, thus conceiving of the notion of polyalphabeticity. Giambattista della Porta provided a modified form of square table and the earliest example of a digraphic cipher in *De furtivis literarum notis* (1563). The *Traicté des chiffres* published in 1586 by Blaise de Vigenère contains the square table commonly attributed to him (Figure 2) and descriptions of the first plaintext and ciphertext autokey systems.

First cipher disk

By 1860 large codes were in common use for diplomatic communications, and cipher systems had become a rarity for this application. Cipher systems prevailed, however, for military communications except for high-command communications because of the difficulty of protecting codebooks from capture or compromise in the field. In the early history of the United States, codes were widely used, as were book ciphers. During the Civil War the Federal Army made extensive use of transposition ciphers, in which a key word indicated the order in which columns of the array were to be read and in which the elements were either plaintext words or code word replacements for plaintext. The Confederate Army primarily used the Vigenère cipher and on occasion monoalphabetic substitution. While the Union cryptanalysts solved most of the intercepted Confederate ciphers, the Confederacy in desperation sometimes published Union ciphers in newspapers, appealing for help from readers in cryptanalyzing them.

**Developments during World Wars I and II.** During the first two years of World War I, the belligerents employed cipher systems almost exclusively for tactical communications; code systems were still used mainly for high-command and diplomatic communications. Figure 8, showing a U.S. Signal Corps cipher disk, points up the lack of sophistication in the field cipher systems. Nevertheless, some complicated cipher systems were used for high-level communications by the end of the war, the most famous of which was the German ADFGVX fractionation cipher.

The communications needs of telegraphy and radio and the maturing of mechanical and electromechanical technology came together in the 1920s to bring about a true

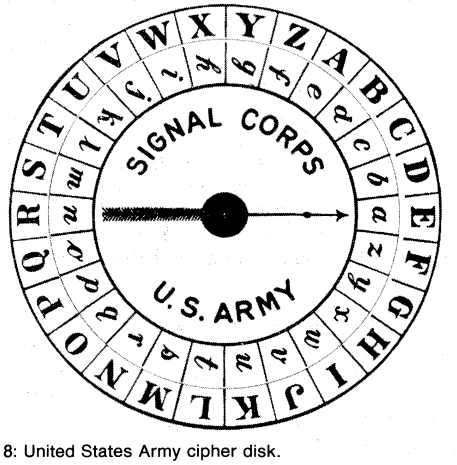



Figure 8: United States Army cipher disk.

revolution in cryptodevices: the development of rotor cipher machines. Although the concept of the rotor had been anticipated in the older mechanical cipher disks, the credit goes to an American, Edward H. Hebern, for first recognizing that by hardwiring a monoalphabetic substitution in the connections from the contacts on one side of an electrical rotor to those on the other side and cascading a collection of such rotors, polyalphabetic substitutions of almost arbitrary complexity could be realized. Hebern also recognized that a permutation in which several letters were shifted by the same amount in the rotor connections, say A to D and B to E, was cryptographically weaker than one in which this partial periodicity was minimized and designed his rotors accordingly. Starting in 1921 and continuing through the next decade, Hebern constructed a series of steadily improving rotor machines that were evaluated by the U.S. Navy and undoubtedly led to the United States' superior position in cryptology as compared to the Axis powers during World War II. The 1920s were marked by a series of challenges by inventors of cipher machines to national cryptologic services and by one service to another, resulting in a steady improvement both of cryptomachines and of cryptanalytic techniques for the analysis of machine ciphers. At almost the same time that Hebern was inventing the rotor cipher machine in the United States, European engineers, notably Hugo A. Koch of The Netherlands and Arthur Scherbius of Germany, independently discovered the rotor concept and designed machines that became the precursors of the best known cipher machine in history, the German Enigma used in World War II (Figure 9). By an indirect path of development, these machines were the stimulus for the TYPEX, the cipher machine employed by the British during World War II. Figure 10 shows a TYPEX with the cover opened to reveal the mechanism of a typical rotor machine. The M-134-C (SIGABA) cipher machine, introduced by the United States during World War II, is shown in Figure 11.

The Japanese cipher machines of World War II have an interesting history linking them to both the Hebern machines and the Enigma. The Washington Conference on naval disarmament (1921–22) had as a primary objective limiting the total tonnage of capital ships (battleships, cruisers, and aircraft carriers) by the major powers—the United States, Great Britain, Japan, France, and Italy. The most difficult problem was the way in which this tonnage was to be allocated among the five countries. The Japanese Foreign Office sent detailed cipher instructions to its ambassador in Washington, D.C., to negotiate for a 10-to-7 U.S.-to-Japanese tonnage ratio, to fall back to 10 to 6.5 if that failed, and only as a last resort to retreat to a lowest acceptable ratio of 10 to 6. In a flash of inspiration Herbert O. Yardley broke the Japanese ciphers (see above), enabling the U.S. representative, Secretary of State Charles Evans Hughes, to press for this lower limit. The Japanese reluctantly accepted the inferior position of 10:10:6:3.3:3.3 (the United States, Britain, Japan, France, and Italy, respectively) laid out in the Five-Powers Treaty. Primarily because of a failure of their cryptography, they had settled for 100,000 tons of shipping less than they might otherwise have obtained, a difference of three capital ships. When Yardley later revealed in 1928 and subsequently published in *The American Black Chamber* the details of the American successes in cryptanalyzing the Japanese ciphers, with the associated costs to Japan, the Japanese government set out to develop the best cryptomachines possible. With this end in mind, it purchased the rotor machines of Hebern and Hagelin and the commercial Enigmas, as well as several other contemporary machines, for study and analysis. In 1930 the Japanese Foreign Office put into service its first rotor machine, which was code-named RED by U.S. cryptanalysts. In 1935–36 the U.S. Army Signal Intelligence Service (SIS) team of cryptanalysts, led by William F. Friedman, succeeded in cryptanalyzing RED ciphers, drawing heavily on its previous experience in cryptanalyzing the machine ciphers produced by the Hebern rotor machines.

It was an ironic twist of fate that the Hebern machines, which were never commercially successful, played such a pivotal role in the design of two widely used rotor machines and in the evolution of the techniques that were

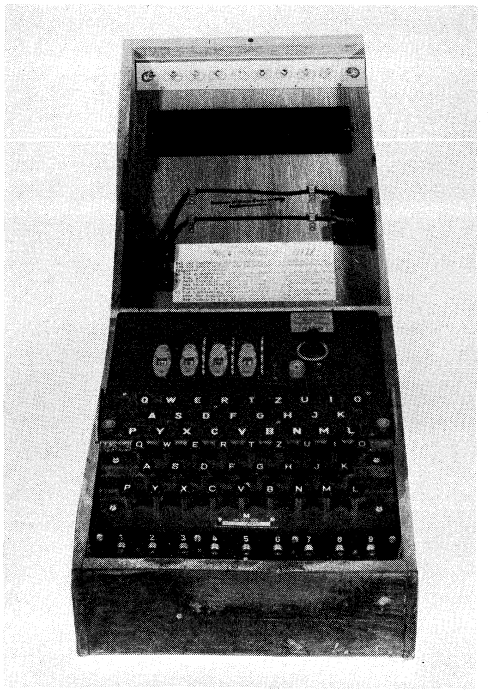


Figure 9: German naval four-rotor Enigma cipher machine.

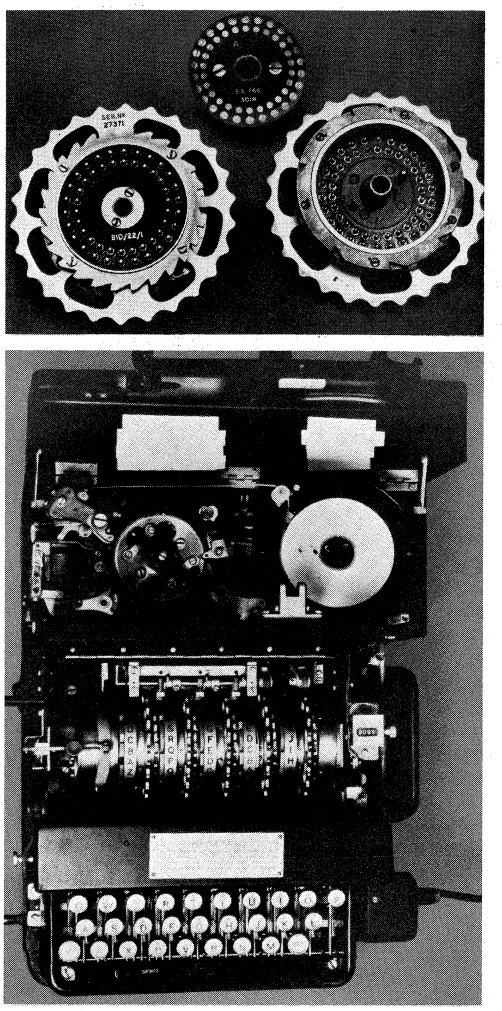


Figure 10: (Top) Rotors for (bottom) TYPEX cipher machine (shown with top open).
From the collection of Louis Kruh

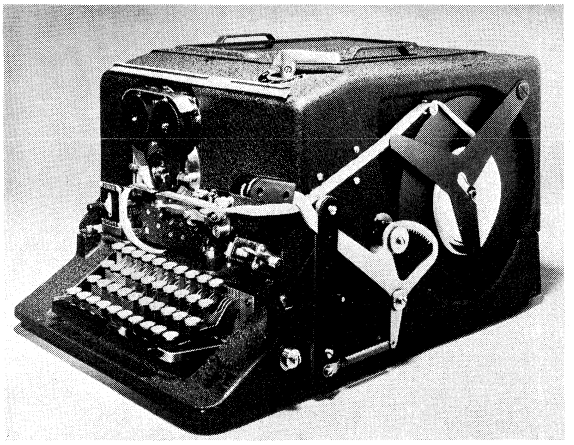


Figure 11: M-134-C (SIGABA) cipher machine used by the United States during World War II.

PURPLE cipher machine

vital to the cryptanalysis of the RED ciphers. In 1939 the Japanese introduced a new cipher machine, code-named PURPLE by U.S. cryptanalysts, in which rotors were replaced by telephone stepping switches. Because the replacement of RED machines by PURPLE machines was gradual, providing an enormous number of cribs between the systems to aid cryptanalysts, and because the Japanese had taken a shortcut to avoid the key distribution problem by generating keys systematically, U.S. cryptanalysts were able not only to cryptanalyze the RED ciphers but also to anticipate keys several days in advance. Figure 12 shows one of the functionally equivalent PURPLE cipher machines constructed by Friedman and his SIS associates and used throughout the war to decrypt Japanese ciphers. Apparently no PURPLE machine survived the war. Another Japanese cipher machine code-named JADE was essentially the same as the PURPLE (Figure 13). It differed from the latter chiefly in that it typed Japanese kana characters directly.

The greatest triumphs in the history of cryptanalysis were the Polish and British solution of the German Enigma ciphers and of two teleprinter ciphers, code-named ULTRA, and the American cryptanalysis of the Japanese RED, ORANGE, and PURPLE ciphers, code-named MAGIC. These developments played a major role in the Allies' conduct of World War II. Of the two, the cryptanalysis of the Japanese ciphers is the more impressive technically, because it was a tour de force of cryptanalysis against ciphertext alone. In the case of the Enigma machines, the basic patents had been issued in the United States, commercial machines were widely available, and the rotor designs were known to Allied cryptanalysts from a German code clerk. Although such factors do not diminish the practical importance of the ULTRA intercepts, they did make the cryptanalysis easier.

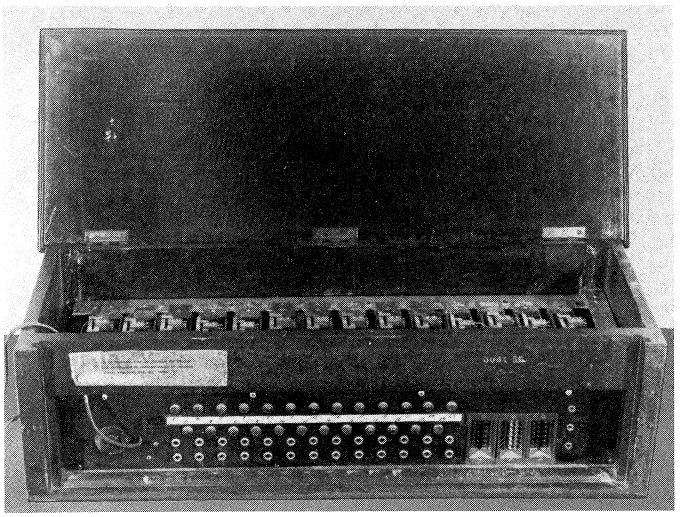


Figure 12: Functional analogue of the Japanese PURPLE cipher machine, *c.* 1940.

**The impact of modern electronics.** In the years immediately following World War II, the electronic technology that had been developed in support of radar and the recently discovered digital computer were adapted to cryptomachines. The first such devices were little more than rotor machines in which rotors had been replaced by electronically realized substitutions. The advantage of these electronic rotor machines was speed of operation; the disadvantages were the cryptanalytic weaknesses inherited from mechanical rotor machines and the principle of cyclically shifting simple substitutions for realizing more complex product substitutions.

It was quickly recognized, however, that electronics made possible the practical realization of far more complex cryptographic functions than had previously been feasible. By the 1960s transistors permitted the implementation of complex transformations in special-purpose circuitry. There is a small amount of information in the open literature about the electronic cipher machines used by the various national cryptologic services, but the most reliable indication of cryptographic developments in the period from the final generation of rotor machines—the KL-7 developed by the United States for the North Atlantic Treaty Organization (NATO)—to the appearance of DES and public-key systems in1976 is to be found in commercial equipment. One class of electronic devices representative of contemporary cryptomachine technology is the Fibonacci generator, named for the Fibonacci sequences of number theory. In the classical Fibonacci sequence of integers ... 21, 13, 8, 5, 3, 2, 1, each succeeding leftmost term is the sum of the two terms to its right; *i.e.*, $x_i = x_{i-1} + x_{i-2}$. By loose analogy, any sequence in which each term is the sum of a collection of earlier terms in fixed (relative) locations is called a Fibonacci sequence. For example, if $x_i = x_{i-1} + x_{i-2} + x_{i-3}$, the sequence ... 68, 37, 20, 11, 6, 3, 2, 1 ... results.

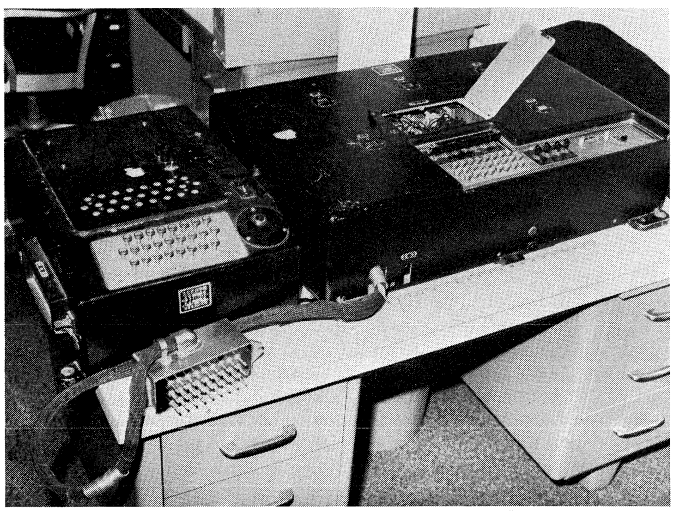


Figure 13: Japanese JADE cipher machine, a member of the PURPLE family of cipher machines.

In an $n$-stage Fibonacci generator (Figure 14), the contents of the register are shifted right one position at each step—the bit at the extreme right is shifted out and lost—and the new left-hand bit is determined by the logical sum of bits occurring in prescribed locations in the shift register before the shift was made. For example, for $n = 5$ and $x_i = x_{i-1} \oplus x_{i-4} \oplus x_{i-5}$ one obtains the cycle

0 1 0 1 1 1 0 1 1 0 0 0 1 1 1 1 1 0 0 1 1 0 1 0 0 1 0 0 0 0 1,

which is a maximal-length sequence realizable with a five-stage generator. The relevance of Fibonacci generators to cryptography is seen if the sequence is read five bits at a time by successively shifting one bit position to give the scrambled ordering of the integers 1 through 31 (Figure 15), which resembles the scrambled orderings seen

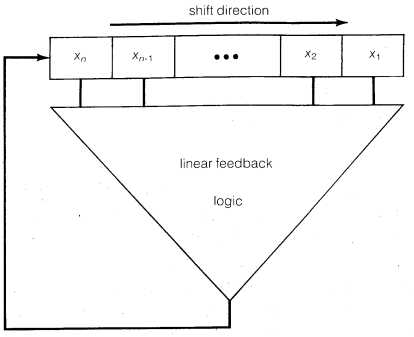



Figure 14: The $n$-stage Fibonacci generator.

earlier in rotors. If $n = 31$, for instance, the sequence could extend for 2, 147, 483, 647 bits before repeating. If one recalls Claude Shannon's dictum that an ideal cryptomachine should offer the possibility of encrypting any plaintext block into any ciphertext block, a maximal-length linear sequence certainly offers this possibility since every nonzero number occurs somewhere in the cycle. The cryptographic problem is that the combining operation used to determine successive states in the sequence from earlier ones is linear and hence easily invertible, even though the sequence can be $2^n-1$ bits in length before repeating. Another problem is how the key is to be used. The obvious choice—namely, to simply use the key to determine the number of steps in the cycle from the plaintext $n$-tuple to the ciphertext $n$-tuple—is cryptographically insecure because a known plaintext cryptanalysis would quickly reveal the key. A frequently reinvented solution to this problem has been to use the number found in selected locations of one maximal-length feedback shift register, in which the key is used as the initial register fill, to control the number of steps from the plaintext $n$-tuple to the ciphertext $n$-tuple in the cycle of another linear feedback shift register. In schemes of this sort the key register is generally stepped forward to hide the key itself before any encryption of plaintext is carried out and then advanced sufficiently many steps between encryptions to ensure diffusion of the keying variables. To encrypt an $n$-bit block of plaintext, the text is loaded into the main shift register and the machine stepped through a specified number of steps, normally a multiple of the number of bits in the key, sufficient to diffuse the information in the plaintext and in the key over all positions in the resulting ciphertext. To decrypt ciphertext, it is necessary to have an inverse combiner function or for the original encryption function to be involutory—*i.e.,* for the encryption of the cipher to be the plaintext again. It is not difficult to design the feedback logic to make an involutory machine, which, when filled with a block of ciphertext and reencrypted, will result in the register's containing the original plaintext. Pictorially, the machine has simply retraced its steps in the cycle(s). Linearity in the logic, though, is a powerful aid to the cryptanalyst, especially if a matched plaintext/ciphertext attack is possible.

Maximal-length feedback shift register

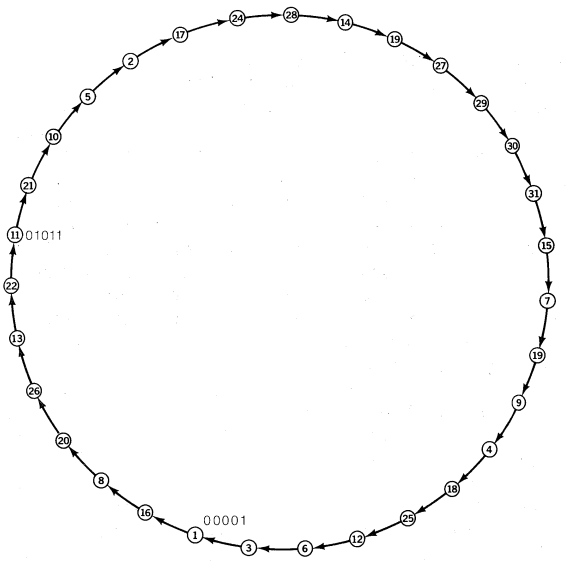



Figure 15: A maximal-length sequence produced by a five-stage Fibonacci generator.

With a slight modification, this approach constitutes the basis of several commercially available cryptographic devices that function in a manner quite similar to the pin-and-lug cipher machines previously described. One such cryptomachine has six maximal-length linear feedback shift registers in which the stepping is controlled by another shift register; the contents of the latter are used to address a (nonlinear) lookup table defined by keys supplied by the user.

Nonlinear feedback logic

As a result, cryptographers have devised a number of nonlinear feedback logics that possess such desirable properties as diffusion of information (to spread the effects of small changes in the text) and large-cycle structure (to prevent exhaustive search) but which are computationally infeasible to invert working backward from the output sequence to the initial state(s), even with many pairs of matched plaintext/ciphertext. The nonlinear feedback logic, used to determine the next bit in the sequence, can be employed in much the same way as linear feedback logic. Nonlinear systems that also satisfy the necessary conditions for invertibility so as to enable the receiver to decrypt the cipher always break up into still very large subcycles. The complicating effect of the key on the ciphertext in nonlinear logics, however, greatly contributes to the difficulty faced by the cryptanalyst. It is generally accepted that no purely cryptanalytic breaks of the cryptoalgorithms developed by the national cryptologic services of the major powers have occurred in recent years, nor are they apt to occur, as a result of the advances in cryptology since World War II.

**BIBLIOGRAPHY.** DAVID KAHN, *The Codebreakers* (1967; abridged ed., 1974), a comprehensive and meticulously researched history of classical single-key cryptology; C.A. DEAVORS and L. KRUH, *Machine Cryptography and Modern Cryptanalysis* (1985), a highly recommended reference, devoted to the history of cipher machines and their cryptanalysis—especially in the period immediately preceding and during World War II; GUSTAVUS J. SIMMONS (ed.), *Secure Communications and Asymmetric Cryptosystems* (1982), a work on public-key (two-key or asymmetric) cryptography, which was first discussed in the open literature by W. Diffie and M. Hellman in 1976. Modern texts in cryptology accessible to the general reader include DOROTHY DENNING, *Cryptography and Data Security* (1982); and ALAN G. KONHEIM, *Cryptography: A Primer* (1981). Classical literature of the field includes LUIGI SACCO, *Manual of Cryptography* (1938, reissued 1977; originally published in Italian, 1936), at one time described by Kahn as "the world's finest unclassified book on cryptology"; and MARCEL GIVIERGE, *Course in Cryptography* (1934, reissued 1978; originally published in French, 1925); and most of the manuals written by the great U.S. cryptanalyst William F. Friedman. For the most recent information about developments in cryptology there is an American journal, *Cryptologia* (quarterly), devoted entirely to the subject; and a continuing series of annual volumes entitled *Advances in Cryptology,* first published in 1982, which can be highly recommended.

(G.J.Si.)

# The Concept and Components of Culture

Culture may be defined as behaviour peculiar to *Homo sapiens,* together with material objects used as an integral part of this behaviour; specifically, culture consists of language, ideas, beliefs, customs, codes, institutions, tools, techniques, works of art, rituals, ceremonies, and so on. The existence and use of culture depends upon an ability possessed by man alone. This ability has been called variously the capacity for rational or abstract thought, but a good case has been made for rational behaviour among subhuman animals, and the meaning of abstract is not sufficiently explicit or precise. The term symboling has been proposed as a more suitable name for man's unique mental ability, consisting of assigning to things and events certain meanings that cannot be grasped with the senses alone. Articulate speech—language—is a good example. The meaning of the word *dog* is not inherent in the sounds themselves; it is assigned, freely and arbitrarily, to the sounds by human beings. Holy water, "biting one's thumb" at someone (*Romeo and Juliet,* Act I, scene 1), or fetishes are other examples. Symboling is a kind of behaviour objectively definable and should not be confused with symbolizing, which has an entirely different meaning.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 512 and 513.

This article is divided into the following sections:


The concept of culture 874
- Various definitions of culture 874
  - Universalist approaches to culture and the human mind
  - Relativist approaches to sociocultural systems
- Culture and personality 876
- Cultural comparisons 876
  - Ethnocentrism
  - Cultural relativism
  - Evaluative grading
- Cultural adaptation and change 877
  - Ecological or environmental change
  - Diffusion
  - Acculturation
  - Evolution

Approaches to the study of culture 878
- Viewing culture in terms of patterns and configurations 878
  - Cultural traits
  - Cultural areas
  - Cultural types
- Viewing culture in terms of institutional structure and functions 878
  - Social organization
  - Economic systems
  - Education
  - Religion and belief
  - Custom and law

Nonurban cultures 881
- Nomadic societies 881
- Settled hunting and gathering societies 883
- Horticultural societies 883
- Herding societies 884
- Peasant societies 885
  - Historical and geographic survey
  - Types of peasant societies

Urban cultures 888
- Definitions of the city and urban cultures 888
- Types of urban cultures 889
  - Urban cultures before the capitalist world system
  - Urban cultures since the capitalist world system
- Cities and cultures 893

Bibliography 893


## The concept of culture

### VARIOUS DEFINITIONS OF CULTURE

What has been termed the classic definition of culture was provided by the 19th-century English anthropologist Edward Burnett Tylor in the first paragraph of his *Primitive Culture* (1871):

> Culture . . . is that complex whole which includes knowledge, belief, art, morals, law, custom, and any other capabilities and habits acquired by man as a member of society.

In *Anthropology* (1881) Tylor made it clear that culture, so defined, is possessed by man alone. This conception of culture served anthropologists well for some 50 years. With the increasing maturity of anthropological science, further reflections upon the nature of their subject matter and concepts led to a multiplication and diversification of definitions of culture. In *Culture: A Critical Review of Concepts and Definitions* (1952), U.S. anthropologists A.L. Kroeber and Clyde Kluckhohn cited 164 definitions of culture, ranging from "learned behaviour" to "ideas in the mind," "a logical construct," "a statistical fiction," "a psychic defense mechanism," and so on. The definition—or the conception—of culture that is preferred by Kroeber and Kluckhohn and also by a great many other anthropologists is that culture is an abstraction or, more specifically, "an abstraction from behaviour."

These conceptions have defects or shortcomings. The existence of behavioral traditions—that is, patterns of behaviour transmitted by social rather than by biologic hereditary means—has definitely been established for nonhuman animals. "Ideas in the mind" become significant in society only as expressed in language, acts, and objects. "A logical construct" or "a statistical fiction" is not specific enough to be useful. The conception of culture as an abstraction led, first, to a questioning of the reality of culture (inasmuch as abstractions were regarded as imperceptible) and, second, to a denial of its existence; thus, the subject matter of nonbiological anthropology, "culture," was defined out of existence, and without real, objective things and events in the external world there can be no science.

Kroeber and Kluckhohn were led to their conclusion that culture is an abstraction by reasoning that if culture is behaviour it, ipso facto, becomes the subject matter of psychology; therefore, they concluded that culture "is an abstraction from concrete behavior but is not itself behavior." But what, one might ask, is an abstraction of a marriage ceremony or a pottery bowl, to use Kroeber and Kluckhohn's examples? This question poses difficulties that were not adequately met by these authors. A solution was perhaps provided by Leslie A. White in the essay "The Concept of Culture" (1959). The issue is not really whether culture is real or an abstraction, he reasoned; the issue is the context of the scientific interpretation.

Importance of the context

When things and events are considered in the context of their relation to the human organism, they constitute behaviour; when they are considered not in terms of their relation to the human organism but in their relationship to one another, they become culture by definition. The mother-in-law taboo is a complex of concepts, attitudes, and acts. When one considers them in their relationship to the human organism—that is, as things that the organism does—they become behaviour by definition. When, however, one considers the mother-in-law taboo in its relationship to the place of residence of a newly married couple, to the customary division of labour between the sexes, to their respective roles in the society's mode of

subsistence and offense and defense, and these in turn to the technology of the society, the mother-in-law taboo becomes, again by definition, culture. This distinction is precisely the one that students of words have made for many years. When words are considered in their relationship to the human organism—that is, as acts—they become behaviour. But when they are considered in terms of their relationship to one another—producing lexicon, grammar, syntax, and so forth—they become language, the subject matter not of psychology but of the science of linguistics. Culture, therefore, is the name given to a class of things and events dependent upon symboling (*i.e.,* articulate speech) that are considered in a kind of extrahuman context.

**Universalist approaches to culture and the human mind.** Culture, as noted above, is due to an ability possessed by man alone. The question of whether the difference between the mind of man and that of the lower animals is one of kind or of degree has been debated for many years, and even today reputable scientists can be found on both sides of this issue. But no one who holds the view that the difference is one of degree has adduced any evidence to show that nonhuman animals are capable, to any degree whatever, of a kind of behaviour that all human beings exhibit. This kind of behaviour may be illustrated by the following examples: remembering the sabbath to keep it holy, classifying one's relatives and distinguishing one class from another (such as uncles from cousins), defining and prohibiting incest, and so on. There is no reason or evidence that leads one to believe that any animal other than man can have or be brought to any appreciation or comprehension whatever of such meanings and acts. There is, as Tylor argued long ago, a "mental gulf that divides the lowest savage from the highest ape" (*Anthropology*).

Symboling

In line with the foregoing distinction, human behaviour is to be defined as behaviour consisting of, or dependent upon, symboling rather than upon anything else that *Homo sapiens* does; coughing, yawning, stretching, and the like are not human.

Next to nothing is yet known about the neuroanatomy of symboling. Man is characterized by a very large brain, considered both absolutely and relatively, and it is reasonable—and even obligatory—to believe that the central nervous system, especially the forebrain, is the locus of the ability to symbol. But how it does this and with what specific mechanisms remain to be discovered. One is thus led to the conclusion that at some point in the evolution of primates a threshold was reached in some line, or lines, when the ability to symbol was realized and made explicit in overt behaviour. There is no intermediate stage, logical or neurological, between symboling and nonsymboling; an individual or a species is capable of symboling, or he or it is not. The life of Helen Keller makes this clear: when, through the aid of her teacher, Anne Sullivan, Keller was enabled to escape from the isolation to which her blindness and deafness had consigned her and to effect contact with the world of human meanings and values, the transformation was instantaneous.

*Evolution of "minding."* But even if almost nothing is known about the neuroanatomy of symboling, a great deal is known about the evolution of mind (or "minding," if mind is considered as a process rather than a thing), in which one finds symboling as the characteristic of a particular stage of development. The evolution of minding can be traced in the following sequence of stages. First is the simple reflexive stage, in which behaviour is determined by the intrinsic properties of both the organism and the thing reacted to—for example, the contraction of the pupil of the eye under increased stimulation by light. Second is the conditioned reflex stage, in which the response is elicited not by properties intrinsic in the stimulus but by meanings that the stimulus has acquired for the responding organism through experience—for example, Pavlov's dog's salivary glands responding to the sound of a bell. Third is the instrumental stage, as exemplified by a chimpanzee knocking down a banana with a stick. Here the response is determined by the intrinsic properties of the things involved (banana, stick, chimpanzee's neurosensory-muscular system); but a new element has been introduced into behaviour, namely, the exercise of control by the reacting organism over things in the external world. And, finally, there is the symbol stage, in which the configuration of behaviour involves nonintrinsic meanings, as has already been suggested.

Increased security and expanse of life

These four stages exhibit a characteristic of the evolution of all living things: a movement in the direction of making life more secure and enduring. In the first stage the organism distinguishes between the beneficial, the injurious, and the neutral, but it must come into direct contact with the object or event in question to do so. In the second stage the organism may react at a distance, as it were—that is, through an intermediate stimulus. The conditioned reflex brings signs into the life process; one thing or event may serve as an indication of something else—food, danger, and so forth. And, since anything can serve as a sign of anything else (a green triangle can mean food, sex, or an electric shock to the laboratory rat), the reactions of the organism are emancipated from the limitations that stage one imposes upon living things, namely, the intrinsic properties of things. The possibility of obtaining life-sustaining things and of avoiding life-destroying things is thus much enhanced, and the security and continuity of life are correspondingly increased. But in stage two the organism still plays a subordinate role to the external world; it does not and cannot determine the significance of the intermediary stimulus: the bark of a distant dog to the rabbit or the sound of the bell to Pavlov's dog. This meaning is determined by things and events in the external world (or in the laboratory by the experimenter). In stages one and two, therefore, the organism is at the mercy of the external world in this respect.

In the third stage the element of control over environment is introduced. The ape who obtains food by means of a stick (tool) is not subordinate to his situation. He does not merely undergo a situation; he dominates it. His behaviour is not determined by the juxtaposition of things and events; on the contrary, the juxtaposition is determined by the ape. He is confronted with alternatives, and he makes choices. The configuration of behaviour in stage three is constructed within the dynamic organism of the ape and then imposed upon the external world.

The evolution of minding is a cumulative process; the achievements of each stage are carried on into the succeeding one or ones. The fourth stage reintroduces the factor of nonintrinsic meanings to the advances made in stages two and three. Stage four is the stage of symboling, of articulate speech. Thus, one observes two aspects of the evolution of minding, both of which contribute to the security and survivability of life: the emancipation of behaviour from limitations imposed upon it by the external world and increased control over the environment. To be sure, neither emancipation nor control becomes complete, but quantitative increase is significant.

*Evolution of culture.* The direction of biologic evolution toward greater expansion and security of life can be seen from another point of view: the advance from instinctive behaviour (*i.e.,* responses determined by intrinsic properties of the organism) to learned and freely variable behaviour, patterns of which may be acquired and transmitted from one individual and generation to another, and finally to a system of things and events, the essence of which is meanings that cannot be comprehended by the senses alone. This system is, of course, culture, and the species is the human species. Culture is a man-made environment, brought into existence by the ability to symbol.

Once established, culture has a life of its own, so to speak; that is, it is a continuum of things and events in a cause and effect relationship; it flows down through time from one generation to another. Since its inception 1,000,000 or more years ago, this culture—with its language, beliefs, tools, codes, and so on—has had an existence external to each individual born into it. The function of this external, man-made environment is to make life secure and enduring for the society of human beings living within the cultural system. Thus, culture may be seen as the most recent, the most highly developed means of promoting the security and continuity of life, in a series that began with the simple reflex.

Absence of culture in primate societies

Society preceded culture; society, conceived as the interaction of living beings, is coextensive with life itself. Man's immediate prehuman ancestors had societies, but they did not have culture. Studies of monkeys and apes have greatly enlarged scientific knowledge of their social life—and, by inference, the scientific conception of the earliest human societies. Data derived from paleontological sources and from accumulating studies of living, nonhuman primates are now fairly abundant, and hypotheses derived from these are numerous and varied in detail. A fair summary of them may be made as follows: The growth of the primate brain was stimulated by life in the trees, specifically, by eye-hand coordinations involved in swinging from limb to limb and by manipulating food with the hands (as among the insectivorous lemurs). Descent to the ground, as a consequence of deforestation or increase in body size (which would tend to restrict arboreal locomotion and increase the difficulty of obtaining enough food to supply increased need), and the assumption of erect posture were other significant steps in biologic evolution and the eventual emergence of culture. Some theories reject the arboreal stage in man's evolutionary past, but this does not seriously affect the overall conception of his development.

The Australopithecines of Africa, extinct manlike higher primates about which reliable knowledge is very considerable today, exemplify the stage of erect posture in primate evolution. Erect posture freed the arms and hands from their earlier function of locomotion and made possible an extensive and versatile use of tools. Again, the eye-hand-object coordinations involved in tool using stimulated the growth of the brain, especially the forebrain. It is not possible to determine on the basis of paleontological evidence the precise point at which the ability to symbol (specifically, articulate speech) was realized, as expressed in overt behaviour. It is believed by some that man's prehuman ancestors used tools habitually and that habit became custom through the transmission of tool using from one generation to another long before articulate speech came into being. In fact, some theorists hold, the customary use of tools became a powerful stimulus in the development of a brain that was capable of symboling or articulate speech.

Immediate effects of symboling

The introjection of symboling into primate social life was revolutionary. Everything was transformed, everything acquired new meaning; the symbol added a new dimension to primate—now human—existence. An ax was no longer merely a tool with which to chop; it could become a symbol of authority. Mating became marriage, and all social relationships between parents and children and brothers and sisters became moral obligations, duties, rights, and privileges. The world of nature, from the stones beside the path to the stars in their courses, became alive and conscious spirits. "And all that I beheld respired with inward meaning" (Wordsworth). The anthropoid had at last become a man.

**Relativist approaches to sociocultural systems.** Thus far in this article, culture has been considered in general, as the possession of all mankind. Now it is appropriate to turn to particular cultures, or sociocultural systems. Human beings, like other animal species, live in societies, and each society possesses culture. It has long been customary for ethnologists to speak of Seneca culture, Eskimo culture, North American Plains culture, and so on—that is, the culture of a particular society (Seneca) or an indefinite number of societies (Eskimo) or the cultures found in or characteristic of a topographic area (the North American Plains). There is no objection to this usage as a convenient means of reference: "Seneca culture" is the culture that the Seneca tribe possesses at a particular time. Similarly, Eskimo culture refers to a class of cultures, and Plains culture refers to a type of culture. What is needed is a term that defines culture precisely in its particular manifestations for the purpose of scientific study, and for this the term sociocultural system has been proposed. It is defined as the culture possessed by a distinguishable and autonomous group (society) of human beings, such as a tribe or a modern nation. Cultural elements may pass freely from one system to another (cultural diffusion), but the boundary provided by the distinction between one system and another (Seneca, Cayuga; United States, Japan) makes it possible to study the system at any given time or over a period of time.

Every human society, therefore, has its own sociocultural system: a particular and unique expression of human culture as a whole. Every sociocultural system possesses the components of human culture as a whole—namely, technological, sociological, and ideological elements. But sociocultural systems vary widely in their structure and organization. These variations are attributable to differences among physical habitats and the resources that they offer or withhold for human use; to the range of possibilities inherent in various areas of activity, such as language or the manufacture and use of tools; and to the degree of development. The biologic factor of man may, for purposes of analysis and comparison of sociocultural systems, be considered as a constant. Although the equality or inequality of races, or physical types, of mankind has not been established by science, all evidence and reason lead to the conclusion that, whatever differences of native endowment may exist, they are insignificant as compared with the overriding influence of the external tradition that is culture.

## CULTURE AND PERSONALITY

The powerful influence of culture

Since the infant of the human species enters the world cultureless, his behaviour—his attitudes, values, ideals, and beliefs, as well as his overt motor activity—is powerfully influenced by the culture that surrounds him on all sides. It is almost impossible to exaggerate the power and influence of culture upon the human animal. It is powerful enough to hold the sex urge in check and achieve premarital chastity and even voluntary vows of celibacy for life. It can cause a person to die of hunger, though nourishment is available, because some foods are branded unclean by the culture. And it can cause a person to disembowel or shoot himself to wipe out a stain of dishonour. Culture is stronger than life and stronger than death. Among subhuman animals, death is merely the cessation of the vital processes of metabolism, respiration, and so on. In the human species, however, death is also a concept; only man knows death. But culture triumphs over death and offers man eternal life. Thus, culture may deny satisfactions on the one hand while it fulfills desires on the other.

The predominant emphasis, perhaps, in studies of culture and personality has been the inquiry into the process by which the individual personality is formed as it develops under the influence of its cultural milieu. But the individual biologic organism is itself a significant determinant in the development of personality. The mature personality is, therefore, a function of both biologic and cultural factors, and it is virtually impossible to distinguish these factors from each other and to evaluate the magnitude of each in particular cases. If the cultural factor were a constant, personality would vary with the variations of the neurosensory-glandular-muscular structure of the individual. But there are no tests that can indicate, for example, precisely how much of the taxicab driver's ability to make change is due to innate endowment and how much to cultural experience. Therefore, the student of culture and personality is driven to work with "modal personalities," that is, the personality of the typical Crow Indian or the typical Frenchman insofar as this can be determined. But it is of interest, theoretically at least, to note that even if both factors, the biologic and the cultural, were constant—which they never are in actuality—variations of personality would still be possible. Within the confines of these two constants, individuals might undergo a number of profound experiences in different chronological permutations. For example, two young women might have the same experiences of (1) having a baby, (2) graduating from college, and (3) getting married. But the effect of sequence (1), (2), (3) upon personality development would be quite different than that of sequence (2), (3), (1).

## CULTURAL COMPARISONS

**Ethnocentrism.** Ethnocentrism is the name given to a tendency to interpret or evaluate other cultures in terms of one's own. This tendency has been, perhaps, more prevalent in modern nations than among preliterate tribes.

The citizens of a large nation, especially in the past, have been less likely to observe people in another nation or culture than have been members of small tribes who are well acquainted with the ways of their culturally diverse neighbours. Thus, the American tourist could report that Londoners drive "on the wrong side of the street" or an Englishman might find some customs on the Continent "queer" or "boorish," merely because they are different. Members of a Pueblo tribe in the American Southwest, on the other hand, might be well acquainted with cultural differences not only among other Pueblos but also in non-Pueblo tribes such as the Navajo and Apache.

Ethnocentrism became prominent among many Europeans after the discovery of the Americas, the islands of the Pacific, and the Far East. Even anthropologists might characterize all preliterate peoples as being without religion (as did Sir John Lubbock) or as having a "prelogical mentality" (as did Lucien Lévy-Bruhl) merely because their ways of thinking did not correspond with those of the culture of western Europe. Thus, inhabitants of non-Western cultures, particularly those lacking the art of writing, were widely described as being immoral, illogical, queer, or just perverse ("Ye Beastly Devices of ye Heathen").

**Cultural relativism.** Increased knowledge led to or facilitated a deeper understanding and, with it, a finer appreciation of cultures quite different from one's own. When it was understood that universal needs could be served with culturally diverse means, that worship might assume a variety of forms, that morality consists in conforming to ethical rules of conduct but does not inhere in the rules themselves, a new view emerged that each culture should be understood and appreciated in terms of itself. What is moral in one culture might be immoral or ethically neutral in another. For example, it was not immoral to kill a baby girl at birth or an aged grandparent who was nonproductive when it was impossible to obtain enough food for all; or wife lending among the Eskimo might be practiced as a gesture of hospitality, a way of cementing a friendship and promoting mutual aid in a harsh and dangerous environment, and thus may acquire the status of a high moral value.

The view that elements of a culture are to be understood and judged in terms of their relationship to the culture as a whole—a doctrine known as cultural relativism—led to the conclusion that the cultures themselves could not be evaluated or graded as higher and lower, superior or inferior. If it was unwarranted to say that patriliny (descent through the male line) was superior or inferior to matriliny (descent through the female line), if it was unjustified or meaningless to say that monogamy was better or worse than polygamy, then it was equally unsound or meaningless to say that one culture was higher or superior to another. A large number of anthropologists subscribed to this view; they argued that such judgments were subjective and therefore unscientific.

The question of subjective judgments

It is, of course, true that some values are imponderable and some criteria are subjective. Are people in modern Western culture happier than the Aborigines of Australia? Is it better to be a child than an adult, alive than dead? These certainly are not questions for science. But to say that the culture of the ancient Mayas was not superior to or more highly developed than the crude and simple culture of the Tasmanians or to say that the culture of England in 1966 was not higher than England's culture in 1066 is to fly in the face of science as well as of common sense.

**Evaluative grading.** Cultures have ponderable values as well as imponderable, and the imponderable ones can be measured with objective, meaningful yardsticks. A culture is a means to an end: the security and continuity of life. Some kinds of culture are better means of making life secure than others. Agriculture is a better means of providing food than hunting and gathering. The productivity of human labour has been increased by machinery and by the utilization of the energy of nonhuman animals, water and wind power, and fossil fuels. Some cultures have more effective means of coping with disease than others, and this superiority is expressed mathematically in death rates. And there are many other ways in which meaningful differences can be measured and evaluations made. Thus, the proposition that cultures have ponderable values that can be measured meaningfully by objective yardsticks and arranged in a series of stages, higher and lower, is substantiated. But, it should be noted, this is not equivalent to saying that man is happier or that the dignity of the individual (an imponderable) is greater in an industrialized or agricultural sociocultural system than in one supported by human labour alone and sustained wholly by wild foods.

Actually, however, there is no necessary conflict between the doctrine of cultural relativism and the thesis that cultures can be objectively graded in a scientific manner. It is one thing to reject the statement that monogamy is better than polygamy and quite another to deny that one kind of sociocultural system contains a better means of providing food or combating disease than another.

## CULTURAL ADAPTATION AND CHANGE

**Ecological or environmental change.** Every sociocultural system exists in a natural habitat, and, of course, this environment exerts an influence upon the cultural system. The cultures of some Eskimo groups present remarkable instances of adaptation to environmental conditions: tailored fur clothing, snow goggles, boats and harpoons for hunting sea mammals, and, in some instances, hemispherical snow houses, or igloos. Some sedentary, horticultural tribes of the upper Missouri River went out into the Great Plains and became nomadic hunters after the introduction of the horse. The culture of the Navajos underwent profound change after they acquired herds of sheep and a market for their rugs was developed. The older theories of simple environmentalism, some of which maintained that even styles of myths and tales were determined by topography, climate, flora, and other factors, are no longer in vogue. The present view is that the environment permits, at times encourages, and also prohibits the acquisition or use of certain cultural traits but otherwise does not determine culture change. The Fuegians living at the southern tip of South America, as viewed by Charles Darwin on his voyage on the *Beagle,* lived in a very cold, harsh environment but were virtually without both clothing and dwellings.

**Diffusion.** "Culture is contagious," as a prominent anthropologist once remarked, meaning that customs, beliefs, tools, techniques, folktales, ornaments, and so on may diffuse from one people or region to another. To be sure, a culture trait must offer some advantage, some utility or pleasure, to be sought and accepted by a people. (Some anthropologists have assumed that basic features of social structure, such as clan organization, may diffuse, but a sounder view holds that these features involving the organic structure of the society must be developed within societies themselves.) The degree of isolation of a sociocultural system—brought about by physical barriers such as deserts, mountain ranges, and bodies of water—has, of course, an important bearing upon the ease or difficulty of diffusion. Within the limits of desirability on the one hand and the possibility of communication on the other, diffusion of culture has taken place everywhere and in all times. Archaeological evidence shows that amber from the Baltic region diffused to the Mediterranean coast; and, conversely, early coins from the Middle East found their way to northern Europe. In aboriginal North America, copper objects from northern Michigan have been found in mounds in Georgia; macaw feathers from Central America turn up in archaeological sites in northern Arizona. Some Indian tribes in northwestern regions of the United States had possessed horses, originally brought into the Southwest by Spanish explorers, years before they had ever even seen white men. The wide dispersion of tobacco, corn (maize), coffee, the sweet potato, and many other traits are conspicuous examples of cultural diffusion.

Evidence of diffusion in the ancient past

**Acculturation.** Diffusion may take place between tribes or nations that are approximately equal in political and military power and of equivalent stages of cultural development, such as the spread of the sun dance among the Plains tribes of North America. But in other instances, it takes place between sociocultural systems differing widely in this respect. Conspicuous examples of this have been

instances of conquest and colonization of various regions by the nations of modern Europe. In these cases it is often said that the culture of the more highly developed nation is "imposed" upon the less developed peoples and cultures, and there is, of course, much truth in this; the acquisition of foreign culture by the subject people is called acculturation and is manifested by the indigenous populations of Latin America as well as of other regions. But even in cases of conquest, traits from the conquered peoples may diffuse to those of the more advanced cultures; examples might include, in addition to the cultivated plants cited above, individual words (*coyote*), musical themes, games, and art motifs.

One of the major problems of ethnology during the latter half of the 19th and the early decades of the 20th centuries was the question "How are cultural similarities in noncontiguous regions to be explained?" Did the concepts of pyramid building, mummification, and sun worship originate independently in ancient Egypt and in the Andean highlands and in Yucatán or did these traits originate in Egypt and diffuse from there to the Americas, as some anthropologists have believed? Some schools of ethnological theory have held to one view, some, to another. The 19th-century classical evolutionists (which included Edward Burnett Tylor and Lewis H. Morgan, among others) held that the mind of man is so constituted or endowed that he will develop cultures everywhere along the same lines. "Diffusionists"—those, such as Fritz Graebner and Elliot Smith, who offered grand theories about the diffusion of traits all over the world—maintained that man was inherently uninventive and that culture, once created, tended to spread everywhere. Each school tended to insist that its view was the correct one, and it would continue to hold that view unless definite proof of the contrary could be adduced.

The tendency nowadays is not to side categorically with one school as against another but to decide each case on its own merits. The consensus with regard to pyramids is that they were developed independently in Egypt and the Americas because they differ markedly in structure and function: the Egyptian pyramids were built of stone blocks and contained tombs within their interiors. The American pyramids were constructed of earth, then faced with stone, and they served as the bases of temples. The verdict with regard to the bow and arrow is that it was invented only once and subsequently diffused to all regions where it has been found. The probable antiquity of the origin of fire making, however, and the various ways of generating it—by percussion, friction, compression (fire pistons)—indicate multiple origins.

**Evolution.** Evolution of culture—that is, the development of forms through time—has taken place. No amount of diffusion of picture writing could of itself, for instance, produce the alphabetic system of writing; as Tylor demonstrated so well, the art of writing has developed through a series of stages, which began with picture writing, progressed to hieroglyphic writing, and culminated in alphabetic writing. In the realm of social organization there was a development from territorial groups composed of families to segmented societies (clans and larger groupings). Sociocultural evolution, like biologic evolution, exhibits a progressive differentiation of structure and specialization of function.

Relationship between diffusion and evolution

A misunderstanding has arisen with regard to the relationship between evolution and diffusion. It has been argued, for example, that the theory of cultural evolution was unsound because some peoples skipped a stage in a supposedly determined sequence; for example, some African tribes, as a consequence of diffusion, went from the Stone Age to the Iron Age without an intermediate age of copper and bronze. But the classical evolutionists did not maintain that peoples, or societies, had to pass through a fixed series of stages in the course of development, but that tools, techniques, institutions—in short, culture—had to pass through the stages. The sequence of stages of writing did not mean that a society could not acquire the alphabet without working its way through hieroglyphic writing; it was obvious that many peoples did skip directly to the alphabet.

## Approaches to the study of culture

### VIEWING CULTURE IN TERMS OF PATTERNS AND CONFIGURATIONS

**Cultural traits.** The concept of culture embraces the culture of mankind as a whole. An understanding of human culture is facilitated, however, by analyzing "the complex whole" into component parts or categories. In somewhat the same sense that the atom has been regarded as the unit of matter, the cell as the unit of life, so the culture trait is generally regarded as the unit of culture. A trait may be an object (knife), a way of doing something (weaving), a belief (in spirits), or an attitude (the so-called horror of incest). But, within the category of culture, each trait is related to other traits. A distinguishable and relatively self-contained cluster of traits is conventionally called a culture complex. The association of traits in a complex may be of a functional and mechanical nature, such as horse, saddle, bridle, quirt, and the like, or it may lie in conceptional or emotional associations, such as the acts and attitudes involved in seclusion in a menstrual hut or retrieving a heart that has been stolen by witches.

**Cultural areas.** The relationship between an actual culture and its habitat is always an intimate one, and therefore one finds a more or less close correlation between kind of habitat and type of culture. This results in the concept of culture area. This conception goes back at least as far as the early 19th century, but it was first brought into prominence by the U.S. anthropologist Clark Wissler in *The American Indian* (1917) and *Man and Culture* (1923). He divided the Indian cultures (as they were in the latter half of the 19th century) into geographic cultural regions: the Caribou area of northern Canada; the Northwest coast, characterized by the use of salmon and cedar; the Great Plains, where tribes hunted bison with the horse; the Pueblo area of the Southwest; and so on. Others later distinguished culture areas in other continents.

**Cultural types.** Appreciation of the relationship between culture and topographic area suggests the concept of culture type, such as hunting and gathering or a special way of hunting—for example, the use of the horse in bison hunting in the Plains or the method of hunting of sea mammals among the Eskimo; pastoral cultures centred upon sheep, cattle, reindeer, and so on; and horticulture (with digging stick and hoe) and agriculture (with ox-drawn plow). Less common are trading cultures such as are found in Melanesia or specialized production of some object for trade, such as pottery, bronze axes, or salt, as was the case in Luzon.

Culture configurations and patterns

Configuration and pattern, especially the latter, are concepts closely related to culture area and culture type. All of them have one thing in common; they view culture not in terms of its individual components, or traits, but as meaningful organizations of traits: areas, occupations, configurations (art, mathematics, physics), or patterns (in which psychological factors are the bases of organization). Clark Wissler's "universal culture pattern" was a recognition of the fact that all particular and actual cultures possess the same general categories: language, art, social organization, religion, technology, and so on.

### VIEWING CULTURE IN TERMS OF INSTITUTIONAL STRUCTURE AND FUNCTIONS

**Social organization.** A sociocultural system presents itself under two aspects: structure and function. As culture evolves, sociocultural systems (like biologic systems) become more differentiated structurally and more specialized functionally, proceeding from the simple to the complex. Systems on the lowest stage of development have only two significant kinds of parts: the local territorial group and the family. There is a corresponding minimum of specialization, limited, with but few exceptions, to division of function, or labour, along sex lines and to division between children and adults. The exceptions are headmen and shamans; they are special organs, so to speak, in the body politic. The headman is a mechanism of social integration, direction, and control, expressing, however, the consensus of the band. The shaman, though a self-appointed priest or magician, is also an instrument of

society; he may be regarded as the first specialist in the history of human society.

Classes and segments

All human societies are divided into classes and segments. Class is defined as one of an indefinite number of groupings each of which differs in composition from the other or others, such as men and women; married, widowed, and divorced; children and adults. Segment is defined as one of an indefinite number of groupings all of which are alike in structure and function: families, lineages, clans, and so on. On more advanced levels of development there are occupational classes, such as farmers, pastoralists, artisans, metalworkers, and scribes, and territorial segments, such as wards, barrios, counties, and states.

Segmentation is a cultural process essential to the evolution of culture; it is a means of increasing the size of a society or a grouping within a sociocultural system (such as an army) and therefore of increasing its power to make life secure, without suffering a corresponding loss of effectiveness through diminished solidarity; segmentation is a means of maintaining solidarity at the same time that it enlarges the social grouping. A tribe could not increase in size beyond a certain point without resorting to segmentation: the formation of lineages, clans, and the like. The word clannish points to one of the functions of segments in general: the fostering of solidarity. Tribes become segments in confederacies; and above the tribal level, the evolution of civil society employs barrios, demes, counties, and states in its process of segmentation. In present-day society, the army and the church offer illuminating examples of increased size and sustained solidarity proceeding hand in hand.

**Economic systems.** Division of labour along occupational lines is rare, although not wholly lacking, in preliterate societies—despite a widespread notion that one member of a tribe specializes in making arrows, which he exchanges for moccasins made by another specialist. Occupational groupings were virtually lacking in all cultural systems of aboriginal North America, for example. Guilds of metalworkers are found in some African tribes and specialists in canoe making and tattooing existed in Polynesia. But it is not until the transition from preliterate society, based upon ties of kinship, to civil society, based upon property relations and territorial distinctions (the state), that division of labour along occupational lines becomes extensive. On this level there are found many kinds of specialists: metalworkers, scribes, astrologers, soldiers, dancers, musicians, alchemists, prostitutes, eunuchs, and so forth.

Distribution and exchange of goods

Production of goods is everywhere followed by distribution and exchange. Among the Kurnai of Australia, for example, game was divided and distributed as follows: the hunter who killed a wallaby, for example, kept the head; his father received the ribs on the right side, his mother the ribs on the left side, plus the backbone, and so on; the various parts of the animal went to various classes of relatives in accordance with fixed, traditional rules.

Distribution along kinship lines constitutes a system of circulation and exchange within the tribe as a whole, for everyone is a relative of everyone else. It takes the form of bestowing gifts to relatives on all sorts of occasions—such as birth, initiation, marriage, death. In some cases there is an exchange of goods on the spot, but more often A gives something to B who gives A a gift at a later date. All this takes place in the network of rights and obligations among kindred; one has both an obligation to give and a right to receive on certain occasions and in certain contexts. The whole process is one of mutual aid and cooperation.

The consequence of this form of distribution and exchange is that the recipient receives kinds of things that he already has; each household has the same kinds of foods, utensils, ornaments, and other things that every other household has. Why, then, it might be asked, does this form of exchange take place? Two reasons may be distinguished. First, this kind of exchange fortifies ties of kinship and mutual aid—as neighbourhood exchange among households in modern American culture initiates friendships that in times of need constitute mutual aid. Second, this system of circulation of goods is in effect a system of social security: a household in need, due to illness or accident, receives help from the community ("No household can starve as long as others have corn," as the Iroquois put it). Here we have an economic system subordinated to the welfare of the society as a whole.

Exchange or circulation of goods and services (a basket is the material form of "a service," that is, human labour) must, of course, take place in sociocultural systems where division of labour finds expression in specialization: the ironworker must obtain food; the horticulturalist needs an iron hoe.

Exchange of goods between sociocultural systems is universal and takes place on the lowest levels of cultural development. In some instances it is the only form of nonhostile communication: in the so-called silent trade the actual exchange takes place in a neutral zone without the presence of the participating parties. Archaeological evidence shows that intergroup exchange occurred in remote times and over great distances, as already noted above in the discussion of diffusion.

An interesting form of the circulation of goods—usually referred to as redistribution—occurs among more highly developed tribes. The head of the sociopolitical system, that is, the chief or priest-chief, imposes levies upon all households, thus acquiring a large amount of goods—food, utensils, art objects, and so on—which he then redistributes to the households of the tribe. This may take the form and occasion of ceremonies and feasts or distribution may be made in cases of need. This widespread and interesting form of redistribution serves the same ends as those served by distribution as a function of the kinship system, namely, fostering solidarity and social security—an equitable distribution that tends to iron out inequalities among households.

Ownership and use

Some economic concepts in modern Western culture do not correspond closely with conceptions and customs in many preliterate societies. Ownership is a case in point. Complete possession of and exclusive right to use something in an economic context, such as land, a dwelling, or a boat, is rare, if not wholly lacking, in preliterate societies (although one might have exclusive rights to a dream, spell, or charm). In general, one has merely the right to use or occupy a tract of land or a house; when its use has terminated, anyone can take it over. In some societies it might be said that a boat "belonged" to the men who made it or even to the individual who initiated its construction. But anyone else in the community would have the right to use it when the "owners" (the men who made it) were not using it. It is the right to use, rather than exclusive and absolute possession, that is significant; there is no such thing as absentee ownership in primitive society.

A band or tribe "holds" the land it occupies; here again, it is tenure rather than ownership that is significant; the land "belongs" to Nature, or Mother Earth; people merely hold and use it. There is usually an intimate relationship between the people and "their" land. Navajo Indians fell on their knees and kissed the earth when they were returned to their former territory after forcible detention in an alien land. Land is defended against outsiders, except when they are accepted as guests, but the significant thing is not that the outsiders do not own the land but that they pose a threat to those who occupy it.

In some tribes there is a distinct conception that the land held "belongs" to the tribe, the chief of which allots plots or tracts to individuals or households for their use. But when use terminates, the land reverts to the tribal domain.

During the latter part of the 19th century there was considerable discussion of "primitive communism." This doctrine came to be interpreted as meaning that private property, the private right to hold or use, was nonexistent in primitive society. It was extended also to communism in wives and children in some tribes; this was interpreted to be a vestige of a former stage of "primordial promiscuity." Many ethnologists, however, launched a vigorous attack upon "the doctrine of primitive communism." Some of the conceptions of earlier anthropologists—such as group marriage—were shown to be unwarranted in the light of later research.

Today, with these polemics well in the past, the situa-

tion with regard to property rights in tribal societies may be summarized as follows. Tenure and use, rather than ownership in fee simple, were the significant concepts and practices. Private, or personal, possession of goods and use of land were recognized. But possession and right were qualified by the rights and obligations of kinship: one had an obligation both to give and to receive within the body of kindred, according to specific rules. In a de facto sense, things belonged to the body of kindred; rights of possession and use were regulated by customs of kinship. In some cultures a borrower was not obliged to return an object borrowed, on the theory that if a person could afford to lend something, he relinquished his right to its possession. The mode of life in preliterate society, based upon kinship and functioning in accordance with the principles of cooperation and mutual aid, did indeed justify the adjective communal; it was the noun communism that was resented—if not feared—because of its Marxist connotation.

One of the most important, as well as characteristic, features of the economic life of preliterate societies, as contrasted with modern civilizations, is this: no individual and no class or group in tribal society was denied access to the resources of nature; all were free to exploit them. This is, of course, in sharp contrast to civil society in which private ownership by some, or a class, is the means of excluding others—slaves, serfs, a proletariat—from the exploitation and enjoyment of the resources of nature. It is this freedom of access, the freedom to exploit and to enjoy the resources of nature, that has given primitive society its characteristics of freedom and equality. And, being based upon kinship ties, it had fraternity as well (see also ECONOMIC SYSTEMS).

Socialization or enculturation

**Education.** In the human species individuals are equipped with fewer instincts than is the case in many nonhuman species. And, as already noted, they are born cultureless. Therefore an infant *Homo sapiens* must learn a very great deal and acquire a vast number of conditioned reflexes and habit patterns in order to live effectively, not only in society but in a particular kind of sociocultural system, be it Tibetan, Eskimo, or French. This process, taken as a whole, is called socialization (occasionally, enculturation)—the making of a social being out of one that was at birth wholly individualistic and egoistic.

Education in its broadest sense may properly be regarded as the process by which the culture of a sociocultural system is impressed or imposed upon the plastic, receptive infant. It is this process that makes continuity of culture possible. Education, formal and informal, is the specific means of socialization. By informal education is meant the way a child learns to adapt his behaviour to that of others, to be like others, to become a member of a group. By formal education is meant the intentional and more or less systematic effort to affect the behaviour of others by transmitting elements of culture to them, be it knowledge or belief, patterns of behaviour, or ideals and values. These attempts may be overt or covert. The teacher may make his purpose apparent, even emphatic, to the learner. But much education is effected in an unobtrusive way, without teacher or learner being aware that culture is being transmitted. Thus, in myths and tales, certain characters are presented as heroes or villains; certain traits are extolled, others are deplored or denounced. The impressionable child acquires ideals and values, an image of the good or the bad.

The growing child is immersed in the fountain of informal education constantly; the formal education tends to be periodic. Many sociocultural systems distinguish rather sharply a series of stages in the education and development of full-fledged men and women. First there is infancy, during which perhaps the most profound and enduring influences of a person's life are brought to bear. Weaning ushers in a new stage, that of childhood, during which boys and girls become distinguished from each other. Puberty rites transform children into men and women. These rites vary enormously in emphasis and content. Sometimes they include whipping, isolation, scarification, or circumcision. Very often the ritual is accompanied by explicit instruction in the mythology and lore of the tribe and in ethical codes. Such rituals as confirmation and Bar Mitzvah in modern Western culture belong to the category of puberty rites.

With marriage come instruction and admonition, appropriate to the occasion, from elder relatives and, in more advanced cultures, from priests. In some sociocultural systems men may become members of associations or sodalities: men's clubs, warrior societies, secret societies of magic or medicine. In some cases it is said that in passing through initiation rites a person is "born again." Women also may belong to sodalities, and in some instances they may become members of secret, magical societies along with men.

**Religion and belief.** Man's oldest philosophy is animism, the doctrine that everything is alive and possesses mental faculties like those possessed by man: desire, will, purpose, anger, love, and the like. This philosophy results from man's projection of his own self, his psyche, into other things and beings, inanimate and living, without being aware of this projection. "To the Omaha," wrote anthropologist Alice Fletcher,

> nothing is without life: . . . He projects his own consciousness upon all things and ascribes to them experiences and characteristics with which he is familiar; . . . akin to his own conscious being.
>
> ("Wakonda," in F.W. Hodge [ed.], *Handbook of American Indians North of Mexico*)

Development of religion

"A belief in spirits is," according to Edward Burnett Tylor, "the minimum definition of religion." Some later students, however, made the same claim for a belief in impersonal, supernatural power, or mana (manitou, orenda, and so on). In any case, these two elements of religion are virtually worldwide and undoubtedly represent a very early stage in the development of religion. In some cultures spirits are virtually innumerable, but, in the course of time, the more important spirits become gods. Thus, there has been a tendency toward monotheism in the history of religion. The German Roman Catholic priest and anthropologist Father Wilhelm Schmidt argued not only that some primitive peoples believe in a Supreme Being but that monotheism was characteristic of the earliest and simplest cultures. Schmidt's thesis, however, has been severely criticized by other ethnologists. Also, as Tylor pointed out many years ago, the Supreme Being of some very primitive peoples is an originator god, or a philosophical explanatory device, accountable only for the existence and structure of the world; after his work was completed, he had no further significance; he was not worshiped and played no part in the daily lives of the people.

Magic and religion

In the past there was much discussion—and debate—about the difference between magic and religion. Both were deemed expressions of a belief in the supernatural. Some argued that religion was social (moral) whereas magic was antisocial (immoral). Another distinction was that magic was the use of supernatural power divorced from a spiritual being. The distinction between religion and magic was so beset with exceptions as to render most definitions of these terms logically imperfect. Another difficulty was the tacit assumption that different entities, religion and magic, exist per se, and therefore that "correct" definitions of them must exist also (Adam called the animal a horse because it was a horse). Much confusion and debate would have been obviated if it had been recognized (as it generally is now) that there is no such thing as a "correct" definition—all definitions are man-made and arbitrary—and that the problem is not what religion or magic are but what beliefs, events, and experiences one wishes to designate with the words religion and magic (see also OCCULTISM: *Magic;* RELIGIONS, THE STUDY AND CLASSIFICATION OF).

**Custom and law.** Sociocultural systems, like other kinds of systems, must have means of self-regulation and control in order to persist and function. In human society these means are numerous and varied. The kinship organization specifies reciprocal and correlative rights, duties, and obligations of one class of relatives to another. Codes of ethics govern the relationship of the individual to the well-being of society as a whole. Codes of etiquette regulate class structure by requiring individuals to conform

to their respective classes. Custom is a general term that embraces all these mechanisms of regulation and control and even more. Custom is the name given to uniformities in sociocultural systems. Uniformities are important because they make anticipation and prediction possible; without them, orderly conduct of social life would not be possible. Custom, therefore, is a means of social regulation and control, of effecting compliance with itself in order to render effective conduct of social life possible.

As in the case of religion and magic, much effort and debate have been spent in attempts to achieve a clean-cut distinction between custom and law. There is little or no difficulty when one is concerned with the extremes of the spectrum of social control. The way that a Hopi Indian holds his corn-husk cigarette in his hand is a matter of custom rather than law, as most ethnologists would probably agree. At the other extreme, a state edict prohibiting the manufacture and sale of alcoholic beverages is a law, not a custom. But in other situations the distinction is far from clear, and disagreement with regard to definitions arises. For example, in marriage the obligation to wed someone within a specified group or class (endogamy) or outside a group or class (exogamy) has been called both law and custom. Probably the most useful distinction between custom and law is the following. If an infraction of a social rule or deviation from a norm is punished merely by expressions of social disapproval, gossip, ridicule, or ostracism, the rule is called custom. If, however, infractions or violations are punished by an agency, designated by society and empowered to act on its behalf, then the rule is called a law. But even here there is difficulty. The same kind of offense may be punished by custom in one society, by law in another—as in, for example, adultery, incest, miscegenation, and black magic.

It is the ethnologist, rather than the historian, who is disturbed by instances of ambiguity with regard to custom and law; in preliterate societies the distinction between the two is not always clear. But in civil societies—that is, states brought into being by the agricultural revolution and their more recent successors—the distinction is usually sharper and more apparent, though instances of sumptuary laws that prohibit the wearing of silk or that limit the length of a garment merge law and custom or reinforce the latter by the former.

One need not be unduly disturbed by the difficulty of making sharp distinctions among sociocultural phenomena and of formulating definitions. The phenomena of culture, like those of the external world in general, are what they are, and if man-made concepts and words do not correspond closely with them, one may regret the lack of fit. But it is better to do this than to distort real phenomena by trying to force them into artificial concepts and definitions. (L.A.W.)

## Nonurban cultures

So great are the variations in ways of life, past and present, that comparisons among them are difficult. Any simple classification of human societies and cultures can only be viewed as arbitrary. From a modern urban point of view, nevertheless, there is the obvious distinction between the primitive and the civilized: between simple and complex societies; between tiny and huge social agglomerations; between scattered and dense populations; and, above all, between prestate societies and societies that have developed states. In general, civilization involves the rise of legal institutions and the acquisition of a legal monopoly of force by a government. Those developments made possible the cities and empires of classical times and the growth of dense populations. Thus "civilized" is nearly synonymous with "urban."

Primitive and civilized

The varieties of nonurban, or primitive, societies may be further classified. One way is by the methods they use to get food. Those who hunt and gather behave quite differently, as societies, from herdsmen and mounted predator-warriors, the pastoralists, who in turn live quite differently from the various kinds of agriculturalists. These distinctions are not sharp, for of course there are societies that combine foraging with some agriculture, others, some agriculture and some herding; and, in a few cases, a class of herders may live in the same society with a class or caste of agriculturalists. A continuum of societies may be constructed, ranging from tiny, simple bands of hunter-gatherers in poor environments to large, dense populations of irrigation agriculturalists—that is, from the entirely nomadic to the fully sedentary. The degree to which societies approach the sedentary deserves prominence in any classification since sedentary ways are accompanied by many other cultural traits and institutions.

### NOMADIC SOCIETIES

Throughout 99 percent of the time that *Homo sapiens* has been on Earth, or until about 8,000 years ago, all peoples were foragers of wild food. There were great differences among them; some specialized in hunting big game, fishing, and shellfish gathering, while others were almost completely dependent on the gathering of wild plants. Broadly speaking, however, they probably shared many features of social and political organization, as well as of religions and other ideologies (in form though not in specific content). The hunting-gathering societies declined with the growth of agricultural societies, which either drove them from their territories or assimilated or converted them.

The later rise of the nation-states, especially after the Industrial Revolution in Europe, resulted in the near extermination of hunting-gathering societies. Today, the remaining ones are confined to desert, mountain, jungle, or Arctic wastelands. Some have been studied and described by anthropologists: the central and northern Australians, the Bushmen of the Kalahari in southern Africa, the Pygmies of the central African forests, the Pygmies of the Andaman Islands in the Indian Ocean, the Ona and Yahgan Indians of southern South America, the "Digger" Indians of Nevada, the Indians of the northern Canadian forests, and the Canadian, Alaskan, and Greenland Eskimos.

All of these peoples inhabit areas representing almost every extreme in climate and environment, but they have one thing in common: their marginality to, or relative isolation from, modern economic systems. Their techniques and forms of acquiring food vary greatly. The Eskimos, for example, are entirely dependent on hunting and fishing; the African San (Bushmen), the Australian Aborigines, and the Nevada Indians are chiefly dependent on the gathering of seeds, nuts, and tubers.

The significance of nomadism to the student of primitive cultures may be suggested by a comparison of the Ona and Yámana (Yahgan) of Tierra del Fuego. The Ona inhabit the interior forests and depend heavily on hunting guanaco (a small New World camel). The Yámana are canoe-using fishermen and shellfish gatherers. Yet, despite their utterly different ecological adaptation, the two Indian societies have cultures that are so similar that anthropologists conventionally group them with the neighbouring Chono and Alakaluf of Chile into one "Fuegian culture area." They are all nomadic, though the Ona are "foot Indians" and the others are "canoe Indians"; they are all relatively sparsely scattered over the landscape and poor in material culture, and they have similar social, political, ceremonial, and ideological customs and institutions.

Common characteristics of nomadic societies

All of the nomads so far mentioned share important general characteristics. The first and most obvious is that their nomadism severely restricts the amount of their "baggage," or material culture. Bows and arrows (except in Australia, where the unique boomerang is used instead) and perhaps a simple spear javelin, or in some areas throwing sticks or clubs, are the usual hunting and fighting weapons. In warmer zones shelter is a simple lean-to or small beehive hut of sticks, twigs, and leaves. In Arctic zones there are the caribou-skin tent and the famous Eskimo igloo—or, in more permanent or revisited places, the stone hut.

Camps are small and impermanent. The nuclear family likes to camp near related families when possible. Usually this group forms the patrilineally extended family consisting of brothers with their own nuclear families and perhaps a few dependent elders. But the size of the camp depends on the season: in times of easily gathered plant food, large groups may come together for ceremonies such as puberty rites. At other times, the constituent families may scatter

widely because food and water are scarce. Patrilineally related men and their families, scattered or not, commonly regard themselves as a group with rights over a particular territory and may be distinguished from neighbours on a territorial basis as well. Marriages are often arranged among territorial groups so that contiguous groups tend to be related, or at least certain members of different groups are related. But this is the only organizing principle that extends beyond the territorial band. Each band may be thought of as part of a larger society composed of distant as well as close relatives—a "tribe" in one of the original meanings of the word.

The social organization looks as though it had been built up from within, so to speak. Family-like statuses and roles, alliances by marriage, and systems of "social distance" based on family relationships are the bones and connective tissues of the society. These are all ingredients of the family itself, however extended or metaphorically construed; it is as though these societies were simply the result of the growth of individual families. But this is only appearance; such societies also grow by accretion. But inasmuch as alliances and the compounding of different groups normally are brought about by arranged marriages, the familistic appearance of the whole is therefore maintained.

The shaman

Almost all status positions rest upon the same criteria of age, sex, and kinship distance. The only achieved status is that of the magical curer, the shaman. Again, with the exception of the shaman, the only division of labour in these societies is on the basis of age and sex—just as in the individual nuclear family unit. Among adults, the hunting of big game is confined to men, whereas the gathering of vegetable foods or small animals, birds' eggs, and so on are women's tasks. This division of labour seems obviously related to men's relative ability to range far from camp, women being too burdened with the tasks of motherhood to track animals wherever they may lead. But the separation of tasks is usually more rigid and confining than the physical and circumstantial differences between men and women dictate, since these would vary among individuals and from society to society—and for that matter, from day to day. Domestic tasks are strictly defined as female and are undertaken only by women even when they seem exceptionally taxing, as attest the following remarks by Lewis Garrard, who traveled with a Cheyenne Indian camp in 1846:

> After a ride of two hours, we stopped, and the chiefs, fastening their horses, collected in circles, to smoke the pipe and talk, letting their squaws unpack the animals, pitch the lodges, build fires, arrange the robes, and when all was ready, these "lords of creation" dispersed to their several homes, to wait until their patient and enduring spouses prepared some food. I was provoked, nay, angry, to see the lazy, overgrown men, do nothing to help their wives; and, when the young women pulled off their bracelets and finery, to chop wood, the cup of my wrath was full to overflowing, and, in a fit of honest indignation, I pronounced them ungallant, and indeed savage in the true sense of the word.
>
> (*Wah-To-Yah and the Taos Trail;* University of Oklahoma Press, Norman, Okla., 1966)

Status within the family is based on age, sex, relationships by blood, or marriageability. Males are regarded as superior to women in most activities; the elders are respected as repositories of both secular and spiritual wisdom; and people, such as cousins who may be of the same genealogical distance, are frequently divided into "marriageable" and "nonmarriageable" groups, with consequent differences in their interpersonal behaviour. But in all other respects hunting-gathering societies are profoundly egalitarian, especially in intergroup relations.

Outside the family there is no system of coercive authority. Some persons may, by their wisdom, physical ability, and so on, rise to positions of leadership in some particular endeavour, such as a raiding party or a hunt. But these are temporary and variable positions, not posts or offices within a hierarchical structure. Social order is maintained by emphasizing correctness in conduct—etiquette—and ritual and ceremony. Ceremonies bring together the scattered members of the society to celebrate birth, puberty, marriage, and death. Such ceremonies have the effect of minimizing social dangers (or the perception of them) and also of adjusting persons to each other under controlled emotional conditions. (It may very well be true that "the family that prays together, stays together.")

The passage rites at birth, marriage, and death are universal in human society, though puberty celebrations are less common in the modern world, except for such survivals as the Jewish Bar Mitzvah. In most hunting-gathering societies, however, male puberty rituals take up more social time and engage more people than do the other three ritual occasions. They may last as long as a month, food supplies permitting. Almost universally, puberty rites include a period of instruction in adult responsibilities, rituals dramatizing the removal of boys from the mothers' care and signalizing the changed social relations between boys and girls of the same generation, and physical ordeals, including scarification or some other mark that will permanently demonstrate the successful passage to manhood.

The Plains Indians

The mounted buffalo hunters of the North American Great Plains, common in popular literature and cowboy movies, constituted a type of nomadic hunting society. But they represented a brief and very special development: an interaction and amalgamation of elements of Indian culture with Spanish horses and the training of them, as well as with metal and guns. The Indians, once mounted, could follow, surround, and kill tremendous numbers of buffalo, where previously the Indians had found the buffalo herds nearly impregnable. So productive was mounted buffalo hunting that tribes of diverse languages and customs were quickly drawn into the Great Plains from all directions. A distinctive, picturesque culture arose among them, reaching its peak about 1800. But from 1850 through the 1870s the tide of white settlers virtually wiped out the buffalo. By that time most of the Indians had been defeated in battle and confined to reservations.

Equestrian Indians can be regarded as a special form of nomadic hunters rather than as a form of pastoralists. Pastoral culture is dominated by the requirements of domesticated livestock and by the relation of herds to pasture. The Plains Indians' nomadism, however, was determined by the habits of the wild buffalo herds. The natural cycle of the buffalo was to concentrate in huge herds in summer and disperse into smaller groups in winter and spring. The Indians accordingly traveled in small camps of a few related families in winter and formed huge encampments in summer and fall for tribal ceremonies and organized cooperative hunts. The summer camps sometimes numbered several thousand people.

The continual intrusion of new groups into the Plains—first Indians, then whites—and the introduction of new weapons constantly altered the balance of power and kept the region in a state of belligerent turmoil. Equestrian bow-and-arrow Indians were superior militarily to those on foot; Indians with guns, of course, were superior to bow-and-arrow Indians; but Indians with both guns and horses—as happened in the Central Plains first—were vastly superior to the others. But the supply of horses and guns and especially ammunition continued to fluctuate wildly as access to sources varied greatly from place to place and time to time.

Nomadism places limitations on property and material technology, and the Plains Indians consequently manufactured no pottery, cloth, or basketry, although leatherwork and beadwork were highly developed. On the other hand, being equestrian, they could carry far more goods than nomadic hunters on foot. Perhaps the most notable thing they carried was the large conical tent (tepee) of decorated buffalo hide.

Sociopolitical organization was informal, probably because of the fluidity of the population. On the other hand, some tribal cohesion and systems of alliance were required because of the constant raiding. Consequently, a large number of pan-tribal associations arose, especially military societies and male age-graded societies.

Religion among the Plains Indians reflected the varying sources of the original religions of the pre-horse tribes. Some elements, however, became widespread in the Plains. The folk hero of a great many myths was the trickster Old Man Coyote. There was a widespread concept of Manitou, the pervasive spirit. Most notable was the nearly universal

importance attached to the Sun—but without the notion of the Sun as a supreme deity. Ordeals and self-torture and mass ritual self-torture were common Plains religious practices. The Indian tortured himself and fasted in order to suffer hallucinations that would reveal a personal guardian spirit for his protection in the hunt and in battle.

As in other nomadic hunting-gathering societies, principal ceremonies were related to the life cycle, with special prominence given to male puberty rites to instill bravery, endurance, and hunting and raiding skills.

### SETTLED HUNTING AND GATHERING SOCIETIES

Outstanding examples of the settled hunters and gatherers were the peoples of the North Pacific Coast of North America, roughly from Oregon to southern Alaska. The resources of the sea and inlets and rivers were of astonishing variety, and some, like the salmon during their runs, were so easy to catch that the word "harvesting" seems more appropriate than "fishing" for this activity. In central and northern California there were numerous sedentary Indian groups, such as the Pomo, Wintun, and Yurok. Their basic food was the acorn, which was ground and stored as flour. Many of the streams had salmon, and the Indians also gathered roots and berries and hunted wild fowl and deer. Other sedentary hunter-gatherer societies are rare and scattered. The most prominent of these are in southwestern New Guinea, as represented by the Asmat. These groups rely on the sago palm, whose starchy pith is easily reduced to flour. Fish, wild birds, and semidomesticated pigs supplement the basic sago.

The effect of food on population and culture

The basic foods of these sedentary peoples had two common characteristics: they were reliable and they could be stored, much as can the products of agriculture. Salmon were smoke-dried and stored in wooden boxes by the Northwest Coast Indians, and acorn flour obviously could be stored just as can grain flour. Sago flour can also be stored, but it has no season; a palm can be cut at any time the food is required. So abundant and reliable are these resources that such peoples are said to practice a "natural agriculture."

Sedentary life makes possible many improvements in material culture. Houses become larger and more elaborate and are improved over time. The Asmat of New Guinea and the Northwest Coast Indians make huge houses of planks and are among the best wood-carvers of the primitive world.

Permanent villages and a consistent abundance of food make possible high population densities. The California tribes are estimated to have reached 11 or 12 persons per square mile, as did those of the Northwest Coast. The Asmat of New Guinea have villages ranging up to 2,000 people, which is from 10 to 20 times the size of the average hunting-gathering settlement. Usually such large villages remain politically independent. Intermarriages occur, of course, and some local cohesion is achieved by secret societies and other clublike associations. But such integration is only incidental.

The Northwest Coast Indians

The Northwest Coast Indians elaborated a hierarchical form of organization, or chiefdom. They were the only hunter-gatherers to have done so. Chiefs or nobles occupied positions of high status that were inherited in a single descent line by primogeniture. Secondary lines of descent, collateral to the above, were of lesser status. Finally, there were the commoners.

Along with chiefly status went the socioeconomic institution of redistribution. Surplus products of family production were passed on to the chief, who in turn gave a large feast (or "potlatch"), during which he distributed gifts to those who needed them. This process of redistribution had the economic function of encouraging specialization and division of labour. The potlatch in late times on the Northwest Coast became famous for its competitiveness. A chief of a lineage or longhouse, for example, would amass as much food and material goods as he could in order to lay on a feast and give presents lavishly in hopes that the guest lineage would be unable to reciprocate on the same scale. One lineage, house, or perhaps village thus might "defeat" the others.

The Northwest Coast Indian type of chiefdom is primarily social and economic. It can be called political only to the extent that a certain amount of personal authority for decision making may reside in a high social status. This authority can serve a purpose, however. The egalitarian nature of hunting-gathering bands tends toward anarchy, which becomes perilous in populous societies. Quarrels can turn into feuds for lack of a higher authority to settle them.

### HORTICULTURAL SOCIETIES

Primitive agriculture is called horticulture by anthropologists rather than farming because it is carried on like simple gardening, supplementary to hunting and gathering. It differs from farming also in its relatively more primitive technology. It is typically practiced in forests, where the loose soil is easily broken up with a simple stick, rather than on grassy plains with heavy sod. Nor do horticulturalists use fertilizer intensively or crop rotation, terracing, or irrigation. Horticulture is therefore much less productive than agriculture. The villages are small—some no larger than many hunting-gathering settlements—and the overall population density is low compared with farming regions.

Forest horticulture

Forest horticulturists use fallowing techniques variously called "slash-and-burn," "shifting cultivation," and "swidden cultivation" (a northern English term now widely used by anthropologists). After about two years of cropping a plot is left fallow for some years and allowed to revert to secondary forest or bush. Before resuming cultivation the bush may be cut, left to dry, and then burned. The ashes bestow some fertilization, but the foremost benefit of this procedure is that the plot will be relatively weed free at first.

Since the fallowing periods of the plots are much longer than the planted periods, the swidden horticulturalists must gradually encroach on more distant land. Sometimes this results in semisedentary villages when the newly arable plots finally are so distant that a few horticulturalists must start to build huts near the newer fields, to be joined later by others. Such a land-hungry system, in a region of competing populations, greatly increases the chances of conflict. Population dispersal thus becomes a grave threat in horticultural regions. Land for expansion inevitably must be found at the expense of neighbours or by shortening the fallowing periods—which eventually results in lower production.

Many forest tribes—typical are the horticulturalists of the South American tropical forest—constantly maintain a military posture. Large-scale warfare is not usual (because of the lack of political leadership) but raids, cannibalism, torturing of captives, and other forms of belligerence are.

Horticulturalists have more material goods than most hunter-gatherers, though not more than such societies as the Indians of the Northwest Coast. This suggests that the accumulation of domestic goods is related not so much to the higher productivity of the horticulturalists as to their greater stability of settlement.

The most highly developed of aboriginal slash-and-burn horticulturalists were undoubtedly the Maya of Guatemala and Yucatán, who had a chiefdom or primitive state. But this was most exceptional, for almost all other primitive horticulturalists did not go beyond simple tribes with egalitarian and nearly autonomous communities. Any regional confederation was likely to be only on the basis of intermarriage and clanship. Sometimes an ephemeral sort of near-chiefdom arises, founded on the capabilities of a charismatic leader. In Melanesia, where a well-established form of personal politics thrives, the leader is called Big Man or Centre Man.

The Big Man of Melanesia

The Big Man in Melanesia is big because he has a following. He begins with his own family and near relatives and friends, who provide goods that he, on behalf of his group, gives away to other groups at a feast on some ceremonial occasion. He and his faction are feasted reciprocally by others at other times. His ability to redistribute on an increasingly lavish scale to larger groups expands his following. He thus amasses what the anthropologist Bronisław Malinowski, in reporting on the Trobriand Islanders, called a "fund of power." With the public esteem gained in this economic contest, the Big Man is sought

out for giving advice, adjudicating quarrels, planning ceremonies, and admonishing and conciliating. But this is influence, not the true authority that inheres in a status or office in an established hierarchy. A really big Big Man may succeed in integrating a region of several villages, but when he loses to a rival or dies, the unity of the region dissolves until some other unusually influential man unites it again.

In many respects the religion of horticultural peoples resembles that of hunting-gathering peoples. Shamans, life-crisis ceremonies—especially puberty rites—totemism (ceremonies for plant or animal species believed to be ancestral to particular human groups like clans or lineages), and the worship of animistic spirits are common in the religion of many kinds of primitive societies. The egalitarian society does not usually practice ancestor worship as does the hierarchical society. Among horticultural peoples with chiefdoms, the chief's ancestors, in time, become gods. The most remote ancestors, the founders of the chiefly lineage, are the most important gods; more recent ancestors and those of related but collateral lines have a lesser status. The result is a hierarchy of gods resembling the political hierarchy on Earth. Furthermore, the chiefdoms tend to be theocracies, with the hierarchy of priests closely and functionally related to the political hierarchy.

## HERDING SOCIETIES

Herding societies are in many respects the direct opposite of forest horticulturalists. They are usually the most nomadic of primitive societies, they occupy arid grasslands rather than rain forests, they have a nearly total commitment to their animals, and their sociopolitical system is nearly always that of a true hierarchical chiefdom rather than of egalitarian villages and tribal segments.

The military advantages of herders

A society largely committed to herding has military advantages that a settled agricultural society does not have. If military power is important to survival, it will increase the commitment to the herding specialization, mainly because of the advantage conferred by mobility. This increased commitment, however, will result in the gradual loss of certain previously acquired material developments such as weaving, metalworking, pottery, substantial housing and furniture, and, of course, variety in the diet. Wealth is a burden in such societies. Successful nomadic pastoralists normally have some kind of symbiotic relationship to a settled society in order to acquire goods they cannot produce themselves. The symbiosis may be through peaceful trade. But often the military advantage of the pastoralists has led to raiding rather than exchange.

The best known and purest pastoral nomads are found in the enormous arid belt from Morocco to Manchuria, passing through North Africa, Arabia, Iran, Turkistan, Tibet, and Mongolia. They include peoples as diverse as the Arabized North Africans and the Mongol hordes. Other less specialized and successful pastoralists include the Siberian reindeer herders, cattle herders of the grasslands of north-central Africa, and the Khoikhoin and Herero of southern Africa.

Classic, full pastoralism with its powerful equestrian warriors seems to have developed around 1500 to 1000 BC in inner Asia. This relatively late full-scale pastoralist specialization may have resulted from population pressure. Horticulture mixed with domestication of animals seems to have predominated until even the least cultivable zones were filled. When warfare became endemic in such zones, many groups were forced to become fully nomadic in the arid grasslands. They might have been the losers, pushed out of their homelands, only to discover later the military power that accrued to the pastoral way of life. Thus, the victims became victors.

The full pastoralism of inner Asia requires the care of animals—including varying combinations of horses, cattle, camels, sheep, and goats—kept in ecological balance with grazing conditions over an enormous area. In some regions, however, the people may depend on a single animal species. In Arabia it is the camel (although most tribes in Arabia or North Africa also keep horses), and in east-central Africa some peoples specialize exclusively in cattle. Among these there is sometimes found a complete symbiosis between a tribe of herders and an adjacent tribe of horticulturalists to the point that they resemble a single society composed of two specialized castes, the herders occupying the superior position.

There are many part-time herding societies and many that have merely borrowed herding techniques. The reindeer breeders of the Siberian tundra have learned to apply to reindeer some of the methods of the horsemen farther south, but hunting remains their main subsistential activity. In the Argentine pampas, Indians learned to domesticate and ride the Spanish horse, which they used to hunt the rhea bird and the wild herds of Spanish cattle. The Navajo and other Indians of the American Southwest have exploited the sheep brought in originally by Spaniards, but mostly as a source of wool for blanket and rug weaving. Llamas and alpacas were domesticated in the South American Andes, but no independent pastoral society ever emerged there.

Fully committed pastoralists manifest a considerable degree of cultural uniformity. In economics, social organization, political order, and even in religion, their livelihood with its functional requirements has ironed out what must have been considerable cultural differentiation among such disparate peoples as Mongols, Arabs, and African Negroes.

Cyclical routes of the pastoralists

The wanderings imposed on pastoralists by the necessities of forage and water tend to be cyclical and to follow long-established routes. The cycles are usually seasonal: to the lowlands in winter, to highlands in summer in temperate zones, to more arid areas in the wet season, to more watered regions in the dry season, and so on. Frequently the seasonal moves are accompanied by cultural and organizational changes. For example, a large group may draw together to pass through hostile territory and disperse later when in their own land; frequently the lush (normally wet) season brings the pastoralists together for ceremonies, trade, and fun, while the dry season requires dispersion and arduous work (as in digging deep wells to water the animals). Anthropologists call such established cyclical movements "transhumance orbits."

Since pastoralists live in so many different environments and since even the same society varies from season to season and in response to wider spaced drought cycles, it is reasonable to expect great variations in population density. These factors, of course, affect political organization. Nomadic pastoralism lends itself to wide fluctuations in the size of political units, political cohesion, and degree of centralization.

The elementary unit of organization is the patrilineally extended family, frequently an elder patriarch and his sons and their families. In addition, if some degree of primogeniture (*i.e.*, the eldest son inheriting most of the decision-making power for the group) prevails, and if it is extended to include other groups in terms of putative birth order and patrilineal descent, the basis of the pastoral social organization is established. This social structure has been called the "conical clan" (for its hierarchical shape). It is a characteristic social organization of chiefdoms everywhere. Its capacity for waxing and waning, fusion and fission, has obvious advantages, especially when a brief makeshift political ordering of a very large horde is militarily necessary. But the large organizations cannot maintain themselves for very long, since they easily split into their parts.

It has been noted that herders are likely to raid settled villages. But herders frequently raid each other as well. Livestock is wealth and can be exchanged for other forms of wealth—including wives. Stock raiding, like most forms of aggression, has two facets: one seems to be to replenish one's wealth at a stranger's expense; the other is to warn strangers against encroachment. But a raid frequently leads to retaliation and then to counterretaliation, until such raiding societies gradually become hereditary enemies.

The militarism of herding societies has played a major role in history. As wealthy agricultural civilizations developed in the Fertile Crescent of the ancient Middle East, in the Indus River Valley, and at the middle bend of the Huang Ho in China, they became easy prey for nomads. Indeed, it is likely that urbanization was stimulated for defensive reasons because of the dangers posed by nomads. These dangers may also have stimulated the formation of

legal and governmental institutions in sedentary societies threatened by the pastoral raiders.

Trend to ancestor worship

To the extent that pastoral nomadic societies achieve wealth and success in herding and in war, they tend to solidify and extend their chiefdom structure. They also add to their religious organization a hierarchical principle together with the content known as ancestor worship. Much of the mythology by which a primitive people explains itself and its customs comes in this way to have an ingredient familiar to readers of the Old Testament—the lengthy story of who begat whom and in what order.

Increased dependence on herds, particularly dependence on one particular species, such as cattle, horses, or camels, is reflected in much of the ideological and ritual content of religion. Sometimes the significance of herding leads not only to the glorification of herds and herding but even to a religious taboo against planting. Some Mongols, so quintessentially pastoral, believe that plowing and planting defile the earth spirit. Among the Nuer, as among other African cattle herders, horticulture may be practiced in time of need, but it is considered degrading toil whereas herding is a very prideful occupation. The ethnologist of the Nuer, E.E. Evans-Pritchard, wrote:

> They are always talking about their beasts. I used sometimes to despair that I never discussed anything with the young men but livestock and girls, and even the subject of girls led inevitably to that of cattle.
>
> (From E.E. Evans-Pritchard, *The Nuer;* Oxford University Press, London, 1940)

(E.R.Se.)

## PEASANT SOCIETIES

The one remaining category of nonurban society is that of the peasantry. Peasants are not nomadic but sedentary (thus distinguishable from both hunting-gathering societies and pastoralists); they are not horticultural tribal societies but more intensively and fully agricultural; and neither are they urban, like populations who lived in the centres of the classic civilizations.

Although writers on peasantry have not agreed on a precise definition, accounts of peasant cultures are likely to include these characteristics: peasant communities tend to be small, tradition bound, and resistant to change. Moreover, and perhaps most significantly, peasant societies are "part-societies" or "part-cultures" in relation to some larger civilization, colony, urban centre, state, or elite class. In this relationship the peasant occupies the inferior position because rural isolation tends to make him ignorant and no match for the sophisticated urbanite, and poverty keeps him dependent generation after generation. It is not simply that the peasant is somehow "exploited" (a difficult point to determine in most cases) but that his village is normally small, poor, ignorant, and backward compared with the urban centre. Horticultural tribes, by contrast, though living in even smaller villages and greater poverty and ignorance, possess in some sense a complete culture. What seems universal is the peasant's low status, with its concomitant ascription of poverty and ignorance, in contrast to other parts of the same culture.

**Historical and geographic survey.** Peasantries emerged with the urban revolution by about 4000 BC in Mesopotamia and about 1000 BC in Middle America. Some agricultural villages by then had begun to become cities, thereby initiating the process that gradually led to the formation of the empires of classical civilization.

One major significance of the classical empire was that it could protect and thus politically incorporate, at least in some areas, scattered villages of simple cultivators. These outlying agricultural villages could not have survived without internal pacification (in the law-and-order sense) and without some kind of frontier militia. In exchange for such protection the rulers of the urban centre exacted heavy tribute from the peasantry. They soon realized that the price of the protection they extended to the peasants could easily be increased to the point of virtual expropriation. As a result, the peasants were reduced to the barest subsistence level. The urban rulers, by contrast, steadily advanced toward a literate culture, thereby widening the gulf between the city elite and peasantry.

Peasants in the empires of classical civilization

In the course of time the preindustrial urban centre, with its statelike protection and intervention in the lives of peasants, made it possible to extend agricultural domains, usually as peasant holdings, into pastoral or other "wild" territory. An enormous part of the world's population became peasants as primitive peoples and nomads were dominated and displaced or transformed.

In the 20th century, with the rise of modern science and industry, peasants are being rapidly displaced in all of the so-called developing or modernizing parts of the world. To understand what peasants are, it is helpful to contrast them with what they are not—farmers in an industrialized world. The modern farmer, operating on a cost-accounting basis, has little in common with the peasant family or villager, to whom the tilling of the soil is a traditional way of life rather than a large-scale enterprise.

Peasants could not simply turn into farmers primarily because they lacked the capital. In some parts of the industrialized world, such as Great Britain, central France, and the Low Countries, peasant villages managed to survive by maintaining a low standard of living and long working hours and often by developing some special handicraft for sale as folk art. More often, however, peasants became hired workers in the fields, or they migrated to the cities or overseas to the Americas.

Distinction between peasant and farmer

Although there are no peasants in the United States and Canada, except for a few widely scattered villages of French Canadians, Appalachian mountaineers, Southern sharecroppers, and perhaps Pueblo and Navajo Indians, peasantries are still extant in large numbers in Europe, the Middle East, West Africa, southern and eastern Asia, and Latin America. Peasant cultures in these disparate regions vary considerably, depending on both ecological and historical factors. (E.R.Se./Ed.)

**Types of peasant societies.** *The community of self-serving households.* Though peasants are usually thought of as living in small, close-knit communities huddled against outside danger, they sometimes are so well-protected in mountainous or insular isolation that they feel secure enough to live more independently. In such circumstances, they dwell in scattered households in close proximity to the land they cultivate. They require a market centre of some sort where they may exchange goods and services. The Irish countryman, the isolated *fermier* of the French Massif Central, the Scottish crofter, the Paraguayan *campesino,* and the Brazilian *caboclo* are examples of such independent peasants. Occasionally the same people will be found living in close communities and also in scattered, more self-sufficient households. In the state of Michoacán in Mexico, for example, some of the tightest and closest knit communities to be found anywhere on Earth ring Lake Pátzcuaro, in immediate proximity to the large modern market town and tourist centre of Pátzcuaro. These are the fishing-, agricultural-, and handicraft-specialist villages of the Tarascan Indians. But many more thousands of Tarascans also live scattered in the adjacent mountains, making only infrequent visits to the market centres.

*The village with internal specialization and exchange.* In certain times and places peasant villages have been able to develop considerable self-sufficiency by creating part-time specialists and even full-time professional occupations. Such a development, however, presupposes an intensive agriculture in support of a fairly large population in order that the specialists may be kept fully occupied. The best examples of this kind of village are found in India, in the European medieval manor, and in some Latin-American haciendas.

The most distinctive, as well as the most clear-cut, specialization occurs in Hindu India, where a typical village may contain as many as 2,500 people. The professional specialties are pottery manufacturing, stone working, barbering, trading, weaving, laundering, and herding. All of these occupations are carried on by separate castes, to which should be added the "twice-born" caste, the Brahman, or wise-man priest, though this is more of a status than an occupation.

Hindu villages

The specialized services of the various castes often are rendered without any immediate payment or return service. The occupational castes all have an obligation to

provide their services. The full-time peasant agriculturalist, for example, expects a new plow or hoe from the carpenter, a pot from the potter, haircuts from the barber, and so on. After the semiannual harvest the peasant distributes appropriate shares of produce to those who have served him.

The caste system of occupations largely determines the status of individuals, but there are ways to attain higher status by acquiring wealth or political office. A wealthy landowner of low caste will continue to observe all the traditional attitudes of deference to those of higher caste; yet his opinion may be important and his power considerable in other than direct interpersonal dealings. And, of course, high and low status may be earned within a given caste depending on individual skill and personality.

Land ownership and tenure patterns are variable and complex. There are owners of large holdings who hire labour by wage or by shares. The majority are family-owners and workers of small plots, but large numbers of agricultural workers are landless, working only for others. Many families own some land and at the same time work other plots by shares or for wages. The usual peasant holding is worked jointly by a father and his sons. When the father dies, the land, stock, and implements are distributed equally among the sons. This practice is the major cause of the small size of the individual peasant holdings.

Many Indian peasant villages are exogamous (marrying outside), which results in ties among several villages as a consequence of giving and receiving wives. In such cases, every person participates in a social network outside his village to a greater extent than he associates with persons of other castes within his own village. These regional relationships are the means by which a common culture is diffused over a wide area. Hindu peasant villages are less alike the farther they are from each other, yet vast areas of rural India are remarkably homogeneous in culture.

The medieval manor

The European feudal estate also tended toward economic self-sufficiency in its local specialized occupations but was unlike the Hindu peasant village in several respects. For one thing, there were no castes. The aristocrats considered both their own and the peasant class to be permanent, God-given arrangements of hereditary status. Thus, to the extent that membership was in fact static, these classes were like Hindu castes (which have frequently been defined as "frozen classes"). But the other occupational classes of medieval times were not so castelike, although a tendency existed for son to succeed father. The occupational guilds resembled, to a certain extent, the wide geographic relationships of the Hindu castes. At the time of greatest stability in the European system, the social and political differences from Hindu practice were perhaps largely those of degree.

Other differences were enormous. Whereas India is overcrowded, medieval Europe, between the 11th and 15th centuries, was almost a wasteland by comparison. In fact, the existence of vast tracts of forest lands gave urgency to the problem of law and order; large groups of outlaws and predators could easily hide out. The essentials of the feudal system were master-client relationships: between kings and nobility and between the nobility. The superior individual gave protection to his clients, who in turn provided crops or services (especially labour and military duty).

Seigneury

The institution most typical of medieval society was the local seigneury, which may be defined as an estate comprising a group of people subjected to a single master. The land in such a system was of two kinds. One was the large home-farm, cultivated under the immediate direction of the seigneur or his supervisors. The other part of the seigneury consisted of various small-to-middle-sized holdings whose tenants occupied and cultivated them freely under the seigneur's protection in return for helping him in the cultivation of his demesne.

There were essentially three kinds of labourers: (1) the tenure holders, who owed regular services to the demesne; (2) wage labourers, normally paid in kind, unless they were imported labour to help out at specific occasions such as the grape harvest; and (3) workers housed and provisioned by the demesne. These workers are called prebendal in English (French *provendiers*) because they were provisioned and housed at the master's expense. The only difference between a prebendal worker and a slave was the freedom of the prebendal worker to leave if he was dissatisfied.

The tenure holder, or peasant, owed the seigneur two basic obligations, rent and services. Rents were highly variable, but services were usually still the greater burden. The basic services were agricultural labour on the demesne land, military duty, and craftwork. Agricultural labour, for example, might be calculated as three man-days a week per tenure holding. Since a family on a holding might be quite large, the three days could be divided among several men. Military duty would be highly variable because it would be a simple response to emergency—ordinarily an "all hands" response.

Craftwork was divided among peasants who had some skill passed on from father to son, especially metalworking. Spinning and weaving, wine making, carpentry, and sometimes milling and baking were duties divided among certain, but not all, peasant families. Probably most craftsmen worked at these tasks only part of the time in addition to the basic form of work.

The crafted products did not pass from peasant to peasant or between different specialists but were usually paid to the seigneur who reassigned them to others. This kind of indirect passage of goods from producer to centre and thence to ultimate consumer is the essence of the redistributional system described earlier as characteristic of chiefdoms. It is a way by which a holder of power can muster goods and serve his people at the same time. In political as well as economic structure, the resemblance of the seigneury to a primitive chiefdom is remarkable.

The seigneur was thus not simply a landowner or an exploiter of labour. He was a leader of men whose political-military status was highly direct and personal. He had command over his tenants, and the system would not have worked unless his subjects generally believed in and accepted him. His protective function gave rise to the seignorial court, which was the recognized place for a hearing of pleas and complaints. All in all, the seigneur served his people in many necessary ways, and they served him in others.

Naturally, the asymmetrical power relationship between seigneur and peasant sometimes resulted in attempts by seigneurs to multiply the services or benefits due them. On the other hand, the peasants, if numerous enough, often found ways to resist. But the power of the seigneurs presumably lay originally in their ability to allow or prevent the occupation by peasants of land under their military power. Similarly, peasants at certain early and insecure epochs might want the security of hereditary tenure; at other more secure and prosperous times, they might want freedom to leave.

By the 12th and 13th centuries in France every tenant was either free or a serf. The norm among free tenants was to be bound to the seigneur only because of their occupation of the land. If the tenant left, all obligations both ways were broken. On the other hand, the serf was not free to leave the land. Otherwise mutual dues and obligations were the same.

The hacienda

Another form of agricultural self-sufficiency is exemplified by the hacienda. In the early colonial period of Latin America the hacienda combined the Iberian and American Indian systems of land use. Pre-Columbian Indians in large areas of Latin America (from Chile north through the Andes and in Middle America) were densely settled on communal village holdings under the suzerainty of absentee aristocratic Indians. Other areas of Latin America were inhabited by more primitive tribes of slash-and-burn horticulturalists and nomadic hunter-gatherers. During colonial times in the areas of densely settled Indian population, the leading Spaniards were granted political control over designated villages. They were allowed to tax the Indian families and in return were supposed to protect them and educate them in the Roman Catholic faith. Sometimes Spaniards were rewarded by the crown with enormous tracts of land, latifundios, usually in areas of lesser population where large-scale herding would be the primary economic resource. Indian labour was also exploited in gold and silver mining and in workshops (*obrajes*).

The economy based on the exploitation of unskilled Indian labour was eventually disrupted by disease. Indians had no immunity to several commonplace European afflictions such as smallpox, typhoid fever, measles, and malaria. Numerous disastrous epidemics occurred, and by about 1600 both Spain and its richest New World possessions were in rapid economic decline.

Meanwhile, a new form of rural estate came into being as the economy of town and city, workshops, mines, and commerce was depressed. A large, privately owned estate could withstand monetary and commercial crises by becoming increasingly self-sufficient. The estate was manned by impoverished Indian workers who needed security and protection. The workers were usually paid in kind, enough for bare subsistence and given credit (against the promise of future labour) for the purchase of other necessities. This debt peonage was the foundation of a permanent labour supply, resembling the serfdom of medieval Europe.

The hacienda had a permanent group of peons settled on its lands, allowed to farm small plots for themselves. There were also house servants, some of whom might reside in the master's home. Other Indians might be residents of neighbouring villages but dependent on the hacienda for protection and often for grazing rights on fallowed rangeland claimed by the hacienda. A hacienda with numerous dependent villages on its periphery could muster a large labour force when needed and not employ it when not needed. The permanent debt peons, however, were more closely bound up in the everyday life of the hacienda. Like the European serf, the peon in difficult times probably welcomed the security of such an arrangement.

The hacienda probably was never completely self-sufficient, but it could take care of its own people in many ways. Large haciendas, some with thousands of peons, could afford numerous specialists, such as metalworkers and leatherworkers, weavers, bakers, masons, carpenters, and sometimes even a resident priest. There might also be a jail and a whipping post. And just as in European seignorial law, the master adjudicated disputes and meted out punishment. The economy of the specialized crafts resembled the European redistributional system insofar as the planning, commissioning, and delivery of all benefits were centralized under the hacienda master and his agents.

Virtual bondage of tenants

Although debt bondage no longer exists in Latin America, the tenant worker on the remaining large haciendas in some of the Andean areas seems as closely bound to the soil as peasants ever were. The Chilean tenant is legally free to move as he pleases, but he cannot, in fact, usually do so. He works his ancestral land, which he understands belongs to the hacienda, whose owner he has been conditioned all his life to regard as his master and protector. Were the worker and his family to leave, the other haciendas would not accept him. And since there is no vacant fertile land he could not become a squatter. Most peasants fear the city, which is already filled with the unemployed younger sons of peasants.

In Mexico, it was not until well into the 20th century that the hacienda system began to yield to modernism and more liberal laws, and the hacienda became increasingly commercialized. But earlier the peasant could not improve his position, legally or economically. By the end of the regime of Porfirio Díaz in 1911, the concentration of ownership of land in the hands of a few hacendados was greater than in any other Latin-American country. But the payment for agricultural labour had not risen appreciably since 1792. Over the same period the price of maize had increased 179 percent and that of beans 565 percent.

*The closed regional market system.* A kind of regional self-sufficiency may be seen among peasants in the Middle American highlands and the Andes, in parts of Indonesia, and in West Africa. These are nearly self-sufficient regions embracing peasant hamlets and villages that trade with each other, usually on a periodic market day or fair. These villages are typically cohesive and tend to be self-governing through ritual and religion. The relations of peasants with the outside world are usually mediated by the community (or its officials), except in the peasant market, where frequently some kind of middleman or representative of a store sells items not produced in the region itself.

The highland Indians of Mexico and Guatemala

The highland Indian communities, especially in Mexico and Guatemala, are quintessentially of this type. The inhabitants of the region are Indians (although there is a heavy overlap of Old Spanish custom, dress, and a folkish variant of Spanish Roman Catholicism). They all speak the same Indian regional dialect. They occupy such a distinctly inferior and helpless position in relation to the outer society that intermediaries of some sort are needed, and as a consequence of the felt inferiority they act toward outsiders in an extremely withdrawn manner. This withdrawal trait of the Indian peasantry has been appropriately labeled the "*encogido* syndrome," meaning a nearly utter lack of self-confidence.

The Indian communities that have legal ejidos (communal holdings) as well as small family properties are not usually subject to outside landowners. Thus, the *encogido* syndrome derives not from any inferior position of the peasantry to a resident owner (as in the hacienda system) but from the simple fact of ethnic stratification. These Indians feel themselves in an inferior position to everyone else in the whole outside world, that is, inferior to everybody except others designated as Indians. And they are usually also inferior in visible ways: extremely poor, uneducated, and ignorant of the manners and customs of the nation's urban citizenry. They have responded by withdrawing into their own community, which only serves to continue or to exacerbate the inferior condition, since much of it is caused simply by economic and social isolation from the main currents of national life.

Within the community of Indians extreme egalitarianism prevails. Prestige is won only in community service, which means giving more than receiving—a pattern very much like that of general reciprocity of primitive society. Giving takes the form of subsidizing one of the traditional fiestas of the community, an obligation that may be expensive in food, liquor, candles, and fireworks. Families work for years to get up the capital to sponsor a magnificent festival, at which they not only spend their savings but go into considerable debt. Lavishness in money and labour equals love (of the community); the poverty of an individual family may be accompanied by the highest prestige. This bears a resemblance to the Big Man system of Melanesia, where prestige is also linked with lavish giving.

Marketing behaviour

The economic egalitarianism of the village does not rule the periodic regional market. There the family, or its agent, tries to maximize its gains. Outsiders who see more of the marketing behaviour than of the more pervasive social economy within the individual villages are often misled by the ferocious haggling and cheating. Many regional markets are held in a smallish city of the national, non-Indian sort. The economically important market is the one in which the Indians exchange their own products with each other, yet it may be overshadowed by the market in which Indian handicrafts are sold to outsiders. Town stores may display their own wares in stalls at the market, and the carrier-middlemen may have brought materials from a considerable distance to sell where such items are rare and will bring a better price. And the same middleman (often simply the owner of a pickup truck) may buy materials that are abundant at one market for sale in another. In Mexico, the weekly Tarascan market in Pátzcuaro or the Otomí at Ixmiquilpan or, in Guatemala, the Mayan markets at Chichicastenango or Panajachel are large and complex, incorporating all of the above elements.

Close-knit peasant communities do change, of course; but it is their resistance to change that commands attention in modern times, when virtually all other institutions are committed to rapid change. Because the peasant community is a rigid structure, the individuals who want change and have the means or capacity to carry it out simply leave the village to find their place in some mixed community. In some peasant communities this is not possible. Thus, in densely settled central Java opportunities for the surplus peasant population to adapt to other modes of life were few. As population increased, putting stress on the traditional network of communal villages, the community tried to fit more people into the traditional system. The social patterns grew more elaborate but remained traditional; cooperative labour and tenancy institutions became

more intricate but otherwise did not change. Just as in the Latin-American Indian commune, egalitarianism continued on a basis of what has been called shared poverty.

In West Africa peasant society is based on full-time intensive agriculture, with a considerable amount of craft specialization and a large amount of trade carried on in great markets. In ancient times the more populated areas were organized as chiefdoms and primitive states, which exacted a tribute or tax from the agriculturalists in exchange for their protection and regulation. Intensive cultivators produce mainly for their own consumption, but with a frequent surplus or specialized handicraft to trade in the market. In modern times, they may produce surpluses to trade with a middleman representing an exporter. African villagers traditionally have been politically dependent on some hierarchical chieftains or patrimonial retainers. But they do not, like peasants elsewhere, feel themselves to be different from or inferior to any other class or culture. Nor are they so regarded by the urbanites with whom they come in contact.

General characteristics of peasant societies

One characteristic of the peasant economy is that the production unit is normally the family. But this does not mean that families are all the same size. Mainly, technical and economic requirements tend to govern the size of the family, which ranges from large three-generational extended families down to the nuclear unit of one set of parents and their unmarried children. Inheritance patterns tend to reflect the requirements of the agricultural operation. Whether the land is split equally among the heirs or passed on as a single unit (commonly through the eldest son) depends on whether farming requires large holdings or whether a small, intensively farmed area is sufficient. In some historical instances, the ecological determinant of the size of holdings has been contravened by ideology or law. For example, the Code Napoléon required that agricultural holdings be inherited equally, with the result that when the fragments of land were obviously becoming too small the French peasants responded with one of the most drastic reductions in the birthrate in all of recorded demographic history.

Some villages, notably among Latin-American Indians, are quite communalized. Individual families work plots of land, but it is the community as a whole that makes the important decisions. Other peasant societies with more independent households find numerous occasions for cooperation and labour exchanges among families. One widespread means of establishing a network of reciprocal obligations and trust within a peasant community is through ritualized ties of fictive kinship, such as the godparenthood common throughout most of peasant Europe and Latin America (in Spanish it is co-parenthood—*compadrazgo*). Other forms of fictive kinship are the familiar blood brotherhood of Balkan Europe, the *mit* of Nepal, and the *oyabun-kobun* of rural Japan. (E.R.Se.)

## Urban cultures

### DEFINITIONS OF THE CITY AND URBAN CULTURES

Research on urban cultures naturally focuses on their defining institution, the city, and the lifeways, or cultural forms, that grow up within cities. Urban scholarship has steadily progressed toward a conception of cities and urban cultures that is free of ethnocentrism, with broad cross-cultural and historical validity.

Well into the 20th century conceptions of the city often proceeded as if there were only one authentic or typical form. From his research on the city in Europe's Middle Ages, Henri Pirenne, for example, argued in *Medieval Cities* (1925) that two characteristics were fundamental to the development of an urban culture: a bourgeoisie, or middle class, that depends on trade for both wealth and political autonomy from nonurban feudal power holders; and a communal organization of the urban citizenry that creates the municipal integration necessary to free the city from control by local feudal lords or religious authorities. Although it has often been taken as a general definition of the city and urban culture (whence the commonsense notion that cities must fulfill commercial functions), Pirenne's formulation was deficient because only the European medieval city and its burgher culture were taken as typical of the "true" city.

Max Weber's definition of the city

Max Weber in *The City* (1921) provided another definition of the city, similar to Pirenne's, when he contrasted "Occidental" with "Oriental" urbanism. According to Weber, five attributes define an urban community: it must possess (1) a fortification, (2) a market, (3) a law code and court system of its own, (4) an association of urban citizenry creating a sense of municipal corporateness, and (5) sufficient political autonomy for urban citizens to choose the city's governors. Weber believed that Oriental cities rarely achieved these essential characteristics because familial, tribal, or sectarian identities prevented urban residents from forming a unified urban citizenry able to resist state control. Even with regard to the Occident Weber's definition would exclude almost all premodern cities, for the urban autonomy he required existed only in northern Europe and Italy and, even there, for very short periods of time at the end of the Middle Ages. The result was an overly limited conception of urban cultures, from which it was extremely difficult to generate a cross-culturally valid understanding.

In the 1940s Robert Redfield, strongly influenced by Louis Wirth and other members of the Chicago school of urban ecology, conceived of the urban as invariably impersonal, heterogeneous, secular, and disorganizing. In the folk-urban model, as set forth in his article "The Folk Society," Redfield contrasted this image of city life with an image of the folk community, which he characterized as small, sacred, highly personalistic, and homogeneous. He presumed that as individuals moved from folk community to city or as an entire society moved toward a more urbanized culture, there would be a breakdown in cultural traditions. Urbanizing individuals and societies would suffer from cultural disorganization and would have higher incidences of social pathologies like divorce, alcoholism, crime, and loneliness.

Redfield's conception of the city depended on the urban research carried on by sociologists in American industrial cities, predominantly Chicago. He ethnocentrically assumed that their findings could be generalized to all urban cultures. Subsequent research indicated that this conception was in many respects wrong even for American industrial cities. In spite of being generally ethnocentric and specifically inadequate for American cities, this conception still holds sway over much popular thinking, which conceives of cities, in all cultures and all times, as centres of bohemianism, social experimentation, dissent, anomie, crime, and similar conditions—whether for good or bad—created by social breakdown.

Gideon Sjoberg (*The Preindustrial City, Past and Present,* 1960), in the next step toward a cross-culturally valid understanding of cities, challenged this conception of urban culture as ethnocentric and historically narrow. He divided the world's urban centres into two types, the preindustrial city and the industrial city, which he distinguished on the basis of differences in the society's technological level. Preindustrial cities, according to Sjoberg, are to be found in societies without sophisticated machine technology, where human and animal labour form the basis for economic production. Industrial cities predominate in the modernized nations of western Europe and America where energy sources from fossil fuels and atomic power phenomenally expand economic productivity. For Sjoberg, preindustrial urban culture differed markedly from its industrial counterpart: the preindustrial city's neighbourhoods were strongly integrated by personalistic ties of ethnicity and sectarian allegiance; it maintained strong family connections, and social disorganization was little in evidence; churches or other sacred institutions dominated the skyline as well as the cultural beliefs of the urban place; and the major urban function was imperial administration rather than industrial production.

Preindustrial and industrial cities

Although Sjoberg's conception of a preindustrial urban type was a major improvement over previous urban definitions, it too suffered from overgeneralization. Sjoberg collapsed urban cultures of strikingly different sorts into a single undifferentiated preindustrial city type—for example, the cities of ancient empires were conflated with

present-day urban places in the Third World. Past urban cultures that did not readily fit the Sjoberg conception, such as the autocephalous (self-governing) cities of early modern Europe, were disposed of as temporary and unusual variants of his preindustrial type rather than important varieties of urban culture.

In "The Cultural Role of Cities," Robert Redfield and Milton Singer tried to improve on all previous conceptions of the city, including the one Redfield had himself used in his folk-urban model, by emphasizing the variable cultural roles played by cities in societies. Redfield and Singer delineated two cultural roles for cities that all urban places perform, although with varying degrees of intensity and elaboration. Cities whose predominant cultural role is the construction and codification of the society's traditions perform "orthogenetic" functions. In such urban cultures, cadres of literati rationalize a "Great Tradition" of culture for the society at large. The cultural message emanating from Delhi, Paris, Washington, D.C., and other capitals of classic empires or modern nation-states functions to elaborate and safeguard cultural tradition. By contrast, cities whose primary cultural role is "heterogenetic," as Redfield and Singer defined it, are centres of technical and economic change, and they function to create and introduce new ideas, cosmologies, and social practices into the society. In cities like London, Marseille, or New York, the intelligentsia challenge old methods, question established traditions, and help make such cities innovative cultural centres.

The city's orthogenetic and heterogenetic functions

Continuing Redfield and Singer's concern for the cultural role of cities within their societies, Paul Wheatley in *The Pivot of the Four Quarters* (1971) has taken the earliest form of urban culture to be a ceremonial or cult centre that organized and dominated a surrounding rural region through its sacred practices and authority. According to Wheatley, only later did economic prominence and political power get added to this original urban cultural role. Wheatley, following Redfield and Singer, established that any conception of an urban culture had to be grounded in the cultural role of cities in their societies; research must specifically address how the urban cultural role organizes beliefs and practices in the wider culture beyond the urban precincts, and, consequently, how this urban cultural role necessitates certain lifeways and social groupings (cultural forms) in the city.

Beginning in the 1970s, David Harvey (*Social Justice and the City,* 1973), Manuel Castells (*The Urban Question,* 1977), and other scholars influenced by Marxism caused a major shift in the conception of urban cultural roles. Although they mainly worked on cities in advanced capitalist cultures, their approach had wide relevance. Rather than looking outward from the city to the urban culture as a whole, the new scholarship conceived the city as a terminus for cultural roles emanating from the wider culture or even the world system. Harvey, for example, linked major changes in American urban lifeways to the urban culture of advanced capitalism: for him, the growth of suburbia developed out of capitalism's promotion of new patterns of consumption in the interests of profit. Castells saw the city as an arena for social conflicts ultimately emanating from the class divisions within capitalist society.

This Marxist scholarship did not contradict the earlier emphasis on the city as the source of cultural roles so much as complement it. Studying the cultural roles of cities must include not only the cultural beliefs and practices that emanate from cities but also the cultural forms that develop within the city as a result of the impact of the urban culture on it. In this way scholarship can bring forward a cross-culturally and historically valid conception of cities, their cultural forms, and the urban cultures in which they are set.

## TYPES OF URBAN CULTURES

The following typology of urban cultures depends on a conception of cities as centres for the performance of cultural roles found only in state-level societies. Such societies, in contrast to the nonurban cultures previously discussed, have inequalities in economic wealth and political power, the former usually evidenced by class divisions, the latter by specialized institutions of social control (ruling elites, government bureaucracies). Because cities do not occur in societies without state organization, the terms "urban cultures" and "state-level societies" are closely linked—the former emphasizing belief patterns, the latter stressing social organization in such societies.

State-level societies and urban cultures

State-level societies differ in the nature and extent of economic and political inequalities, and this variability accounts for the different types of urban cultures and cultural roles adduced below. The labels for the types of urban cultures denote the predominant cultural role played by cities in this urban culture—thus, "ritual city" or "administrative city." Obviously, cities in any society combine some amount of ritual role with administrative functions. The rationale for the labels used below, however, is that given particular constellations of inequalities, certain urban cultures come to exist and certain cultural roles of cities come to characterize or typify them. Thus, the label "administrative city" typifies the major (but not exclusive) cultural role played by cities in agrarian empires, whereas "industrial cities" represents the dominant urban cultural role in capitalist nation-states.

The typology below draws a major distinction between urban cultures that existed before the development of the world capitalist system in the 16th century and those that came after. Before the world capitalist system developed, state-level societies were not integrated in an economically unequal relationship. The advent of the capitalist world system led to a specialized world economy, in which some state-level societies represented the core and others represented the economically, and often politically, subservient periphery. Before the world system, urban cultures differed mainly on the basis of internal differences in political and economic inequality. After the world system, urban cultures, in addition, differed according to their placement in either the core or the periphery.

**Urban cultures before the capitalist world system.** *The ritual city.* Ritual cities represented the earliest form of urban centre, in which the city served as a centre for the performance of ritual and for the orthogenetic constitution and conservation of the society's traditions. Ritual was the major cultural role of such cities, and through the enactment of ritual in the urban locale, rural regions were bound together by ties of common belief and cultural performance.

The early forms of urbanism in the pristine civilizations of the Old World and Meso-America, which Wheatley refers to as "cult centres," conform to the ritual city type. Other examples of ritual cities can be drawn from ethnographies of the urban culture of the Swazi in southeast Africa, Dahomey in West Africa, and Bali before the Dutch conquest. In most areas of the world this form of urban culture was quickly succeeded by more complex types.

Ritual cities were found in urban cultures that have been called "segmentary states" or "primitive states." Such states had minimal development of class stratification and political coercion. Although segmentary states had rulers, such as a chiefly lineage or a priesthood, control over land and other means of production remained with clans, lineages, or other kin-based groups outside the rulers' domination. Political authority and economic wealth were therefore widely dispersed.

Limited political centralism and economic coordination meant that the ritual, prestige, and status functions of the state loomed large. Segmentary state rulers were symbolic embodiments of supernatural royal cults or sacred ritual ones. They—their courts and temples—provided a model of the proper political order and status hierarchy that was adhered to throughout the otherwise weakly cohered segmentary state. Through the awe they inspired, they extracted gifts from the rural populace with which to sustain their royal or priestly election.

The cultural forms of ritual cities centred on the cult centres, temple complexes, or royal courts that dominated their physical space and defined their urban role. As the rulers' habitation, the ritual city spatially embodied the role of the sacred and ceremonial in defining the urban culture. The everyday population of the city consisted of those bound to court or temple by family, official duties,

or craft and ritual specializations; at ceremonial times, people from the surrounding rural areas temporarily swelled the urban area. Therefore, rather than individualism, secularism, or impersonality, the calendrical round of state rituals, kingly ceremonies, divine sacrifices, sacred celebrations, feasts, funerals, and installations defined urban life, rendering it sacred, corporate, and personalistic.

The unifying effect of the ritual city

The city as ritual centre made for strong rural–urban solidarity. Because in the segmentary state power and wealth were dispersed rather than concentrated in the city, there existed no intrinsic antagonism between country and city. Consequently the orthogenetic message of tradition and sacredness broadcast from the city throughout the urban culture had a unifying effect, forging a solid rural–urban bond.

*The administrative city.* Like ritual cities, administrative cities were the habitations of the state rulers. Their major cultural role was to serve as the locus of state administration. State offices and officers had an urban location, from which they exercised a political control and economic exploitation of the surrounding rural areas quite unknown in ritual cities. Administrative cities also had a qualitatively different demographic and social complexity. They contained large populations, densely settled, often ethnically varied, with heterogeneous occupations. Such cities were nodes of communication and transportation and centres of commerce, crafts, and other economic functions for the surrounding countryside.

Administrative cities occurred in agrarian empires, state-level societies associated with the early civilizations of Hindu and Muslim India, China, and Egypt, as well as the Mamlūk Middle East, Tokugawa Japan, Alexandrine Greece, and other expansive territorial states before the advent of the world capitalist system. These states had rulers with great powers of political coercion, which they used to maintain a high level of inequality in wealth between the state ruling elite and the primary producers, the peasantry. This type of urban culture rested on how effectively the state could exploitatively control peasant agricultural productivity for maintaining the elite. The urban administrative cultural role was the major means to this end.

The administrative city brought together the political, economic, transport, and communications functions and institutions necessary for this rural rapine. For just as the state elite preyed on the peasant, so the administrative city's flamboyant architecture and monumental public works ultimately rested on what could be taken from the rice paddies of the Japanese cultivator or the wheat field of the Indian peasant. There also grew up urban populations that converted the wealth taxed from the rural area into a sumptuous life-style for the urban-resident state elite: artisans and artists, of various levels of reputation. This gave rise to the poor of the city and, often, institutions to help govern and subdue them, such as municipal governments. Merchants also were necessary to convert the peasant's grain payments into cash. Administrative cities commonly tried to restrain the wealth of urban merchants from fear that such riches might be converted into political power.

As the links between coercive state and oppressed peasant grew stronger (that is, as the two became more unequal), the urban cultural practices (for the elite) became more separated from those of the countryside. The urban area concentrated a sophistication, an elaboration of custom and ideology that marked it off from the rural, which now was defined as rustic. Alongside the elaborate, the monumental, and the beautiful, which distinguished the administrative city's architecture, elite entertainments, and general cultural forms from those of the countryside, however, there was also an overwhelming poverty in the city's artisan and servant wards.

Growing discrepancy between urban and rural ways of life

The administrative city had some of the properties commonly attributed to cities: it was a locale for cultural elaboration and monumental building, a repository of great wealth but also of extensive poverty, and a heterogeneous locale, both occupationally and in terms of ascriptive identities based on ethnicity, religion, caste, or race. But it was not disorganized or impersonal. Family, guild, and ethnic group framed the allegiances that defined the basic unit of urban cultural practice, the city quarter, which for the urban nonelite functioned with many of the characteristic cohesions of the peasant village.

*The mercantile city.* Mercantile cities appeared at the geographic margins or at times of dissolution of agrarian empires—for example, in medieval and early modern Europe, after a decentralized feudalism had fully replaced the Roman Empire. This urban type is thus a variant form that appeared, under particular conditions, in the urban cultures that also contained administrative cities. The mercantile city's links with the wider culture were disjunctive rather than, as with the administrative city, supportive. A class of powerful and wealthy merchants not completely beholden to the state rulers grew up in such cities and, left unchecked, could grow strong enough to effectively challenge the state rulers. This merchant class, and the mercantile cities it occupied, depended for their wealth and political autonomy on the profits of international trade, moneylending, or investment in cash cropping of export agricultural commodities (as, for example, vineyards and olive groves in the Mediterranean). The city produced wealth and capital in its own right rather than simply sucking it from rural agriculture. Such wealth provided an avenue for political power separate from that offered by the revenues derived from the peasantry. Often, therefore, urban magnates and state power holders or rural gentry stood in strong opposition, each trying to control—or absorb—the wealth and power of the other.

Mercantile cities varied in the extent of legal, fiscal, and martial autonomy they enjoyed. At their most developed, they conformed to the definitions of "true" cities provided by Weber and Pirenne. They enjoyed independent municipal government, sported urban fortifications, fielded citizen armies, and even subdued surrounding rural magnates. In less developed (generally earlier) mercantile cities, urban independence was not so great: for example, urban trading capital depended on intermarriage with rural magnates or came from rural moneylending. Even in such cases, however, rural resources were put to novel uses in the urban setting.

The cultural role of mercantile cities grew out of their independent economic productivity and their political autonomy. They played a very strong heterogenetic role. They were strongholds of a merchant class and other social strata based on acquired wealth, against the landed aristocracy of the agrarian empire. Because they were often under attack from the aristocracy, these cities came to symbolize freedom and social mobility: "city air makes one free." Being embattled, mercantile cities also became bastions of cultural innovation. Urban cultural form emphasized achievement, and urban politics involved shifting factional alignments. Given the volatility of commercial operations, leading families rose and fell rapidly, and plutocracies, quite fluid in membership, came to rule these cities. The poor artisans and small traders too were more independent than in administrative cities, and through occupational or sectarian associations, like guilds, they demanded and won political concessions.

Mercantile cities as seedbeds of cultural innovation

Although places of innovation, achievement, freedom, and mobility—traits that they share with industrial cities—mercantile cities were neither impersonal nor secular. The extended family was the major institution organizing business firms, political coalitions, and much elite social life. Other corporate institutions, like guilds and religious fraternities, joined city dwellers into highly personalized, ritualized associations that downplayed individualism and secularism in the city.

Given the commercial conditions and the difficult class oppositions that set the cultural context for mercantile cities, they proved evanescent and fragile, usually reverting under state intervention to administrative cities, in which the merchant magnates and their wealth came under the control of state rulers.

**Urban cultures since the capitalist world system.** Beginning in the 15th century, the Age of Discovery, Europeans carried the capitalist system burgeoning at home to distant places, whose labour and productivity were harnessed to the European core in an unequal, colonial relationship. The result was the capitalist world system, as Immanuel M. Wallerstein in *The Modern World-System* (1974) terms

it. There was increasing economic and productive specialization among the world's regions, as a pattern of unequal exchange developed between the industrial commodities of the advanced European nations (at the world system's core) and the raw materials from underdeveloped Asia, Africa, and the New World (at the world system's periphery). By the 18th century a worldwide urban culture had come into existence. It took variant forms of economic, political, and urban organization in the colonizing core and in the colonized periphery. Although the following discussion treats urban cultures in the core and in the periphery separately, it must be remembered that they—and the urban cultural roles that typify them—form an interactive unit.

*The industrial city.* Industrial cities appeared after the full development of industrial capitalism in the core nation-states of the late 18th-century world system. Their urban cultural role fit well with the capitalist economic order that came to dominate all other social institutions. Capitalism depended on the production of commodities through wage labour in the interests of capital accumulation. The city became a centre of such production processes and the location for the industrial factories in which this production most typically took place. It was also the residence for the other "commodity" necessary to its productivity, wage labourers. Ancillary urban functions—banking, wholesale and retail trade, transportation and communications nodality—grew up to expedite the factory production or the provisioning of the labour force.

Factories and wage labourers in the industrial city

Rapid population increase through in-migration characterized the growth of the industrial city. The most salient aspects of urban cultural forms grew up in the neighbourhoods that housed the newly urbanized labour. Populations with very different cultural characteristics came together in the city, such as the Irish in the British Midlands or the many ethnic groups that formed the urban American melting pot. Ethnic and racial ties often provided the links for migration chains, and they helped recent migrants find jobs, housing, and friendship in a new environment. These ties often resulted in ethnically segregated urban neighbourhoods among the working class.

Two contradictory patterns of organization and conflict characterized this urban population. One pattern grew out of the dense settlement of the working class in the industrial city. Residential aggregation helped organize large-scale working-class protest in the interests of better working conditions and wages. The other, contradictory, pattern consisted of ethnic or racial exclusiveness and competition within the working class. Ethnic or racial residential segregation often provided the base for competition among members of the working class for jobs and urban locations convenient to the workplace. Characteristically, one ethnic population in the industrial city guarded its neighbourhood against invasion by another—or, in times of rapid economic growth and social mobility, ethnic succession, wherein an upwardly mobile ethnic population would leave its neighbourhood to a newly urbanizing ethnic grouping, would occur. The retention or strengthening of ethnic or racial identities in industrial cities commonly took place under these conditions.

The industrial city is the terminus for two conflicting processes emanating from the capitalist character of the wider society: capitalist investment in urban property for profit making, and class conflict. The former process subjects the human and natural environment to the interests of capital accumulation; the latter makes for the formation of urban neighbourhood associations, ethnic associations, and other sorts of class alliances that organize local resistance to this profit taking. The city then becomes a battleground for these opposing forces. Castells in *The City and the Grassroots* (1983) has studied a range of social movements in present-day American and European industrial cities that arose in resistance to capitalist rationalization of the urban environment. The resistance can take different forms but includes attempts to preserve public services or public spaces for their use value against a capitalist rationality that would privatize and put a price tag on them—that is, this resistance aims at making municipalities rather than private enterprise responsible for provisioning good schools, recreational facilities, museums, and parks. Other forms of resistance consist of attempts to preserve the cultural identity of neighbourhoods and subcultures against residential blockbusting and attempts to develop neighbourhood decentralization so that the urban population takes control over its own living environment.

*The mass-communications city.* The industrial city, consonant with the rise and consolidation of capitalism in the western European and North American core nations, appears to be rapidly giving way to what has been termed the mass-communications city in the advanced industrial nations. Cities such as New York, London, Tokyo, and other metropoles increasingly perform a primary cultural role as centres of managerial control, based on high-technology mass communication and data processing, over far-flung manufacturing activities. Old urban manufacturing centres in the core of the capitalist system, like Birmingham, Eng., Detroit, and Glasgow, have declined as their role in industrial production has become less important.

The movement toward the mass-communications city has to do with changes in the urban culture of the core brought on by changes in the world system since the 20th century began. This development of "late capitalism," "monopoly capitalism," or the "welfare state," as it is usually labeled, depended on the investment of capital from the core to generate industrial production in the periphery, usually through the institution of multinational corporations. The cultural role of core cities is shifting away from manufacturing as they come to house the advanced means of communication and data analysis necessary to manage this worldwide industrial production.

The mass-communications city ceases to be primarily a habitation of the industrial working class. Instead, those working mainly in high technology industry and service (the middle class) define urban cultural forms. For example, suburbanization and gentrification, two characteristic urban residential patterns of the middle class, become important cultural forms in such cities. Both show the emerging importance of the new social class and the provisioning of new urban spaces (the suburbs) or the renovation of old ones (gentrified inner cities) for it. Again, these new urban locales represent the larger capitalist society, in that they are locales for profit making as well as arenas of class resistance. Harvey in *Consciousness and the Urban Experience* (1985) argues, for example, that the suburbanization process typical of American cities, especially after World War II, was motivated by the need to foster a new life-style of consumption to negate problems of capitalist overproduction. It also minimized class violence by spreading population out from the old, dense, inner-city neighbourhoods. These suburbs, however, once created in the service of capitalist profit making, can become the locales for resistance against further capitalist rationalization of urban space and against the inroads of welfare statism on local decision making.

Suburbanization and gentrification

As the mass-communications urban cultural role further develops in the advanced industrial societies and industrial production is exported, whatever urban manufacturing continues must meet the competition of imported commodities. Various new means of urban labour use develop to make production cost-effective. For example, manufacturing is left to the lowest strata of the urban population, either illegal migrants, such as Mexicans or Haitians in the United States, or the most underprivileged of the national population, such as American blacks, or foreign workers, such as eastern Europeans or Turks in France, who do not have full citizen rights. Often, manufacturing that once was done in factories is now done in homes as a way of minimizing costs, especially by avoiding government regulations and taxation. Thus, because they work at home or because they are illegal migrants or because they are subject to racial prejudice, the labourers have little legal protection and welfare support. In the face of this massive insecurity they depend on extensive mutual-aid networks, in which the poor share the risks of poverty among themselves. Their abject condition—and their attempts at security—mirror the practices of poor shantytown dwellers in neocolonial cities, as described below.

*The colonial city.* Colonial cities arose in societies that

fell under the domination of Europe and North America in the early expansion of the capitalist world system. The colonial relationship required altering the productivity of the colonial society in order that its wealth could be exported to the core nations, and colonial cities centralized this function. Their major cultural role was to house the agencies of this unequal relationship: the colonial political institutions—bureaucracies, police, and the military—by which the core ruled the colony, and the economic structure—banks, merchants, and moneylenders—through which wealth drained from colony to core.

Bombay and Calcutta under the British, the European trading cities in China and West Africa, the British East African and Dutch East Indian urban centres for the collection of plantation crops—from the 18th through the mid-20th centuries—represent this urban type. The core capitalist nations implanted colonial cities as new growths into preexisting precapitalist state societies in many world regions, just as they altered the societies by making them unequal participants in world capitalism. The resulting urban culture represented a novel amalgam of the core and the periphery, with qualities not found in either parent culture.

Amalgam of core and periphery

This new combination was most in evidence in the elite population of the colonial city and its cultural forms. For example, new classes and urban lifeways appeared among the indigenous population. Most of the time the cultural role of the colonial city required the creation of an indigenous urban lower-middle class of merchants, moneylenders, civil servants, and others who were educated to serve the colonial political and economic establishment. For instance, Thomas Babington Macauley, a British Indian administrator in the mid-19th century, hoped to create an elite through Western-style education that was "Indian in blood and colour, but English in taste, in opinion, in morals and intellect." The colonial educated lower-middle class often attempted to reform its culture in line with that of the colonizing power, most often through new urban institutions like schools, welfare associations, and sectarian or secular reform groups. A generation or so later, this class transformed by these urban institutions, commonly formed the leadership of nationalist, anticolonial movements. Thus, the colonial city, which began as an instrument of colonial exploitation, became a vehicle of anticolonial protest through this lower middle class and the cultural institutions, schools, newspapers, and other urban cultural forms it had constructed.

After World War II many new nations in Asia and Africa gained independence. Although no longer the direct political colonies of Western countries, these urban cultures and their cities continued in a dependent economic relationship with the advanced industrial nations.

*The neocolonial city.* The latest type of urban development in the periphery of the capitalist world system, or what is often called the Third World, is the neocolonial city. This urban type has arisen in relation to the development of monopoly capitalism and the mass-communications city in the core. Export capital from advanced industrial nations has created enclaves of industrial production in Third World cities, thus replicating in these urban places many of the cultural roles played by the industrial city in the core. There are urban factories and urban-resident wage labourers. There is a developing infrastructure of urban transport and communication by which these commodities and labourers are allocated. There is massive urban-ward migration from neighbouring rural areas.

The neocolonial city, however, does not exactly duplicate the cultural role of the industrial urban type precisely because of its dependent relationship with the core. One major difference is that the commodities produced in neocolonial cities generally are destined for export rather than for home consumption, except perhaps by a small home elite. The neocolonial city does not serve an indigenous hinterland; it serves the wider world economy. Its rural environs are important only because they provide a large and readily available labour supply.

Distinction between industrial and neocolonial city

The large-scale urbanization in the neocolonial city differs from the urbanization that characterized the industrial city earlier. It gives rise to what has been called the informal economy in these cities. The informal economy consists of urban services and products provided by the neocolonial city's poorest denizens, the petty hawkers, the shoeshine boys, the household help, the rag pickers, and others who form a class of petty commodity producers and sellers. The common image of these people is highly pejorative: they are marginal to the city, usually unemployed and often criminal, unmotivated and dysfunctional to urban life, characterized by a "culture of poverty" that, at the same time, makes them accept their wretched condition and keeps them in it. Their marginality is often said to be exemplified in the shantytowns, tin can cities, or squatter settlements that they build at the borders of the city and that blight it. This "myth of marginality" as Janice Perlman calls it (*The Myth of Marginality,* 1976) obscures the importance of shantytown inhabitants in defining the nature of the neocolonial city.

To compete successfully in the world market, commodities manufactured in Third World cities have to be less expensive than the comparable items produced in the core. Wage labour in the industrial sector of these cities is considerably cheapened because many services and small commodities that wage labourers require are supplied through the informal economy. As Larissa Lomnitz indicates in *Networks and Marginality: Life in a Mexican Shantytown* (1977), recent rural migrants and shantytown dwellers act as maids, gardeners, and handymen to the industrial workers and the middle class at costs well below what would be charged if the formal sector supplied these services (comparable to domestic labour and baby-sitting supplied well below minimum wage in the core nations).

The informal urban economy never provides security, and the inhabitants of shantytowns in neocolonial cities have had to develop cultural means of survival over the hard times that commonly befall them. Rather than being places of anomie, shantytowns are made up of highly intimate webs of relationship and mutual dependence, based on carefully fostered kinship, ethnic, sectarian, or friendship networks. These networks succor those temporarily out of money and provide some security for those otherwise economically unprotected, who have neither job security nor welfare institutions to fall back on, given the informal sector work that they do.

Shantytowns

These networks, which are in fact adaptations to the exigencies of neocolonial cities, often appear as survivals from the peasant or rural backgrounds of the shantytown dwellers—they are said to be "peasant urbanites" rather than truly urbanized, and this image incorrectly strengthens the notion of their marginality. The tribal identities found among recent urban migrants in African cities are actually instances of "retribalization," a strengthening or redefinition of tribal identity to form networks among urban migrants for mutual aid. Similarly, extended family networks may not disappear in the city; they became wider and stronger among Mexican shantytown inhabitants, for example. New sectarian identities can play an equivalent role: Bryan Roberts in *Cities of Peasants* (1978) shows that the growth of Pentecostal and other Protestant sects in Guatemala fulfills needs for mutual support networks in poor neighbourhoods and for those without kin ties.

Although shantytown inhabitants in the informal economy are impoverished and insecure, it is not certain that they can organize and struggle for better urban living conditions, as did wage labourers in industrial cities. Whereas some scholars have argued for the revolutionary potential of this class, others are persuaded that it does not form a proletariat and will not engage in revolutionary confrontation. The fact that the people who live in shantytowns are "self-employed" and do not enter a wage relationship with the urbanites whom they provide with services apparently limits class antagonisms. Furthermore, both the middle class and shantytown dwellers often perceive their real enemies as the Western imperialist nations or the national government said to be in league with international capital. This perception recognizes that the travails of all classes in the neocolonial city have more to do with external economic relationships in the world economy than class exploitation within the city.

CITIES AND CULTURES

The subject matter of urban anthropology

In the 1970s anthropologists debated whether they should proceed with micro-studies of the city's poor or its recent migrants—an anthropology "in the city," as it was called—or with macro-studies of the city as a whole—an anthropology "of the city." Ten years later the debate was resolved by a tide of studies that focused neither at the micro-level nor at the macro-level but rather at the links in between, that is, the webs of cultural, economic, and political relationship binding the shantytown, ghetto, or neighbourhood to the city and even beyond, to the world economy.

In urban cultures after the establishment of the capitalist world system these webs consist of the economic, political, and cultural strands linking mass-communications cities in the core with neocolonial cities in the Third World into a world system of unequal political and economic relationships. For precapitalist urban cultures these webs consisted of power and wealth inequalities and cultural domination within the urban culture. These different webs effect variant urban cultural roles and cultural forms.

Urban anthropologists in the 1970s also worried over the contribution their studies of urban cultures would make to the general anthropological concept of culture. Oscar Lewis initiated a debate about the nature of culture when he put forward his notion of an urban "culture of poverty." He believed the culture of poverty socialized the poor into political apathy, immediate gratification, broken families, and passive responses to their economic plight, and he argued that the poor could not lose this debilitating culture even if they ceased to be poor. A massive scholarly critique of the culture of poverty concept also exposed the limitations in the traditional anthropological conception of culture on which it was based. This critique argued that the poor's marginality was not a result of their internalized culture but rather of their abject material conditions given the world system (as in the case of the shantytown research cited above). In the face of this critique, the traditional notion of culture—that it was a weighty set of traditions compelling individuals to act in certain ways—gave way to a conception of the constant production of cultures (urban or nonurban) through continual human action—people working with their hands and minds—in response to the material conditions of everyday life. (R.G.F)

**BIBLIOGRAPHY.** For a general account of man and culture, see WILLIAM A. HAVILAND, *Cultural Anthropology,* 5th ed. (1987); RICHARD A. BARRETT, *Culture and Conduct* (1984); MARC J. SWARTZ and DAVID K. JORDAN, *Culture: The Anthropological Perspective* (1980); and ELMAN R. SERVICE, *Primitive Social Organization: An Evolutionary Perspective,* 2nd ed. (1971). The unique capacity for symboling that distinguishes humans from primates is discussed by LESLIE A. WHITE, "The Symbol: The Origin and Basis of Human Behavior," in his *Science of Culture,* 2nd ed., pp. 22–39 (1969); ERNST CASSIRER, *An Essay on Man: An Introduction to a Philosophy of Human Culture* (1944, reprinted 1974); and TERENCE DIXON and MARTIN LUCAS, *The Human Race* (1982). The many conceptions of culture are discussed in A.L. KROEBER and CLYDE KLUCKHOHN, *Culture: A Critical Review of Concepts and Definitions* (1952, reprinted 1978). See also LESLIE A. WHITE and BETH DILLINGHAM, *The Concept of Culture* (1973); and CLIFFORD GEERTZ, *The Interpretation of Cultures: Selected Essays* (1973, reissued 1975). The history of theory and method in social and cultural anthropology is traced in FRED W. VOGET, *A History of Ethnology* (1975). (L.A.W./Ed.)

For information on nomadic hunting-gathering societies, see ELMAN R. SERVICE, *The Hunters,* 2nd ed. (1979); RICHARD B. LEE and IRVEN DEVORE (eds.), *Kalahari Hunter-Gatherers: Studies of the !Kung San and Their Neighbors* (1976), and *Man the Hunter* (1969); LEWIS H. GARRARD, *Wah-To-Yah and the Taos Trail* (1850, reprinted 1982), an account of Plains Indian life; E. ADAMSON HOEBEL, *The Cheyennes,* 2nd ed. (1978); F.R. SECOY, *Changing Military Patterns on the Great Plains (17th Century Through Early 19th Century)* (1953, reprinted 1971), an analysis of the rise of Plains Indian nomadic equestrian societies; and PHILIP DRUCKER, *Cultures of the North Pacific Coast* (1965), an account of sedentary fishing societies. An analysis of the forms of economic exchange among those societies is HELEN CODERE, *Fighting with Property: A Study of Kwakiutl Potlatching and Warfare, 1792–1930* (1950, reprinted 1970). Pastoral societies are depicted by OWEN LATTIMORE, *Inner Asian Frontiers of China,* 2nd ed. (1951, reissued 1967); LAWRENCE KRADER, "Ecology of Central Asian Pastoralism," *Southwestern Journal of Anthropology,* 11(4):301–326 (Winter 1955); and E.E. EVANS-PRITCHARD, *The Nuer* (1940, reprinted 1974). The consequences of plant and animal domestication is discussed in ROBERT J. BRAIDWOOD and GORDON R. WILLEY (eds.), *Courses Toward Urban Life: Archeological Considerations of Some Cultural Alternates* (1962); and modern horticulturalists are analyzed by MARSHALL D. SAHLINS, *Tribesmen* (1968).

Important general works on peasant societies are ROBERT REDFIELD, *Peasant Society and Culture* (1956, reprinted with *The Little Community,* 1973); ERIC R. WOLF, *Peasants* (1966); EMMANUEL LE ROY LADURIE, "Peasants," in *The New Cambridge Modern History,* vol. 12, *Companion Volume,* edited by PETER BURKE (1979), pp. 115–163; and two valuable essay collections: JACK M. POTTER, MAY N. DIAZ, and GEORGE M. FOSTER (eds.), *Peasant Society* (1967); and TEODOR SHANIN (ed.), *Peasants and Peasant Societies* (1971, reprinted 1984). Peasantry in medieval Europe is described and analyzed by BARBARA A. HANAWALT, *The Ties That Bound: Peasant Families in Medieval England* (1986); and GEORGES DUBY, *Rural Economy and Country Life in the Medieval West* (1968, reprinted 1976; originally published in French, 2 vol., 1962).

Studies of peasantries have generated important theoretical arguments. Perhaps most notable is the "dual society" concept proposed by J.H. BOEKE, *Economics and Economic Policy of Dual Societies, as Exemplified by Indonesia* (1953, reprinted 1978). See also J.S. FURNIVALL, *Netherlands India: A Study of Plural Economy* (1939, reprinted 1983); and CLIFFORD GEERTZ, *Agricultural Involution: The Process of Ecological Change in Indonesia* (1963, reprinted 1968). With regard to Latin-American studies, ROBERT REDFIELD was an innovator in comparative field studies, introducing a conception of historical stages in *The Folk Culture of Yucatan* (1941, reprinted 1968). Other important analyses of Latin-American Indian peasantries are GEORGE M. FOSTER, *Tzintzuntzan: Mexican Peasants in a Changing World,* rev. ed. (1979); GEORGE MCCUTCHEN MCBRIDE, *Chile: Land and Society* (1936, reprinted 1971); and CHARLES WAGLEY, *The Latin American Tradition: Essays on the Unity and the Diversity of Latin American Culture* (1968). CONRAD M. ARENSBERG initiated anthropological study of Irish peasantry in *The Irish Countryman* (1937, reissued 1968). On Asia and the Far East, see HSIAO-TUNG FEI, *Peasant Life in China: A Field Study of Country Life in the Yangtze Valley* (1939, reprinted 1980); S.C. DUBE, *India's Changing Villages* (1958, reissued 1967); and THOMAS C. SMITH, *Agrarian Origins of Modern Japan* (1959, reissued 1966). See also DAVID BROKENSHA and MARION PEARSALL (eds.), *The Anthropology of Development in Sub-Saharan Africa* (1969); and GEORGE DALTON, *Economic Anthropology and Development: Essays on Tribal and Peasant Economies* (1970). On the rise of peasantry in relation to urban centres see WALTER GOLDSCHMIDT, *Man's Way: A Preface to the Understanding of Human Society* (1959, reprinted 1966). (E.R.Se./Ed.)

Overviews of the anthropological study of cities and urban cultures can be found in ULF HANNERZ, *Exploring the City* (1980); RICHARD G. FOX, *Urban Anthropology* (1977); and EDWIN EAMES and JUDITH GRANICH GOODE, *Anthropology of the City* (1977). Urban anthropological research papers are collected in GEORGE GMELCH and WALTER P. ZENNER (eds.), *Urban Life* (1980); THOMAS W. COLLINS (ed.), *Cities in a Larger Context* (1980); and IRWIN PRESS and M. ESTELLIE SMITH (eds.), *Urban Place and Process* (1980). The integration of micro-level and macro-level urban research is best seen in studies of industrial or mass-communications cities in advanced capitalist cultures, such as in MANUEL CASTELLS, *The Urban Question* (1977, reprinted 1979; originally published in French, 1972), and *The City and the Grassroots: A Cross-Cultural Theory of Urban Social Movements* (1983); DAVID HARVEY, *Consciousness and the Urban Experience* (1985), and *The Urbanization of Capital* (1985); MICHAEL PETER SMITH (ed.), *Cities in Transformation: Class, Capital and the State* (1984); WILLIAM K. TABB and LARRY SAWERS (eds.), *Marxism and the Metropolis,* 2nd ed. (1984); and DOLORES HAYDEN, *The Grand Domestic Revolution: A History of Feminist Designs for American Homes, Neighborhoods, and Cities* (1981, reprinted 1983). This theoretical integration is also to be found in studies of neocolonial cities and Third World urban culture, such as LARISSA ADLER LOMNITZ, *Networks and Marginality: Life in a Mexican Shantytown* (1977; originally published in Spanish, 1976); ALAN GILBERT and JOSEF GUGLER, *Cities, Poverty, and Development* (1982, reprinted 1984); BRYAN ROBERTS, *Cities of Peasants* (1978, reissued 1979); PETER LLOYD, *A Third World Proletariat?* (1982), and *Slums of Hope? Shanty Towns of the Third World* (1979); WAYNE A. CORNELIUS, *Politics and the Migrant Poor in Mexico City* (1975); OSCAR J. MARTÍNEZ, *Border Boom Town: Ciudad Juarez Since 1848* (1978); and JANICE E. PERLMAN, *The Myth of Marginality: Urban Poverty and Politics in Rio de Janeiro* (1976, reprinted 1979). (R.G.F.)

# Cyprus

The island of Cyprus (Greek Kípros, Turkish Kıbrıs) lies in the eastern Mediterranean, about 40 miles (64 kilometres) south of Turkey, about 60 miles west of Syria, and some 480 miles southeast of mainland Greece. Its maximum length, from Cape Arnauti in the west to Cape Andreas at the end of the northeastern peninsula, is 140 miles; the maximum north–south extent is 60 miles. With an area of 3,572 square miles (9,251 square kilometres), it is the third largest Mediterranean island (after Sicily and Sardinia). The general pattern of its 486-mile coastline is indented and rocky, with long, sandy beaches.

The first human settlement was in the Neolithic period, at a date now calculated as well before 6000 BC. The immigration of settlers from Greece, which began in about 1200 BC, led to the foundation of Greek kingdoms covering almost the whole of the island and to the dominance of the Greek language. Since then Cyprus has come under the influence or control of the various peoples that have exercised power in the eastern Mediterranean—Phoenicians, Assyrians, Egyptians, Persians, the Greek monarchies of Alexander the Great and his successors, the Roman Empire from its successive capitals of Rome and Constantinople, French crusaders, Genoese, Venetians, Turks, and, more recently, the British. Britain maintains two military bases, at Akrotiri and Dhekelia, in the south of the island. They have a combined area of 99 square miles.

The British occupation began in 1878 and ended in August 1960, when Cyprus became independent as the Republic of Cyprus (Greek Kipriakí Demokratía, Turkish Kıbrıs Cumhuriyeti), with its capital at Nicosia. The long-standing conflict between the Greek-Cypriot majority and the Turkish-Cypriot minority intensified; in 1974 an invasion by Turkish troops produced an effective although unrecognized partition of the island and led to the declaration in 1975 of a separate Turkish-Cypriot state in the northern part. The Turkish-Cypriot state made a unilateral declaration of independence in 1983 and adopted the name Turkish Republic of Northern Cyprus. Its independence was recognized only by Turkey.

This article is divided into the following sections:


Physical and human geography 894
- The land 894
  - Relief
  - Drainage
  - Climate
  - Plant and animal life
  - Settlement patterns
- The people 896
  - Ethnic composition
  - Linguistic composition
  - Religions
  - Demographic trends
- The economy 896
  - The economy after independence
  - Effects of partition
  - Resources
  - Agriculture, forestry, and fishing
  - Industry
  - Finance and trade
  - Transportation
- Government and social conditions 897
  - Government
  - Education
  - Health
  - Cultural life

History 898
- Bronze Ages 898
- Greek immigration 898
- External political influences 898
  - Assyrian and Egyptian domination
  - The Persian Empire
  - Hellenistic and Roman rule
  - Byzantine Empire
  - The Lusignan kingdom, Genoese rule, and Venetian rule
  - Ottoman rule
  - British rule
- The Republic of Cyprus 900
  - Turkish invasion
  - Intercommunal talks
  - Declaration of an independent Turkish state

Bibliography 901


## Physical and human geography

### THE LAND

**Relief.** The saucepan shape of Cyprus results from its topography, which, in turn, reflects its geology. The 100-mile-long Kyrenia Mountains, known also as the Pentadaktylos range for the five-fingered peak that is one of its main features, runs parallel to and just inland from the northern coast. It is the southernmost range of the great Alpine–Himalayan chain in the eastern Mediterranean; like much of that extensive mountain belt it is formed largely of thrust masses of Mesozoic limestone.

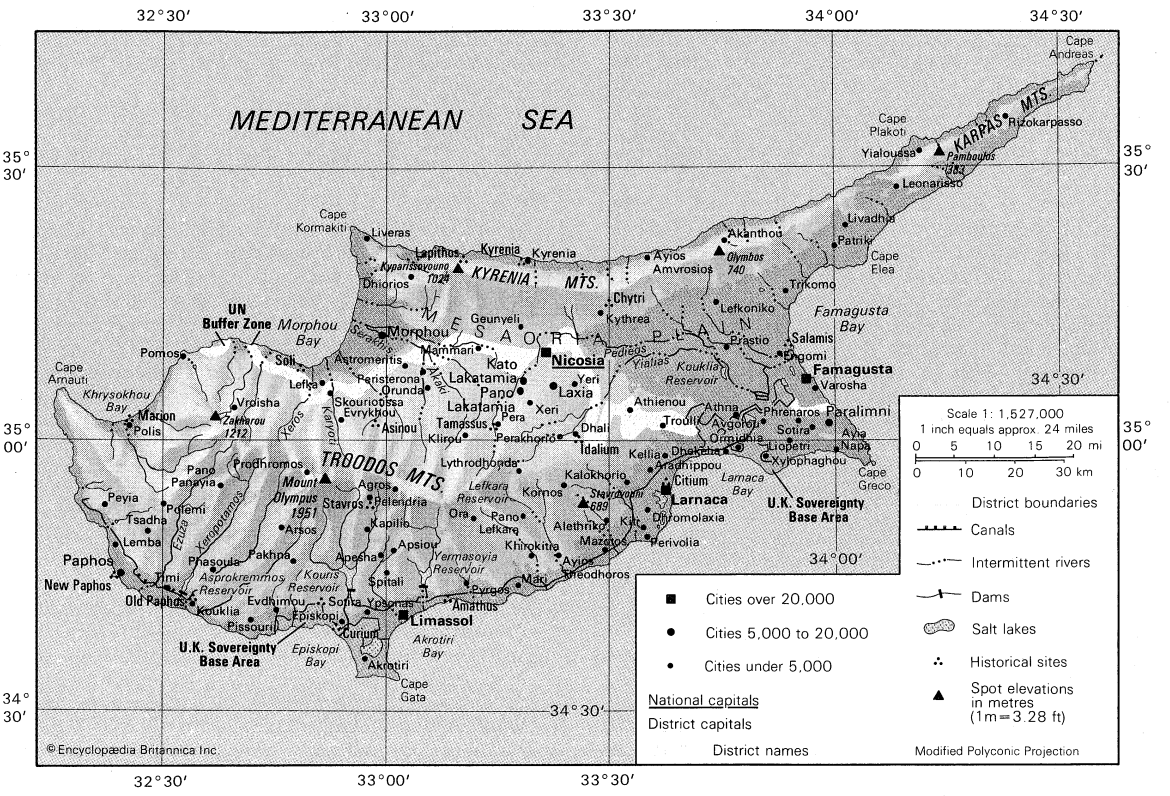
ormation f the roodos Mountains

The Troodos Mountains in the south and southwest are of great interest to geologists, who have concluded that the range, made up of igneous rock, was formed from molten rock beneath the deep ocean (Tethys) that once separated the continents of Eurasia and Afro-Arabia. The range stretches eastward about 50 miles from near the island's west coast to Stavrovouni peak (2,260 feet [689 metres]), about 12 miles from the southeast coast. The range's summit, Mt. Olympus (also called Mt. Troodos), reaches an altitude of 6,401 feet (1,951 metres).

Between the two ranges lies the Mesaoria Plain (its name means "Between the Mountains"). The plain, which is flat and low-lying, extends from Morphou Bay in the west to Famagusta Bay in the east. Roughly in the centre of the plain is the capital, Nicosia (Greek Levkosía; Turkish Lefkoşa). The plain is the principal cereal-growing area in the island. About half of its 465,500 acres are irrigated; the remainder are devoted to dryland farming.

**Drainage.** All of the major rivers in Cyprus originate in the Troodos Mountains. The Pedieos, which is the largest, flows eastward toward Famagusta Bay; the Karyoti flows westward to Morphou Bay; and the Kouris flows southward to Episkopi Bay. The rivers are dependent on winter rainfall; in summer they become dry courses.

**Climate.** Cyprus has an intense Mediterranean climate with a typically strongly marked seasonal rhythm. Hot dry summers from June to September and rainy, rather changeable, winters from November to March are separated by short autumn and spring seasons of rapid change in October and in April and May. Autumn and winter rain, on which agriculture and water supply in general depend, is variable. Average annual rainfall is about 20 inches (500 millimetres). The lowest average precipitation is 14 inches at Nicosia and the highest 41 inches on Mt. Olympus. At Nicosia summer temperatures range between an average daily maximum of 98° F (37° C) and an average daily minimum of 70° F (21° C); in winter the range is between 59° F (15° C) and 41° F (5° C). From December to March the Troodos range experiences several weeks of below-freezing night temperatures.




**MAP INDEX**

**Political subdivisions**

Famagusta .....35 20 N 33 52 E
Kyrenia .....35 15 N 33 16 E
Larnaca .....34 57 N 33 55 E
Limassol .....34 47 N 32 55 E
Nicosia .....35 04 N 33 12 E
Paphos .....34 50 N 32 35 E

**Cities and towns**

Agros .....34 55 N 33 01 E
Akanthou .....35 22 N 33 45 E
Akrotiri .....34 36 N 32 57 E
Alethriko .....34 51 N 33 30 E
Apesha .....34 47 N 32 59 E
Apsiou .....34 48 N 33 01 E
Aradhippou .....34 57 N 33 35 E
Arsos .....34 50 N 32 46 E
Astromeritis .....35 08 N 33 02 E
Athienou .....35 04 N 33 32 E
Athna .....35 03 N 33 47 E
Avgorou .....35 02 N 33 50 E
Ayia Napa .....34 59 N 34 00 E
Ayios Amvrosios ..35 20 N 33 35 E
Ayios Theodhoros .....34 48 N 33 23 E
Dhali .....35 01 N 33 25 E
Dhekelia .....34 59 N 33 44 E
Dhiorios .....35 18 N 33 03 E
Dhromolaxia .....34 52 N 33 35 E
Engomi .....35 09 N 33 53 E
Episkopi .....34 40 N 32 54 E
Evdhimou .....34 41 N 32 46 E
Evrykhou .....35 02 N 32 54 E
Famagusta .....35 07 N 33 57 E
Geunyeli .....35 13 N 33 18 E
Kalokhorio .....34 55 N 33 32 E
Kapilio .....34 50 N 32 57 E
Kato Lakatamia ...35 06 N 33 19 E
Kellia .....34 58 N 33 37 E
Khirokitia .....34 48 N 33 23 E
Kiti .....34 50 N 33 34 E
Klirou .....35 01 N 33 11 E
Kornos .....34 55 N 33 24 E
Kouklia .....34 42 N 32 34 E
Kyrenia .....35 20 N 33 19 E
Kythrea .....35 15 N 33 29 E
Larnaca .....34 55 N 33 38 E
Laxia .....35 06 N 33 22 E
Lefka .....35 07 N 32 51 E
Lefkoniko .....35 15 N 33 44 E
Lemba .....34 49 N 32 25 E
Leonarisso .....35 28 N 34 08 E
Limassol .....34 40 N 33 02 E
Liopetri .....35 00 N 33 53 E
Livadhia .....35 24 N 34 02 E
Liveras .....35 23 N 32 57 E
Lythrodhonda ....34 57 N 33 18 E
Mammari .....35 10 N 33 12 E
Mari .....34 44 N 33 18 E
Mazotos .....34 48 N 33 29 E
Morphou .....35 12 N 32 59 E
Nicosia .....35 10 N 33 22 E
Ora .....34 51 N 33 12 E
Ormidhia .....34 59 N 33 47 E
Orunda .....35 06 N 33 06 E
Pakhna .....34 46 N 32 48 E
Pano Lakatamia .....35 06 N 33 18 E
Pano Lefkara .....34 52 N 33 18 E
Pano Panayia ....34 55 N 32 38 E
Paphos .....34 45 N 32 25 E
Paralimni .....35 02 N 33 59 E
Patriki .....35 22 N 33 59 E
Pelendria .....34 54 N 32 58 E
Pera .....35 02 N 33 15 E
Perakhorio .....35 01 N 33 23 E
Peristerona .....35 08 N 33 05 E
Perivolia .....34 49 N 33 35 E
Peyia .....34 53 N 32 23 E
Phasoula .....34 45 N 32 38 E
Phrenaros .....35 02 N 33 55 E
Pissouri .....34 40 N 32 42 E
Polemi .....34 53 N 32 30 E
Polis .....35 02 N 32 25 E
Pomos .....35 09 N 32 33 E
Prastio .....35 10 N 33 45 E
Prodhromos .....34 57 N 32 50 E
Pyrgos .....34 44 N 33 11 E
Rizokarpasso ....35 36 N 34 23 E
Skouriotissa .....35 05 N 32 53 E
Sotira .....35 02 N 33 57 E
Spitali .....34 46 N 33 00 E
Timi .....34 44 N 32 31 E
Trikomo .....35 17 N 33 52 E
Troulli .....35 02 N 33 37 E
Tsadha .....34 50 N 32 28 E
Varosha .....35 06 N 33 57 E
Vroisha .....35 04 N 32 40 E
Xeri .....35 04 N 33 19 E
Xylophaghou .....34 58 N 33 51 E
Yeri .....35 06 N 33 25 E
Yialoussa .....35 32 N 34 11 E
Ypsonas .....34 41 N 32 58 E

**Physical features and points of interest**

Akaki, *river* .....35 10 N 33 05 E
Akrotiri Bay .....34 38 N 33 05 E
Amathus, *historical site* ....34 42 N 33 08 E
Andreas, Cape ...35 42 N 34 35 E
Arnauti, Cape ....35 06 N 32 17 E
Asinou, *historical site* .....35 02 N 32 58 E
Asprokremmos Reservoir .....34 44 N 32 33 E
Chytri, *historical site* .....35 15 N 33 29 E
Citium, *historical site* .....34 55 N 33 38 E
Curium, *historical site* .....34 40 N 32 53 E
Elea, Cape .....35 19 N 34 04 E
Episkopi Bay .....34 38 N 32 50 E
Ezuza, *river* .....34 44 N 32 27 E
Famagusta Bay ...35 15 N 34 10 E
Gata, Cape .....34 34 N 33 02 E
Greco, Cape .....34 56 N 34 05 E
Idalium, *historical site* .....35 01 N 33 25 E
Karpas Mountains .....35 32 N 34 15 E
Karyoti, *river* .....35 09 N 32 54 E
Khrysokhou Bay .....35 06 N 32 25 E
Kormakiti, Cape ..35 24 N 32 56 E
Kouklia Reservoir .....35 07 N 33 45 E
Kouris Reservoir ..34 45 N 32 55 E
Kyparissovouno, *mountain* .....35 19 N 33 10 E
Kyrenia, *historical site* ....35 20 N 33 19 E
Kyrenia Mountains .....35 16 N 33 30 E
Lapithos, *historical site* ....35 20 N 33 10 E
Larnaca Bay .....34 35 N 33 45 E
Lefkara Reservoir .....34 51 N 33 19 E
Marion, *historical site* .....35 02 N 32 26 E
Mediterranean Sea .....35 30 N 33 00 E
Mesaoria Plain ....35 10 N 33 35 E
Morphou Bay ....35 10 N 32 50 E
New Paphos, *historical site* ....34 45 N 32 24 E
Old Paphos, *historical site* ....34 42 N 32 34 E
Olympos, *mountain* .....35 21 N 33 45 E
Olympus, *mountain* .....34 56 N 32 52 E
Pamboulos, *mountain* .....35 32 N 34 14 E
Pedieos, *river* ....35 09 N 33 45 E
Plakoti, Cape ....35 33 N 34 10 E
Salamis, *historical site* ....35 10 N 33 54 E
Serakhis, *river* ....35 13 N 32 55 E
Soli, *historical site* .....35 08 N 32 48 E
Sotira, *historical site* .....34 43 N 32 52 E
Stavros, *historical site* ....34 54 N 32 58 E
Stavrovouni, *mountain* .....34 53 N 33 26 E
Tamassus, *historical site* ....35 02 N 33 15 E
Troodos Mountains .....34 55 N 32 52 E
Xeropotamos, *river* .....34 42 N 32 33 E
Xeros, *river* .....35 08 N 32 50 E
Yermasoyia Reservoir .....34 45 N 33 05 E
Yialias, *river* .....35 09 N 33 44 E
Zakharou, *mountain* .....35 03 N 32 37 E

**Plant and animal life.** There is a narrow fertile plain along the northern coast where the vegetation is largely evergreen but also includes olive, carob, and citrus trees. The Troodos range has a covering of pine, dwarf oak, cypress, and cedar forest. The southern and western slopes are extensively planted with vines. Between autumn and spring the Mesaoria Plain is green and colourful, with an abundance of wildflowers and flowering bushes and shrubs; there are also patches of woodland in which eucalyptus and many types of acacia, cypress, and lowland pine are found. At the island's western end in the area around Morphou there are orange plantations.

Fossil remains of elephant and hippopotamus have been found in the Kyrenia area, and in classical times there were large numbers of deer and boar, but the only large wild animal now surviving is the *agrino,* a species of wild sheep related to the mouflon of the western Mediterranean. It is under strict protection in a small forested area of the Troodos range. Small game is abundant but keenly hunted. Snakes, in classical times so ubiquitous as to earn the island the name of Ophiussa, "the Abode of Snakes," are now comparatively rare.

Migration route for birds

Cyprus lies on major migration routes for birds, and in spring and autumn many millions pass through. Many species also winter on the island. There are many resident species, including francolin and chukar partridges.

**Settlement patterns.** The Cypriots were traditionally a largely rural people, but a steady drift to the towns began early in the 20th century. The census of 1973 recorded six towns, defined as settlements of more than 5,000 inhabitants, and almost 600 villages. This pattern was altered after the Turkish invasion of 1974 by the need to resettle in the southern part of the island some 180,000 Greek-Cypriot refugees from the Turkish-occupied north. The accommodation built for them was situated mainly in the neighbourhood of the three towns south of the line of demarcation, and especially in the part of the Nicosia suburban area still controlled by the government of the Republic of Cyprus. In contrast, the northern portion of the island is now more thinly populated in spite of the influx of Turkish Cypriots transferred from the south and of immigrants from Turkey.

Cyprus' principal towns

The six towns recorded in the 1973 census, under the undivided republic, were the headquarters of the island's six administrative districts. Of these Kyrenia (Turkish Girne), Famagusta (Greek Ammókhostos, Turkish Maǧusa), and the northern half of Nicosia are to the north of the demarcation line drawn in 1974 and are in Turkish-Cypriot hands.

Limassol, Larnaca, Paphos, and the southern part of Nicosia remained in Greek-Cypriot hands after 1974; the northern part of Nicosia became the administrative centre of the Turkish-Cypriot sector.

## THE PEOPLE

**Ethnic composition.** The people of Cyprus represent two main ethnic groups, Greek and Turkish. The Greek Cypriots, who constitute the majority, are descended from a mixture of aboriginal inhabitants with immigrants from the Peloponnese who colonized Cyprus about 1100 BC and assimilated subsequent settlers up to the 16th century. The Turkish Cypriots are the descendants of the soldiers of the Ottoman army that conquered the island in 1571 and of immigrants from Anatolia brought in by the Sultan's government shortly thereafter. Since 1974 additional immigrants from Anatolia, with their families, have been brought in to work vacant land and increase the total labour force.

**Linguistic composition.** The language of the majority is Greek and of the minority Turkish. English is widely spoken and understood as a second language. Illiteracy is low, thanks to the excellence of the educational system.

**Religions.** The Greek Cypriots are Eastern Orthodox Christians. Their church, the Church of Cyprus, is autocephalous—*i.e.,* not under the authority of any patriarch; this privilege was granted to Archbishop Anthemius in AD 488 by the Byzantine emperor Zeno. Under the Ottoman Empire the archbishop of the Church of Cyprus was made responsible for the secular as well as the religious behaviour of the Orthodox community and given the title ethnarch. The Turkish Cypriots are Sunnī Muslims. There are also Maronites, Armenians, Roman Catholics, and Anglicans on the island.

Emigration of Cypriots

**Demographic trends.** Cypriots at times have emigrated in large numbers, and it is estimated that as many live abroad as on the island itself. The great majority of emigrants has always gone to the United Kingdom and the rest mainly to the English-speaking countries: Australia, South Africa, the United States, and Canada. Waves of heavy emigration followed the negotiation of independence in 1960 and the Turkish invasion in 1974. As a result of emigration and other factors, such as war losses and a temporary decline in fertility, the population decreased by about 5 percent between mid-1974 and 1977. The years since 1974 also have been marked by an increase in persons leaving the island in search of work, especially in the Middle East.

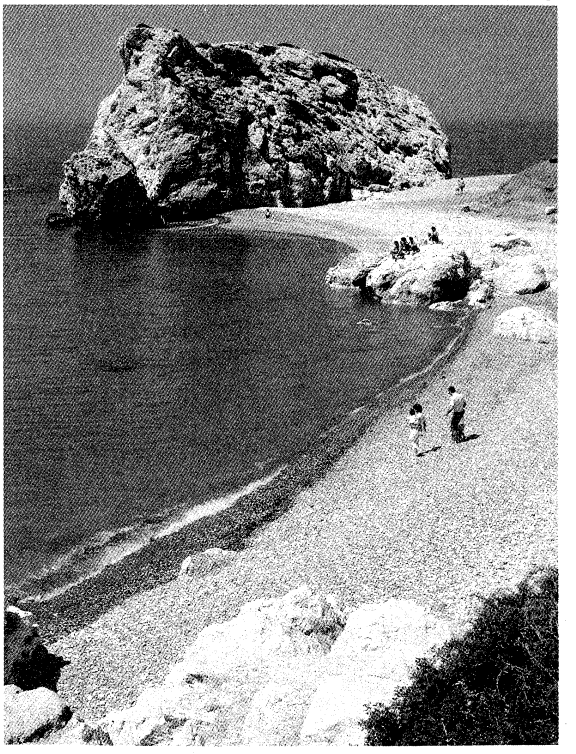
## THE ECONOMY

**The economy after independence.** Between 1960 and 1973 the Republic of Cyprus, operating a free enterprise economy based on agriculture and trade, achieved a standard of living higher than most of its neighbours, with the exception of Israel. This progress was substantially assisted by various agencies of the United Nations, operating through the UN Development Program. Generous financial assistance was given by the World Bank and the International Monetary Fund in the form of loans for specific development projects (electricity supply, port development, and sewerage, among others). Aid was also made available by individual foreign countries. Experts were provided to advise on economic planning and to initiate productive projects, and training for Cypriot specialists was encouraged by scholarships and grants. During this period the gross domestic product grew at an average annual rate of over 7 percent, and per capita national income by about 6 percent annually. Agricultural production doubled; industrial production and exports of goods and services more than trebled. Tourism became the largest single earner of foreign exchange.

Foreign economic aid

**Effects of partition.** The Turkish occupation of 37 percent of the country in 1974, involving the displacement

© Berlitz—CLICK/Chicago



Petra tou Romiou, the legendary site of Aphrodite's emergence from the sea, on the southwest coast of Cyprus near Old Paphos.

of about a third of the population, dealt a serious blow to economic development. Losses of land and personal property in the occupied areas were very great. The gross domestic product of the Greek-Cypriot sector dropped sharply, the reduction amounting to 33 percent (at constant 1973 prices) between 1973 and 1975. By vigorous efforts real growth was resumed in the area left under the control of the government of the Republic of Cyprus, and between 1975 and 1983 the annual rate of growth was estimated to average about 8 percent.

The Turkish-occupied area did not enjoy the same prosperity, and its economy was supported by subsidies from the Turkish government. Trade between the two areas ceased and the two economies became entirely independent, although the southern zone continued to supply the northern with certain services such as electricity and the northern zone supplied water to the south.

Copper

**Resources.** Cyprus was for many centuries a noted producer of copper; in Greek the name of the island and the name of the metal are identical. As early as 2500 BC its mines were being exploited, and traces of prehistoric and Roman workings and surface slags are still to be seen. With the discovery of other sources the mines remained neglected for many centuries until they were reopened shortly before World War I. They were exploited more seriously from 1925 until they were closed by the Great Depression of the 1930s. After World War II they were brought back into production, and since then copper and other minerals—iron pyrites, asbestos, gypsum, chrome ore—have contributed moderately to the export trade. Reserves of copper ore have declined but there are substantial reserves of asbestos, chrome, gypsum, and iron pyrites. There are also extensive quarries of good building stone. The island's most important copper mines are located in the area of Skouriotissa in the Turkish-occupied zone.

**Agriculture, forestry, and fishing.** Of the arable land on the island, about one-fourth is irrigated, mainly in the Mesaoria Plain and around Paphos in the southwest. Pastures occupy about 10 percent of the total land area. Landholdings are generally small, highly fragmented, and dispersed under traditional laws of inheritance. A program of land consolidation was enacted in 1969, but it met with resistance, particularly from Turkish-Cypriot landowners, and was only very slowly implemented.

The major crops of the Greek-Cypriot sector include grapes, deciduous fruits, vegetables, olives, and carobs. The area under Turkish occupation produces the bulk of the country's citrus fruits, wheat, barley, carrots, tobacco, and green fodder.

Livestock—especially sheep, goats, and pigs—and livestock products historically have accounted for about one-third of the island's total agricultural production. Some cattle are also raised.

Cyprus was once famous for its extensive forests, but the demand for timber for shipbuilding by successive conquerors from the 7th century BC onward, and extensive felling for building and for fuel, have destroyed the greater part. Under the British administration a vigorous policy of conservation and reforestation was pursued, and the Cyprus Forestry College was established at Prodhromos, on the western slopes of Mt. Olympus. Forests cover some 520 square miles, most of them being found in the mountain areas and in the Paphos district.

The fishing industry is small, in part because coastal waters are deficient in nutrients and associated plankton. Although the industry has shown some growth in the Greek-Cypriot sector, most fish is imported.

Manufacturing and tourism

**Industry.** Resources of raw materials on Cyprus are very limited, restricting the scope for industrial activity. Before the partition of the island most manufacturing was of goods produced for the domestic market by small, owner-operated plants, and a considerable number of those plants were located in the area that was occupied by the Turks in 1974. Industries in the Republic of Cyprus were subsequently reoriented toward production for export, and a number of larger factories were built in the south. Petroleum refining, cement and asbestos pipe production, and thermal electricity production are the republic's heavy industries, and its light industries produce goods such as clothing, footwear, and machinery and transport equipment.

Tourism became one of Cyprus' major industries after 1960. About 65 percent of tourist accommodation, however, was in the portion of the island that was occupied by the Turks in 1974. After partition the tourist trade recovered rapidly in the Greek-Cypriot sector: to counter the loss of Kyrenia and the Famagusta–Varosha area, which had been the leading seaside resorts, the southern coastal towns of Limassol, Larnaca, and Paphos were further developed to accommodate tourists.

**Finance and trade.** The Republic of Cyprus began to expand financial services, including offshore banking, in 1982. Light manufactures, particularly clothing and footwear, and foodstuffs, including potatoes and citrus fruit, constitute the Republic of Cyprus' major exports. Machinery and transport equipment, petroleum and petroleum products, and foodstuffs and live animals are imported. Chronic trade deficits are offset by receipts from tourists, remittances sent home by expatriate Greek Cypriots, and receipts from the British military bases on the island. In the Turkish sector, citrus fruits, potatoes, and carobs are the principal exports.

**Transportation.** In Roman times the island had a well-developed road system, but by the time of the British occupation in 1878 the only carriage road was between Nicosia and Larnaca. An extensive new road network was built under the British administration. A narrow-gauge public railway proved uneconomical and was closed in the early 1950s. Because there are no public railways, inland travel depends entirely upon roads, and motor transport has greatly increased.

International air services provide connections to all parts of Europe and the Middle East and to some points in Africa. Nicosia International Airport was closed in 1974, and the airport at Larnaca was developed in its stead. An airport at Paphos, opened in 1985, is also used for international flights. An airport at Geçitiikale (Lefkoniko) in the Turkish-occupied sector is used by flights coming from or through Turkey.

There is no coastal shipping, and much of the merchant marine registered to Cyprus is foreign-owned. The great bulk of the island's international trade remains seaborne, however, the main ports being Limassol and Larnaca; Turkish shipping uses Famagusta.

## GOVERNMENT AND SOCIAL CONDITIONS

**Government.** The constitution of the Republic of Cyprus, adopted in 1960, provided that the executive power be exercised by a Greek-Cypriot president and a Turkish-Cypriot vice president, elected to five-year terms by universal suffrage, and that there be a Council of Ministers (Cabinet) comprising seven Greek-Cypriot and three Turkish-Cypriot members. There was also to be an elected House of Representatives with 50 seats, divided between Greek and Turkish Cypriots in the proportion of 35 to 15 and elected for five years.

Objections to the constitution of 1960

The constitution, derived from the negotiations in Zürich in 1959 between representatives of the governments of Greece and Turkey, did not inspire enthusiasm among the citizens of the new republic, however. The Greek Cypriots, whose struggle against the British had been for enosis (union with Greece) and not for independence (see below *History: British rule*), regretted the failure to achieve this national aspiration. As a result it was not long after the establishment of the republic before the Greek-Cypriot majority began to regard many of the provisions, particularly those relating to finance and to local government, as unworkable. Proposals for amendment were rejected by the Turkish government and, after the outbreak of fighting between the two Cypriot communities in late 1963, the constitution went largely into abeyance. In the territory controlled by the government of the Republic of Cyprus after the Turkish occupation of 1974, the constitution's provisions are considered as still in force where practicable; the main formal change has been the gradual increase of the number of seats in the House of Representatives, all of which are held by Greek Cypriots.

On the Turkish side of the demarcation line there have

been, since 1974, an elected president, prime minister, and legislative assembly, all serving five-year terms of office. A new constitution was approved for the Turkish Republic of Northern Cyprus by its electorate in 1985.

*Local government.* Local government in the Republic of Cyprus is at the district, municipal, rural municipality, and village level. District officers are appointed by the government; local councils are elected, as are the mayors of municipalities.

*Justice.* The legal code of Cyprus is based on Roman law. In the Greek-Cypriot zone judges are appointed by the government, but the judiciary is entirely independent of the executive power. There are a supreme court and an appeals court, district assize courts handling criminal matters, and district courts exercising summary jurisdiction. The Turkish-Cypriot zone has a similar system of justice.

*Political parties.* The oldest established political party in the Republic of Cyprus is the Anorthotiko Komma Ergazomenou Laou (AKEL; Progressive Party of the Working People), founded in 1941. It is a pro-Moscow Communist party, and it controls the principal trade union federation; its share of the vote in the first 25 years of the Republic of Cyprus was usually in the neighbourhood of 30 percent. Other parties have had varying success. Among them are the Eniea Demokratiki Enosis Kyprou (Cyprus National Democratic Union), a Socialist party; the Enosi Kentrou (Centre Union); the Demokratikos Synagermos (Democratic Rally); and the Demokratiko Komma (Democratic Party). In the Turkish-Cypriot zone the major parties are the Ulusal Birlik Partisi (National Unity Party) and the Toplumcu Kurtuluş Partisi (Communal Liberation Party).

Com-
pulsory
elementary
education

**Education.** Six grades of free and compulsory elementary education are provided for children beginning at age five. At least three years of the five-year secondary education program are free, and all secondary education at technical schools is free. Post-secondary facilities include schools for teacher training, technical instruction, hotel and catering training, nursing, and midwifery. The education system in the Turkish sector is administered separately. Cyprus has no university, but many students attend universities abroad, especially in Greece, Turkey, Britain, or the United States.

**Health.** Health standards are high because of a favourable climate and well-organized public and private health services. Since the eradication of malaria shortly after World War II and, later, of echinococcosis (hydatid disease), the island has been free from major diseases.

**Cultural life.** The very ancient cultural traditions of Cyprus are maintained partly by private enterprise and partly by government activity, especially on the part of the Cultural Service of the Republic of Cyprus' Ministry of Education. The Cultural Service publishes books and awards prizes for literature. Mobile libraries operate in rural areas. The government-sponsored Cyprus Theatrical Organization stages plays by contemporary Cypriot dramatists as well as classical works. The ancient theatres of Salamis, Curium, and Soli have been restored and are used for the staging of a variety of plays, and a Greek theatre has been built at Nicosia.

Literature, theatre, and art

Many painters and sculptors work in Cyprus, and the Cultural Service keeps the state's collection of modern Cypriot art on permanent exhibition. In the village of Lemba near Paphos the Cyprus College of Art runs courses for postgraduate art students. Government encouragement is given to young composers.

Television and radio are controlled by the semigovernmental Cyprus Broadcasting Corporation and are financed by advertising. The Turkish sector receives broadcasts from Turkey. Languages used are Greek, Turkish, English, and Armenian. Many daily and weekly newspapers are published in Greek, Turkish, and English.

For statistical data on the land and people of Cyprus, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL. (H.W.G./D.W.S.H.)

## History

The earliest evidence of human habitation in Cyprus comes from the Neolithic Period. The settlement at Khirokitia (near the southern coast), which is now dated to well before 6000 BC, is one of the most remarkable Neolithic communities ever excavated in Europe. It was a town of about 2,000 inhabitants, living in well-built round stone houses of two stories. The Khirokitians made little use of pottery, using stone, and presumably wood, for utensils and stone for tools. The presence of flakes of obsidian, which is not native to the nonvolcanic island of Cyprus, is the only sign of contact with other cultures. Khirokitia and a few smaller associated settlements appear to have died out after a few centuries, leaving the island uninhabited again for some 2,000 years. The beginning of the next period of habitation, known as the Sotira culture, is dated to between 4500 and 4000 BC; small villages of this culture are found not only at Sotira (near the southern coast, north of Curium) but also in the Kyrenia range. Small ornaments of picrolite (a variety of soapstone) and a progressively more attractive pottery distinguish the Sotira culture; toward the end of the period copper came into use.

Khirokitia and Sotira settlements

### BRONZE AGES

After the Chalcolithic Age, dating from 3000 to 2500 BC, began the Bronze Ages, which in Cypriot archaeology are treated as separate from the Chalcolithic and which lasted for about 1,500 years. The Middle Bronze Age (1900–1600 BC) produced several styles of well-made and competently decorated pottery, and its bronze implements show a well-advanced craftsmanship; imports from Crete, Anatolia, Syria, and Egypt prove that external trade had begun. It has been conjectured that the name Alashiya or Alasia, which occurs in Hittite and Egyptian records in connection with the supply of copper, refers to Cyprus. These trade links probably account for the foundation of the new settlements in the east of the island that were to develop into international urban emporiums.

The Late Bronze Age (1600–1050 BC) was one of the most formative periods of the life of ancient Cyprus. The island's international contacts extended from the Aegean Sea to the Levant and the Nile Delta. (Thutmose III of Egypt claimed Cyprus as one of his conquests in about 1500 BC.) Writing, in the form of a linear script known as Cypro-Minoan, was borrowed from Crete. Cypriot craftsmen were distinguished for fine jewelry, ivory carving, and bronze figures. From about 1400 BC a profusion of Mycenaean pottery was imported from mainland Greece, and it is possible that Mycenaean artists accompanied the merchants. After 1200 BC, with the collapse of Mycenaean civilization, there is evidence of Greek immigration from the Peloponnese. The principal city, and port, was Engomi (west of Famagusta); its massive city walls and houses of hewn stone are evidence of a high degree of prosperity.

### GREEK IMMIGRATION

The mass immigration of Greek-speaking peoples from the Peloponnese began with the Iron Age (1100–700 BC). From the start of the 1st millennium the Greek language has been dominant in Cyprus; the fact that the dialectal form in which it first appears is known as Arcado-Cypriot confirms traditions of the Peloponnesian origin—and specifically of the Arcadian origin—of the immigrants. They founded new cities, which became the capitals of six ancient Greek kingdoms on Cyprus: Curium (Greek Kourion), Paphos, Marion, Soli (Greek Soloi), Lapithos, and Salamis. In about 800 BC a Phoenician colony was founded at Citium (Greek Kition), near modern Larnaca. The colony was a dependency of the mother city, Tyre. A seventh kingdom, that of Amathus, remained for some time under the control of the earlier indigenous inhabitants; the language used there was called Eteo-Cypriot ("True Cypriot") by the Greeks. Amathus was active politically, especially in external trade relations. The later Iron Age was a period of advancing civilization, as evidenced by the spectacular chariot-burials of the royal family of Salamis, which so closely match descriptions in the Homeric poems as to suggest inspiration by them.

Immigration from the Peloponnese

### EXTERNAL POLITICAL INFLUENCES

**Assyrian and Egyptian domination.** In 709 BC Sargon II of Assyria erected a stela at Citium recording the fact that

seven Cypriot kings had paid him homage; subsequent Assyrian documents speak of 11 tributary kingdoms, the seven already mentioned plus Citium, Kyrenia, Tamassos, and Idalium. This subordination to Assyria, probably rather nominal, lasted until about 663 BC. For the next hundred years Cyprus enjoyed a period of complete independence and exuberant development. Epic poetry was greatly popular, as it had always been, and much was written on the island; Stasinus of Cyprus, credited with the authorship of the lost epic poem *Cypria,* was reckoned among the most important poets in this style in the 7th century. Bronze work and ironwork, a spirited style of ceramic decoration, and delicate jewelry and ivory work are characteristic of this period; among outstanding works are the sumptuous ivory throne and bedstead excavated from a royal tomb at Salamis dated from about 700 BC.

Egyptian dominance

When the Assyrian Empire finally broke up at the end of the 7th century BC, Egypt, under the Saite dynasty, became the predominant power in the eastern Mediterranean. In about 569 BC the Cypriot kingdoms recognized the pharaoh Ahmose II as their overlord. Direct Egyptian influence was not always apparent, although many limestone sculptures reproduce Egyptian conventions in dress and some statues are directly inspired by Egyptian models. A more important influence in the last years of the Archaic period (750–475 BC) came from the artistic schools of Ionia. From the same source probably came the inspiration for the issue of coinage; the first Cypriot coins were struck for Euelthon, king of Salamis (560–525 BC).

**The Persian Empire.** In 525 BC the Cypriot kings transferred their allegiance to the Achaemenid (Persian) conquerors of Egypt. The Cypriots retained their independence until the accession of Darius I (522 BC) but were then incorporated into the fifth satrapy of the Persian Empire. When the Ionians revolted in 499 BC all the kingdoms of Cyprus except Amathus joined them; the revolt was suppressed in about a year's campaigning, culminating in sieges of Paphos and Soli. In Xerxes I's invasion of Greece in 480 BC the Cypriot kings, like the Ionians, contributed naval contingents to his forces. During the 5th century Cyprus remained under Persian rule in spite of a major Athenian expedition there in 450/449 BC. Evagoras, who became king of Salamis in 411 BC, maintained a pro-Hellenic policy, with some help from Athens, and succeeded in extending his rule over a large part of the island. He was defeated by the Persians in 381 BC and was assassinated in 374 BC. After the victory of Alexander the Great over the last Achaemenid ruler, Darius III, at Issus in 333 BC, the Cypriot kings rallied to Alexander and assisted him at the siege of Tyre. During the period from 475 to 325 BC, known conventionally as the Classical Period, Cypriot art came under strong Attic influence.

**Hellenistic and Roman rule.** Alexander allowed the Cypriot kingdoms to continue but took from them the right of coinage. After his death in 323 BC Cyprus was contested by his successors; the eventual victor was Ptolemy I of Egypt, who suppressed the kingdoms and made the island a province of his Egyptian kingdom. He forced the last king of Salamis, Nicocreon, to commit suicide in 310 BC, together with all his family; their cenotaph, a particularly fine specimen containing ornaments and clay effigies of the royal families, has been discovered. For two and a half centuries Cyprus remained a Ptolemaic possession, ruled by a strategus, or governor-general.

Rule of the Ptolemies

*Cyprus as a Roman province.* Cyprus was annexed by the Roman Republic in 58 BC and, along with Cilicia on the coast of Anatolia, was made into a Roman province. The orator and writer Cicero was one of its first proconsuls. Cyprus was briefly retroceded to Cleopatra VII of Egypt by Julius Caesar, and this status was confirmed by Mark Antony, but after the victory of Caesar's heir, Octavian (subsequently the emperor Augustus), over Mark Antony and Cleopatra at Actium in 31 BC it became a Roman possession again. It was originally administered as part of the "imperial" province of Syria but became a separate "senatorial" province in 22 BC in consequence of the constitutional settlement of the previous year. Its governors resumed the old republican title of proconsul, although there is evidence that Augustus could, and on one occasion did, influence the Senate's choice. For the next 600 years Cyprus enjoyed a profound peace, disturbed only by occasional earthquakes and epidemics and by a Jewish uprising put down by a lieutenant of the future emperor Hadrian in AD 116. Many large public buildings were erected: among them were a gymnasium and theatre at Salamis, a theatre at Curium, and the governor's palace at Paphos.

Introduction of Christianity

*Early Christianity.* Undoubtedly the most important event in the Roman period was the introduction of Christianity. The Apostle Paul, accompanied by Barnabas (later St. Barnabas), a native of the Cypriot Jewish community, preached there in about AD 45 and converted the proconsul, Sergius Paulus. By the time of Constantine I the Great, Christians were numerous in the island and may have constituted a majority.

**Byzantine Empire.** After the division of the Roman Empire (AD 395) Cyprus remained subject to the Eastern, or Byzantine, Empire at Constantinople, being part of the Diocese of the Orient governed from Antioch. In ecclesiastical matters, however, the Church of Cyprus was autocephalous—*i.e.,* independent of the Patriarchate of Antioch—having been given that privilege in 488 by the emperor Zeno. The archbishop received the rights, still valued and practiced, of carrying a sceptre instead of a crozier and writing his signature in ink of imperial purple.

There was a break in direct rule from Constantinople in 688 when Justinian II and the caliph 'Abd al-Malik signed an unusual form of treaty neutralizing the island, which had been subject to Arab raids. For almost 300 years Cyprus was a kind of condominium of the Byzantine Empire and the Caliphate, and although the treaty was frequently violated by both sides, the arrangement lasted until 965, when the emperor Nicephorus II Phocas gained Cyprus completely for the Byzantines.

This appears to have been a period of modest prosperity. A remarkable mosaic of the 6th century, at Kiti, is the best example of Eastern Roman art of that date, comparable with works at Ravenna in Italy. Another equally remarkable mosaic of roughly the same date, at Lythrangomi, was destroyed in 1974. Wall paintings demonstrate close contact with Constantinople: those at Asinou, in particular, are noteworthy as being the earliest of an unparalleled series of mural paintings showing successive developments of Byzantine art.

In about 1185 a Byzantine governor of Cyprus, Isaac Comnenus, rebelled and proclaimed himself emperor. Isaac resisted attacks from the Byzantine emperors Andronicus I Comnenus and Isaac II Angelus, but in 1191, on engaging in hostilities with an English crusader fleet under King Richard I the Lion-Heart, he was defeated and imprisoned. The island was seized by Richard, from whom it was acquired by the crusading order of the Knights Templar; because they were unable to pay his price he took it back and sold it to Guy of Lusignan, the dispossessed king of Jerusalem.

**The Lusignan kingdom, Genoese rule, and Venetian rule.** Guy, who called himself lord of Cyprus, invited families that had lost their lands in Palestine after the fall of Jerusalem to take up land in Cyprus. He thereby laid the basis for a feudal monarchy that survived to the end of the Middle Ages. His brother and successor, Amalric, obtained the title of king from the Holy Roman emperor Henry VI. The earliest kings of the Lusignan dynasty were involved in the affairs of the small territory still left to the Kingdom of Jerusalem, and this commitment caused a heavy drain on the resources of Cyprus until the kingdom was extinguished in 1291 with the fall of Acre. Over the next hundred years Cyprus enjoyed a reputation in Europe for immense riches where its nobles and merchants (especially Famagusta merchants) were concerned. Famagusta's opulence derived from its position as the last entrepôt for European trade adjacent to the Levant.

The kings of Cyprus had kept alive the crusading idea, and the island remained a base for counterattack against the Muslims. In 1361 the Cypriot king Peter I (reigned 1359–69) devoted himself to the organization of a crusade. He captured Adalia (Antalya) on the Cilician coast of Anatolia, and in 1365, after collecting money and mer-

ivalry for ontrol of 'yprus' ade

cenaries in western Europe, seized and sacked Alexandria. He was not able to maintain the conquest, however, and was soon forced to abandon Alexandria. At his son's accession rivalry between Genoa and Venice, vying for control of Cyprus' valuable trade, resulted in Genoa's seizing Famagusta and holding it for almost a hundred years. This led to a rapid decline in the island's prosperity. In 1426 a marauding expedition from Egypt overran the island, which from then on paid tribute to Cairo. The last Lusignan king, James II (reigned 1460–73), a bastard of the royal house, seized the throne with the help of an Egyptian force and in 1464 expelled the Genoese from Famagusta. He married a Venetian noblewoman, Caterina Cornaro, who on his death, which was followed by that of his posthumous son, succeeded him as the last queen of Cyprus (1474–89). During her reign she was under strong Venetian pressure and was eventually persuaded to cede Cyprus to the Venetian Republic. It remained a Venetian possession for 82 years until its capture by the Ottomans.

Many noteworthy buildings survive from the Lusignan and Venetian periods, in particular the Gothic cathedrals at Nicosia and Famagusta and the Abbey of Bellapais near Kyrenia. There are other Gothic churches throughout the island. Orthodox Christians also built numerous churches in a distinctive style, one often influenced by the Gothic, and the interiors illustrate the continued development of Byzantine art. Cyprus also has imposing examples of Medieval and Renaissance military architecture, such as the castles of Kyrenia, St. Hilarion, Buffavento, and Kantara and the elaborate Venetian fortifications of Nicosia and Famagusta.

**Ottoman rule.** A Turkish invading force landed in Cyprus in 1570 and captured Nicosia; the following year Famagusta fell after a long siege. Ottoman rule lasted more than three centuries. The Latin church was suppressed and the Orthodox hierarchy restored; with the abolition of feudal tenure the Greek peasantry acquired inalienable and hereditary rights to land. Taxes were at first reduced but very soon were greatly increased and arbitrarily levied. In the 18th century the Orthodox archbishop was made responsible for tax collection.

About 20,000 Muslims (including the garrison, nominally 3,666 strong) were settled in the island in the immediate aftermath of the Ottoman conquest. Cyprus was an unimportant province to the sultans; its governors were slothful, inefficient, occasionally oppressive, and always venal. There were Turkish uprisings in 1764 and 1833; in 1821 the Orthodox archbishop was hanged on suspicion of sympathy with the rebellion in mainland Greece. The sultanate's various imperial proclamations in the 19th century promising reform had no effect in Cyprus, where local opposition prevented their application.

**British rule.** The Cyprus Convention of 1878 between Britain and Turkey provided that Cyprus, while remaining under Turkish sovereignty, should be administered by the British government. Britain's aim in occupying Cyprus was to secure a base in the eastern Mediterranean for possible operations in the Caucasus or Mesopotamia as part of the British guarantee to preserve the Sultan's Asian possessions from threat by Russia. In 1914, however, Britain and Turkey being at war, the former proclaimed the island annexed; Turkish recognition was granted under the Treaty of Lausanne (1923), and the position was regularized in 1925 when Cyprus was declared a crown colony.

British occupation was initially welcomed by the Greek population, who from the start expected the British to transfer Cyprus to Greece. The Greek Cypriots' demand for enosis (union with Greece) and a corresponding hostility to it on the part of Turkish Cypriots constituted almost the sole division in politics; almost annual petitions demanding enosis were matched by counter-petitions, and demonstrations, from the Turkish Cypriots. An offer to transfer the island had been made in 1915, on condition that Greece fulfill its treaty obligations toward Serbia when attacked by Bulgaria. The Greek government refused and the offer was not renewed. In 1931 the demand for enosis led to riots in Nicosia.

Demands or union with Greece

Cyprus was untouched by World War II apart from a few air raids. In 1947 the governor, in accordance with the British Labour Party's declaration on colonial policy, published proposals for greater self-government. They were rejected in favour of the slogan "enosis and only enosis." In 1955 Lieut. Col. Georgios Grivas (known as Dighenis), a Cypriot who had served as an officer in the Greek Army, began a concerted campaign for enosis. His National Organization of Cypriot Struggle (Ethnikí Orgánosis Kipriakoú Agónos; EOKA) bombed public buildings and attacked and killed opponents of enosis, both Greek-Cypriot and British. Proposals for self-government were put forward at different times; the most advanced were those of the British jurist Lord Radcliffe in 1956. All were rejected and the attacks continued. In March 1956 the archbishop, Makarios III, who as ethnarch considered it his duty to champion the national aspirations of the Greek Cypriots, was deported to the Seychelles. He was released from exile in March 1957 and left the Seychelles in April, but, being forbidden to return to Cyprus, he made his headquarters in Athens. By this time the operations of EOKA were much reduced, but on the other hand the Turkish-Cypriot minority, led by Fazıl Küçük, began to express alarm and demanded either retrocession to Turkey or partition. Public opinion in Greece and Turkey was much aroused in support of the two communities, resulting in riots and expulsions of Greek residents in Turkey. Frequent recourse to the United Nations produced no agreed solution.

The decisive step was taken by the Greek and Turkish governments, which in February 1959 reached agreement between themselves in Zürich. Later the same month, at a conference in London, the Greek–Turkish compromise was accepted by the British government and by representatives of the Greek-Cypriot and Turkish-Cypriot communities, led by Makarios and Küçük, respectively. In 1960 it was ratified by treaties agreed to in Nicosia. Cyprus became an independent republic, with Britain retaining sovereignty over the two military bases at Akrotiri and Dhekélia. According to the terms of the treaties, the new republic would not participate in a political or economic union with any other state, nor would it be subject to partition. Greece, Turkey, and Britain guaranteed the independence, integrity, and security of the republic, and Greece and Turkey undertook to respect the integrity of the areas remaining under British sovereignty. Makarios became president and Küçük vice president in elections held in December 1959. Decisions of the council of ministers would be binding on the president and vice president, either of whom could, however, exercise a veto in matters relating to security, defense, and foreign affairs. Cyprus' ethnic balance was to be reflected not only in the legislature but also in the civil service and army, in which Turkish Cypriots would make up 30 percent and 40 percent, respectively. A joint Greek and Turkish military headquarters was to be established to train the Cypriot army.

Negotiations for independence

### THE REPUBLIC OF CYPRUS

The first general election took place on July 31, 1960. Of the 35 seats allotted to the Greek Cypriots, 30 were won by supporters of Makarios, five by agreement being allotted to the Communist-led Progressive Party of the Working People (AKEL). All 15 Turkish-Cypriot seats were won by supporters of Küçük. The republic came into being on Aug. 16, 1960, and Cyprus was admitted as a member of the United Nations. The British government agreed to pay £12,000,000 in financial assistance over five years. Cyprus was admitted to membership in the Commonwealth in March 1961.

The difficulties experienced in implementing some of the complicated provisions of the constitution, particularly over local government and finance, led Makarios late in 1963 to propose to Küçük 13 amendments. These were rejected by the Turkish government and the Turkish Cypriots, and in the next month fighting broke out between the two Cypriot communities. As a result, the area controlled by the Turkish Cypriots was reduced to a few enclaves and Nicosia was divided by a cease-fire line, policed to begin with by British troops. In March 1964 the UN Security Council agreed to send to Cyprus a multinational force known as the United Nations Peace-Keeping Force in Cyprus, or UNFICYP. Its mandate was

Conflict between Greek Cypriots and Turkish Cypriots

extended repeatedly in the course of the continuing conflict. In 1964 intensified fighting in the northwest caused the Turkish air force to intervene; at the same time a full-scale invasion was threatened. Contingents of troops from Greece and Turkey were brought into the island clandestinely together with officers to command and train the forces raised by the two communities. Grivas, who had been promoted to lieutenant general in the Greek army, returned from Greece to command the Greek-Cypriot National Guard. In 1967 an incident in the southeast led to a Turkish ultimatum to Greece, backed by the threat of invasion. The military junta then ruling Greece complied by withdrawing the mainland contingents together with General Grivas. An uneasy peace was established, but intercommunal talks failed to produce a solution.

Makarios was reelected president in 1968 by an overwhelming majority, and in 1973 his reelection was not even contested. He played a conspicuous role on the world political stage, being especially associated with the nonaligned movement. Although Makarios had originally been a leader in the campaign for enosis, he was thought by many Greek Cypriots and mainland Greeks to be content with Cyprus' independence after he became president. Dissidents angered by that perception are assumed to have been responsible for attempts to assassinate him in 1970 and 1973. In the latter year Makarios was denounced by the three suffragan bishops ecclesiastically subordinate to him; they demanded that he renounce the presidency because it was in conflict with his role as a spiritual leader. Makarios circumvented them, however, by calling a synod of the Eastern Orthodox churches presided over by the Patriarch of Alexandria. Meanwhile Grivas had returned secretly to Cyprus in 1971 to resume the campaign for enosis under a newly formed EOKA-B; he died, still in clandestinity, in Limassol in 1974, at the age of 75.

**Turkish invasion.** On July 15, 1974, a detachment of the National Guard, led by officers from mainland Greece, launched a coup aimed at assassinating Makarios and establishing enosis. They laid the presidential palace in ruins, but Makarios narrowly escaped, eventually flying via Malta to London. A former EOKA member, Nikos Sampson, was proclaimed president of Cyprus. Five days later Turkish forces landed at Kyrenia with the expressed aim of overturning Sampson's government. Vigorous resistance was offered, but after suffering substantial casualties the Turks were successful in establishing a bridgehead around Kyrenia and linking it with the Turkish sector of Nicosia. On July 23 Greece's junta fell and was replaced by a democratic government under Konstantinos Karamanlis. At the same time Sampson was replaced in Cyprus by Glafkos Clerides, who as president of the House of Representatives was the person stipulated by the constitution to carry out the head of state's duties in the latter's absence. The three guarantor powers, Britain, Greece, and Turkey, as required by the treaty, met for discussions in Geneva, but it proved impossible to halt the Turkish advance until August 16. By that time Turkey controlled the northern 37 percent of the island. In December Makarios returned to resume the presidency.

**Intercommunal talks.** Talks between Clerides and Rauf Denktash, representing the Greek and Turkish Cypriots, respectively, had begun in 1968. They continued inconclusively until 1974, the Turks demanding and the Greeks rejecting a bizonal federation with a weak central government. In February 1975 the Turkish Cypriots proclaimed the Turkish-occupied area the Turkish Federated State of Cyprus (a body calling itself the Provisional Cyprus-Turkish Administration had been in existence among Turkish Cypriots since 1967); Denktash announced that their purpose was not independence but federation. Talks were resumed in Vienna in April and June 1975 and February 1976 under UN auspices. In January and February 1977 Makarios met with Denktash; they agreed upon guidelines according to which Makarios could accept the principle of a bizonal federation.

Proclamation of the Turkish Federated State of Cyprus

In August 1977 Makarios died at the age of 64. Spyros Kyprianou, president of the House of Representatives, became acting president of the republic, and he was returned unopposed to that office for a five-year term in January 1978. An alliance between the President's Democratic Party and AKEL ensured Kyprianou's reelection as president in February 1983; Turkish Cypriots, however, took no part in the election.

**Declaration of an independent Turkish state.** In May 1983 Denktash broke off the intercommunal talks, and in November he proclaimed the Turkish Republic of Northern Cyprus. The UN Security Council condemned the move and repeated its demand, first made in 1974, for the withdrawal of all foreign troops from the Republic of Cyprus. (D.W.S.H.)

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

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 912, 962, and 972.

BIBLIOGRAPHY

*Physical and human geography:* (*General sources*): A comprehensive study is FREDERICA M. BUNGE (ed.), *Cyprus: A Country Study*, 3rd ed. (1980); and H.D. PURCELL, *Cyprus* (1969). For brief overviews, the following guidebooks are useful: COLIN THUBRON, *Journey into Cyprus* (1975); IAN ROBERTSON, *Cyprus* (1981); and SUZANNE PATTERSON, *Cyprus* (1985). (*Geology, flora, and fauna*): A. PANAYIOTOU (ed.), *Ophiolites* (1980), is a collection of symposium papers discussing the rock formations of Cyprus. Flora and fauna of the island are discussed in JENS HOLMBOE, *Studies on the Vegetation of Cyprus* (1914); B.F. OSORIO-TAFALL and GEORGE M. SERAPHIM, *List of the Vascular Plants of Cyprus* (1973); ESTHER F. CHAPMAN, *Cyprus Trees and Shrubs* (1949); and DAVID A. BANNERMAN and W. MARY BANNERMAN, *Birds of Cyprus* (1958).

(*Economics*): General information can be found in DIAMOND JENNESS, *The Economics of Cyprus: A Survey to 1914* (1962); A.J. MEYER and SIMOS VASSILIOU, *The Economy of Cyprus* (1962); and L.W. ST. JOHN-JONES, *The Population of Cyprus* (1983). Current developments are presented in REPUBLIC OF CYPRUS, DEPARTMENT OF STATISTICS AND RESEARCH, *Statistical Abstract* (annual), and *Economic Report* (annual).

(*Cultural life*): Art and literature are discussed in CAMILLE ENLART, *Gothic Art and the Renaissance in Cyprus* (1987; originally published in French, 1899); TONY SPITERIS, *The Art of Cyprus* (1971; originally published in French, 1970); VASSOS KARAGEORGHIS, *Ancient Cypriote Art in the Pierides Foundation Museum* (1985); ANDREAS STYLIANOU and JUDITH A. STYLIANOU, *The Painted Churches of Cyprus* (1985); and NICOS S. SPANOS (ed.), *Cypriot Prose-Writers, from Antiquity to 1950,* trans. from Greek by DAVID BAILEY *et al.* (1983).

*History:* For the earliest periods of the island's history, see VASSOS KARAGEORGHIS, *The Civilization of Prehistoric Cyprus,* trans. from the Greek (1976), *View from the Bronze Age: Mycenaean and Phoenician Discoveries at Kition* (1976; U.K. title, *Kition: Mycenaean and Phoenician Discoveries in Cyprus*), and *Cyprus: From the Stone Age to the Romans* (1982); and CLAUDE DELAVAL COBHAM, *Excerpta Cypria: Materials for a History of Cyprus* (1908, reprinted 1969). For current research in early history, see publications of the REPUBLIC OF CYPRUS, DEPARTMENT OF ANTIQUITIES. SIR GEORGE HILL, *A History of Cyprus,* 4 vol. (1940–52, reprinted 1972); and SIR DAVID HUNT (ed.), *Footprints in Cyprus: An Illustrated History* (1982), are both comprehensive surveys. Background of Cyprus' political history is provided in J.C. HUREWITZ, *The Middle East and North Africa in World Politics,* 2nd rev. ed., 2 vol. (1975).

Studies of the conflict of Turkish and Greek Cypriots include FRANÇOIS CROUZET, *Le Conflit de Chypre, 1946–1959*, 2 vol. (1973); NANCY CRAWSHAW, *The Cyprus Revolt: An Account of the Struggle for Union with Greece* (1978); ROBERT STEPHENS, *Cyprus, a Place of Arms: Power Politics and Ethnic Conflict in the Eastern Mediterranean* (1966); J.T.A. KOUMOULIDES (ed.), *Cyprus in Transition: 1960–1985* (1986); and JOSEPH S. JOSEPH, *Cyprus: Ethnic Conflict and International Concern* (1985). The effect of the conflict upon a Cypriot community is studied in PETER LOIZOS, *The Greek Gift: Politics in a Cypriot Village* (1975), and *The Heart Grown Bitter: A Chronicle of Cypriot War Refugees* (1981). Historical roots of the conflict are explored in STEPHEN G. XYDIS, *Cyprus: Conflict and Conciliation, 1954–1958* (1967), and *Cyprus: Reluctant Republic* (1973); and POLYVIOS G. POLYVIOU, *Cyprus: The Tragedy and the Challenge* (1975), and *Cyprus, Conflict and Negotiation, 1960–1980* (1980). For the Turkish-Cypriot point of view, see HALIL IBRAHIM SALIH, *Cyprus, the Impact of Diverse Nationalism on a State* (1978); PIERRE OBERLING, *The Road to Bellapais: The Turkish Cypriot Exodus to Northern Cyprus* (1982); ZAIM M. NEDJATIGIL, *The Cyprus Conflict: A Lawyer's View* (1981); and R.R. DENKTASH, *The Cyprus Triangle* (1982).

(D.W.S.H./Ed.)

# Czechoslovak Literature

The beginnings of vernacular writing in what is now Czechoslovakia are connected with the mission sent to Moravia in 863 by the Byzantine emperor Michael III at the request of Prince Rostislav. Christianity had reached the Slavs of Moravia from the west under the political aegis of the Frankish Empire. In order to counter Frankish influence Rostislav, prince of Great Moravia, sought "a bishop and teacher" from the Byzantine east. The mission was led by an experienced scholar and diplomat, Constantine (better known as Cyril), and his brother Methodius. The brothers translated the greater part of the Scriptures and the essential liturgical texts into a Slavonic literary language of Cyril's devising, based on the Slavonic vernacular of his native Salonika but enriched from other sources, notably Greek and the Slavonic of Moravia. The most notable literary monuments of this language (now known as Old Church Slavonic) are the *Lives* of the two brothers, which were almost certainly written before 900 though they are preserved only in later copies. It is not known for certain whether these texts were written in the territory where the brothers carried on their mission. Other Church Slavonic texts, however, can be assigned to the Czech era, notably the probably 10th-century *Legends* about St. Wenceslas (Václav, prince of Bohemia, 920–929) and his grandmother Ludmila. Similar hagiographical works in Latin from this period bear witness to a notable level of literary culture in the monastic centres of the early Bohemian state. The Old Church Slavonic language ceased to be used after 1097 when Latin was established as the liturgical language of the country.

This article is divided into the following sections:

Czech literature 902
Medieval period 902
Influence of the Hussite movement 902
Developments of the 16th and 17th centuries 902
The national revival 902
After 1918 903
Slovak literature 904
Attempts to establish a literary Slovak language 904
20th-century tendencies and concerns 904
Bibliography 904

## CZECH LITERATURE

arliest
reserved
xts

**Medieval period.** The earliest preserved texts in the Czech language were written in the latter part of the 13th century. The courts of the Přemyslid kings Wenceslas I (Václav I; died 1253) and Otakar II (died 1278) gave encouragement to courtly literature in German; but there is no certain evidence of a significant body of literature in Czech before the end of the century.

The 14th century saw a continuous stream of Czech literary production. A group of verse "Legends," written in the first quarter of the century, already showed considerable maturity of technique. The earliest secular work was the epic *Alexandreis,* a life of Alexander the Great, showing contact with a German Alexander epic by Ulrich von Eschenbach. The Czech poem is a minor masterpiece of courtly literature. Another striking epic, "The Dalimil Chronicle," portrayed the history of the Czechs in a vivid, forthright style. Under the emperor Charles IV learning and literature flourished in Bohemia in Czech and German. The greatest achievement of this period was a verse "Legend of St. Catherine."

From about 1350 prose genres began to be cultivated, first lives of saints and chronicles and then, no doubt for a growing urban middle class, versions of several popular medieval tales. From the last part of the century dated a group of verse satires and didactical poems showing great originality and character, especially the anonymous satires of the Hradec manuscript and a political allegory, *Nová rada* ("The New Council"), written by Smil Flaška of Pardubice to defend the rights of the Bohemian nobility against the power of the crown.

**Influence of the Hussite movement.** A preoccupation with social and moral questions coloured much Czech writing in the last part of the 14th century. Thomas of Štítný wrote treatises in the vernacular that anticipated one of the main tendencies of the Hussite movement. This reform movement, inspired by Jan Hus, dominated the life of Bohemia for the first half of the 15th century; amid the religious controversy and civil strife of this period it was natural for vernacular writing to be directed largely to practical and polemical purposes. Hus's own importance lay not only in his vernacular sermons and in his letters but also in the reform of the Czech orthography advocated in the treatise *De orthographia Bohemica.* The simple and moving hymns composed by the Hussites were a direct product of the reform movement, while the events of the Hussite wars were reflected in the poems of the Bautzen manuscript and in a number of chronicles and other prose works.

Among the successors of Jan Hus the most original was Petr Chelčický, who developed his radical social ideas and his pacifism in a series of striking treatises and homilies. It was from Chelčický's ideas that the Unitas Fratrum, or Unity of Brethren, drew their inspiration: this sect, the prototype of the Moravian Church, was to become one of the most important sources of vernacular Czech literature in the next two centuries.

Predominance of didactic and scholarly prose

**Developments of the 16th and 17th centuries.** In the 16th century there was a mass of prose literature in which didactic and scholarly writing predominated. The influence of Humanism was apparent in the elegant style of the legal historian Viktorin Kornel of Všehrdy. Of considerable interest is the small group of travel books from this period. The development of Czech prose culminated in the translation of the Bible made by a group of scholars of the Unitas Fratrum and known as the Kralice Bible (1579–93). The language of this version became the model of classical Czech.

The defeat of the Protestant Bohemian forces by the Habsburgs in 1620 at the Battle of the White Mountain brought the integration of the Czech provinces into the Habsburg empire and the eradication of Protestantism from the country. The nobility was replaced by newcomers with little knowledge of Czech. It was only among the émigrés that Czech literary traditions continued. Among these exiles Jan Ámos Komenský (John Amos Comenius) was preeminent. His Latin works on educational and theological problems and his works in Czech revealed him as a writer and thinker of European stature. His *Labyrint světa a ráj srdce* (1631; *Labyrinth of the World and Paradise of the Heart*) was one of the supreme achievements of Czech prose literature. Some of the new Jesuit literature was of high quality, notably the poetry of Bedřich Bridel and the Latin works of Bohuslav Balbín, whose Czech patriotism expressed itself in studies of his country's past and in a remarkable defense of the Czech language.

By the beginning of the 18th century, Czech had practically ceased to be used for higher literary purposes. Rough vernacular dramas and chapbooks were produced, together with popular poetry, which included not only folk songs but forceful expressions of peasant discontent.

**The national revival.** The historicism and antiquarianism of the 18th century led many Czech scholars to investigate the earlier literature and history of their country; a revival of Bohemian patriotism was, moreover, a natural reaction to the centralizing tendencies of the Habsburg government under Maria Theresa and was encouraged

The establishment of a modern Czech literary language

by the local nobility in the interests of their own power. Josef Dobrovský became a codifier of a revived Czech literary language, and in his history of older Czech literature he brought to light neglected models that Czech writers might follow. He himself wrote in Latin and German; but several of his contemporaries restored Czech to literary currency, notably Antonín Jaroslav Puchmajer with his poetical almanacs. At the same time, social and political developments of the late 18th century created a public for Czech literature. Vernacular plays performed in Prague from 1786 and the beginnings of modern Czech journalism represented by the newspapers of Václav Matěj Kramerius were directed at a new middle class of growing Czech towns.

Dobrovský had contented himself with codifying the Czech language: Josef Jungmann set out to extend its possibilities so that it could become a modern literary language and achieved this in translations (including that of Milton's *Paradise Lost,* 1811) and above all in a monumental Czech–German dictionary (1835–39). The revival of Czech culture was also greatly furthered by František Palacký, whose history of Bohemia and Moravia was the last masterpiece of classical Czech.

Some Slovaks were being drawn into the Czech literary revival; foremost was Ján Kollár, whose allegorical sonnet cycle *Slávy dcera* ("The Daughter of Slava") was the first significant work in the revived language.

A special position in this period belonged to the *Rukopis Královédvorský* ("Manuscript of Králové Dvůr") and *Rukopis Zelenohorský* ("Manuscript of Zelená Hora"), ostensibly poems from the early Middle Ages but in fact forgeries executed by a brilliant and unscrupulous poet, Václav Hanka, and by others. Though their spurious character was finally proved toward the end of the century, they have held their own as examples of Romantic poetry, with echoes of popular poetry of southern Slavs.

Czech Romanticism

The greatest poet of Czech Romanticism, perhaps of all Czech poets, was Karel Hynek Mácha. His lyrics, prose fragments, and a lyrical epic, *Máj* (1836; *May*), showed the influence of Byron, Sir Walter Scott, and the Polish Romantics, but it was an influence transcended by its intensity of poetic vision and perfection of language.

In the 1840s a reaction against Romanticism became evident in a number of authors, notably Karel Havlíček Borovský and Božena Němcová. Both gave attention to practical issues and were masters of prose style who did much to bring Czech prose nearer to the everyday language. Havlíček was a political journalist whose achievement lay in critical articles in which he stood out for Czech rights against Habsburg absolutism, and in brilliant satirical poems, especially those written in exile just before his death: *Král Lávra* ("King Lavra"), *Tyrolské elegie* ("Tyrolese Elegies"), and *Křest svatého Vladimíra* ("The Baptism of St. Vladimir"). Božena Němcová, the first woman to make a significant contribution to Czech literature, became best known for *Babička* ("The Grandmother"), a portrayal of country life in a series of sketches held together by the central figure. Descriptions are precise and realistic; and the language has a firmness and individuality new in Czech prose.

The work of Havlíček, Němcová, and the popular poet Karel Jaromir Erben, all at their peak in the 1850s, formed a bridge from a first, Romantic phase of revived Czech literature to the second half of the century, in which authors, writing for an expanding public, tried to establish a literature that might occupy a worthy place among other literatures of Europe. In 1858 an almanac, *Máj,* appeared, dedicated to the memory of Mácha and containing the work of young authors, among whom Jan Neruda and Vitězslav Hálek were preeminent. A significant novelist of the *Máj* group was Karolina Světlá (Johanna Mužáková); her novels of early 19th-century Prague society and of the north Bohemian countryside contained Romantic elements of theme and plot but showed considerable skill in psychological analysis. Moral and social problems and, in particular, the position of women in 19th-century society were dominant themes of her work. It was under the auspices of the *Máj* group that Adolf Heyduk began his literary career. His patriotic poetry best stood the test of time, notably the poems in which he was inspired by the kinship between Czechs and Slovaks.

By the 1870s Czech literature had fully established itself in poetry and the novel but not yet in the drama. Literary life crystallized around two groups associated with the periodicals *Lumír* and *Ruch* ("Stir"). The former stressed the necessity to Europeanize Czech literature, the latter looked to the strength of native traditions and themes. The leading representative of the cosmopolitan tendency was Jaroslav Vrchlický (Emil Frída), probably the most prolific of all Czech writers. His lyrics showed an amazing mastery of language, with a vast cycle of historical epics containing his best work. Perhaps his greatest influence was exercised by numerous translations of major European writers, and he was also instrumental in forming a modern Czech poetical language. Cosmopolitanism found expression, too, in the poetry of Julius Zeyer's Neoromantic lyrics and epics, though his novels and stories, particularly the semi-autobiographical *Jan Maria Plojhar* (1888), were more original.

The *Lumír* and *Ruch* groups

The principal figure of the "Stir" group was Svatopluk Čech, whose work was inspired by love of his nation and its traditions and by a liberal humanism. One of his most popular works, *Ve stínu lípy* ("In the Shade of the Lime Tree"), gave an idyllic picture of Czech country life; in his prose satires he created Mr. Brouček, who satirized the Czech middle classes' philistinism.

The Czech novel of the last quarter of the century exhibited a trend toward realism of description, which was apparent in two major historical novelists, Alois Jirásek and Zikmund Winter (whose novels, such as *Mistr Kampanus,* were by-products of his work as a historian). Both men presented a romanticized version of Czech history, but the details of reconstruction were based, in both instances, on scholarly research. The novels of Jirásek presented an entire history of the Czechs up to his own time, and his mature craftsmanship was seen at its best in a novel of the period of national decline, *Temno* ("Darkness").

In the last decade of the century new trends shaped poetry. Lyrics of more subtle content and more complex form were written by Antonín Sova, and the turn of the century saw the best work of two remarkable Czech poets: Otakar Březina (Václav Jebavý) and Petr Bezruč (Vladimír Vašek). Březina expressed a personal religion in a language of subtlety and originality; his metrical structure (including free rhythms) influenced later poets. Bezruč was a regional poet, obsessed by the oppression of the Czechs of Silesia.

In the 1890s, critical reaction to the romantic idealism of the national revival had inspired the historical and sociological writings of T.G. Masaryk. A more sophisticated approach to literary criticism showed itself in the work of František Xaver Šalda, whose sensitive interpretations of his country's greatest writers have not been surpassed.

**After 1918.** With the establishment of an independent Czechoslovak state in 1918, new possibilities lay before Czech writers. Some of the best achievements of Czech literature sprang from this period, especially in lyrical poetry of the utmost variety. The Czech drama came into its own, notably, in the idealistic and satirical plays of Karel Čapek and in the versatile talent of František Langer. Shortly after World War I Jaroslav Hašek began his sequence of novels called *Osudy dobrého vojáka Švejka za světové války* (1921–23; *The Good Soldier Švejk*), widely praised satirical writings set in the collapsing Austro-Hungarian Empire. The realistic novel of Prague life had been successfully cultivated by Karel Matěj Čapek-Chod and in the lighter vein by Ignát Herrmann, but Czech narrative prose reached a peak in the work of three very different writers: Karel Čapek, Ivan Olbracht (Kamil Zeman), and Vladislav Vančura. Čapek wrote Wellsian fantasies, subtle psychological studies, and stories, such as a trilogy *Hordubal* (1933; Eng. trans. 1934), *Povětroň* (1934; *The Meteor*), and *Obyčejný život* (1934; *An Ordinary Life*); *Krakatit* (1924; Eng. trans. 1925), and *Válka s mloky* (1936; *War with the Newts*). His best plays exposed the problems of a centralized, machine society of the 20th century: *R.U.R.* (1920), from which the word robot was introduced into English, and *Ze života hmyzu* (1921; *The*

The satire of Karel Čapek and Jaroslav Hašek

*Insect Play*). In his later plays, pacifism gave way to an acceptance of patriotic duty. Olbracht's most successful works had a Ruthenian background. Vančura cultivated a complex, mannered style with success. The novels of Jan Otčenášek stand out as the most effective from a mass of fiction dealing with World War II. (Ro.Au.)

During the war years the psychological novel reached its peak in the work of Václav Řezáč. After the establishment of a Communist government in 1948, he and others adopted Socialist Realism, with inferior results. When strict political controls were relaxed in the early 1960s, the novels of Josef Škvorecký, Milan Kundera, and Ludvík Vaculík, the short stories of Bohumil Hrabal and Arnošt Lustig, the poetry of Miroslav Holub, and the plays of Ivan Klíma and Václav Havel were among the literature that soon attracted worldwide attention. Other writers, such as the dynamic innovator Vladimír Páral, enjoyed both critical success and great popularity at home. In the wake of the Soviet-led invasion in 1968, many of these writers were suppressed in Czechoslovakia, although abroad their reputations continued to grow. Of the large number of writers who emigrated to the western nations, Kundera and Škvorecký soon attained prominence in their adopted countries without forfeiting their national origin in the choice of literary subject and treatment. In quality their work surpassed most literature published in Czechoslovakia in the 1970s. During the same period, belated interest was shown abroad in one of the highest achievements of Czech 20th-century literature: the lyrical poetry of the interwar years written by Josef Hora, František Halas, Vítězslav Nezval, and Jaroslav Seifert. Seifert survived well into the 1980s and saw much of his later poetry translated into English and other languages. He was awarded the Nobel Prize for Literature for 1984. (I.H.)

## SLOVAK LITERATURE

**Attempts to establish a literary Slovak language.** Until the 18th century there was no systematic attempt to establish a literary language from the Slovak dialects, which, though closely related to Czech, developed a separate identity from the early Middle Ages. The decline of literary Czech in the early 18th century caused an increase of local colouring in devotional texts produced by both Catholics and Calvinists. With Anton Bernolák these tendencies were consolidated in a grammar (1790) and a dictionary (6 vol., 1825–27), which codified a literary language based mainly on western Slovak usage. In an era of reviving national consciousness, this language was taken up by a number of writers. Preeminent was Ján Hollý, who wrote lyrics, idylls, and national epics. The didactic novel *René* (1783–85) was written in a strongly Slovakized Czech by Jozef Ignác Bajza.

Bernolák's language was not accepted by the Protestants of Slovakia, who used Czech as their liturgical language. By the 1840s, however, Slovak national feeling among the Catholic majority was so strong that there was little possibility of their accepting Czech as their written language. The western Slovak of Bernolák was inadequate; but L'udovít Štúr codified a new form of literary Slovak based on the central dialects that was immediately taken up by a group of talented poets. Andrej Sládkovič (Andrej Braxatoris) produced, in *Marína* (1846), a national epic that had individuality and freshness. The most significant figure was Janko Kráľ, whose exploits in the revolution and war of 1848–49 made him a legend. His ballads, epics, and lyrics are among the most original products of Slavonic Romanticism.

The first Slovak dramatist of note was Ján Chalupka, whose works, including a lively satire, *Kocúrkovo* (1830), were written in Czech. Ján Palárik used the new literary Slovak in his comedies, produced in the 1850s and 1860s.

**20th-century tendencies and concerns.** The chief strength of Slovak literature continued to be in the lyric, and the preeminent prewar poet was Hviezdoslav (Pavol Országh). He enriched Slovak poetic language by original work and by translations, especially from English, Russian, German, and Hungarian. Other notable poets were the Russophile Svetozár Hurban Vajanský and Ivan Krasko (the pseudonym of Ján Botto), whose volumes of verse, *Nox et solitudo* (1909) and *Verše* (1912), were among the finest achievements of Slovak literature.

Supremacy of lyric poetry

After World War I leading lyric poets were Martin Rázus, Janko Jesenský, Emil Boleslav Lukáč, Ján Smrek (Ján Čietek), Ján Poničan, and Laco Novomeský. In the novel, country tales by Timrava (Božena Slančíkova), a vast chronicle of 20th-century Slovakia by Milo Urban, and the lyrical prose of Margita Figuli were outstanding. The problems of World War II and its aftermath of Communist government found vivid, personal expression in the work of Ladislav Mňačko and Alfonz Bednár. (Ro.Au.)

The more liberal political climate of the 1960s brought fresh independence and self-confidence to Slovak literature, but in the 1970s little Slovak writing of significance emerged either in Czechoslovakia or among writers in exile. (I.H.)

### BIBLIOGRAPHY

*Czech:* Czech literature before World War I is covered in FRANTIŠEK CHUDOBA, *A Short Survey of Czech Literature* (1924, reprinted 1969); and EWALD OSERS and J.K. MONTGOMERY (eds. and trans.), *Modern Czech Poetry: An Anthology* (1945). ARNE NOVÁK, *Czech Literature* (1976; trans. from the Czech), although considerably dated in both style and conception, provides an essential overview, which can be usefully supplemented by the anthologies *Czech Prose,* trans. and ed. by WILLIAM E. HARKINS (1983), and *Anthology of Czech Poetry,* ed. by ALFRED FRENCH (1973). Poetry written between World Wars I and II is examined in ALFRED FRENCH, *The Poets of Prague* (1969); and MILADA SOUČKOVÁ, *A Literary Satellite: Czechoslovak–Russian Literary Relations* (1970), deals with postwar developments. Also of interest are PAUL I. TRENSKY, *Czech Drama Since World War II* (1978); and MARKETA GOETZ-STANKIEWICZ, *The Silenced Theatre: Czech Playwrights Without a Stage* (1979). ALFRED FRENCH, *Czech Writers and Politics, 1945–1969* (1982), supplies an explanatory background to the symposium *Czech Literature Since 1956,* ed. by WILLIAM E. HARKINS and PAUL I. TRENSKY (1980).

*Slovak:* Critical sources in Slovak include ANDREJ MRÁZ, *Dejiny slovenskej literatúry* (1948); JAROSLAV VLČEK, *Dejiny literatúry slovenskej,* 5th ed., 3 vol., ed. by OLGA SVEJKOVSKÁ and FRANTIŠEK SVEJKOVSKÝ (1960); and MILAN PIŠÚT *et al., Dejiny slovenskej literatúry,* 3rd rev. ed. (1984). English-language sources include DIONÝZ ĎURIŠIN, *Sources and Systematics of Comparative Literature,* trans. by PETER TKÁČ (1974); JOSEPH M. KIRSCHBAUM, *Pan-Slavism in Slovak Literature* (1966), and *Slovak Language and Literature* (1975); PETER BROCK, *The Slovak National Awakening* (1976); and ANDREW CINCURA (comp.), *An Anthology of Slovak Literature* (1976).

Anthologies that treat Czech and Slovak literatures together are PAUL SELVER, *Czechoslovak Literature* (1942); SELVER (comp. and trans.), *An Anthology of Czechoslovak Literature* (1929, reprinted 1969); and SELVER (ed. and trans.), *A Century of Czech and Slovak Poetry* (1946); JEANNE W. NĚMCOVÁ (ed. and trans.), *Czech and Slovak Short Stories* (1967); *New Writing in Czechoslovakia,* comp. by GEORGE THEINER (1969); and *The Writing on the Wall,* ed. by PETER KUSSI and ANTONIN LIEHM (1983). GEORGE J. KOVTUN, *Czech and Slovak Literature in English* (1984), is a bibliography of translations.

(Ro.Au./I.H.)

# Czechoslovakia

The position of Czechoslovakia in the heart of Europe has been of great significance, affecting the economic, political, and cultural development of a country in which the most varied of influences and traditions have encountered one another. The elongated shape of the country—its east–west extent is more than 465 miles (748 kilometres), whereas its maximum north–south extent is barely 171 miles—has further differentiated its internal physical and human geography, western European influences being strong in Bohemia while those of eastern Europe predominate in eastern Slovakia. The nation's position on the boundary between Soviet, reformed Socialist, and capitalist systems has added a new element to this situation.

An inland country, Czechoslovakia (Czech and Slovak Federal Republic; Czech: Česká a Slovenská Federativní Republika; Slovak: Česká a Slovenská Federatívna Republika) lies across the great ancient trade routes of Europe, following the Elbe (Czech, Labe) River to the North Sea, the Oder (Odra) River to the Baltic, and the Danube (Dunaj) River to the Black Sea. There were also routes through the Carpathians (Czech, Karpaty). Modern Czechoslovakia came into being at the end of World War I, following the collapse of the Austro-Hungarian Empire. Its area of about 49,380 square miles (127,900 square kilometres) is subdivided into the Czech Republic (Česka Republika), of about 30,450 square miles (78,860 square kilometres), comprising the historic lands of Bohemia and Moravia (often called the Czech Lands) and the Austrian portion of Silesia; and the Slovak Republic (Slovenská Republika) of about 18,930 square miles (49,040 square kilometres), which before World War I was a part of Hungary inhabited mainly by Slovaks.

Following the overthrow of the Communist government in 1989, Czechoslovakia's continued membership in the Council for Mutual Economic Assistance (Comecon) and also in the Warsaw Treaty Organization became uncertain. The country has frontiers with Poland in the north, the Soviet Union in the east, Hungary in the southeast, Austria in the southwest, and Germany in the west and north. The capital is Prague.

This article is divided into the following sections:


Physical and human geography 905
- The land 905
  - Relief
  - Drainage
  - Soils
  - Climate
  - Plant and animal life
  - Settlement patterns
- The people 910
  - Ethnic composition
  - Linguistic composition
  - Religions
  - Demographic trends
- The economy 910
  - Resources
  - Agriculture and forestry
  - Industry
  - Finance
  - Trade
  - Transportation
- Government and social conditions 912
  - Government
  - Education
  - Health and welfare
- Cultural life 912
  - Literature
  - Music
  - Film
  - Fine, applied, and folk arts
  - Libraries and museums
  - Press and broadcasting

History 914
- Origins and early history 914
  - Bohemia
  - Unification of Great Moravia
  - Slovakia
- The Přemysl rulers of Bohemia (895–1306) 915
- Late Middle Ages (1310–1526) 916
  - The Luxembourg dynasty (1310–1437)
  - The Hussite preponderance (1437–71)
  - The Jagiellonian kings (1471–1526)
- Habsburg rule (1526–1914) 919
  - The Habsburgs to 1848
  - From absolutism to constitutionalism (1848–1914)
- The Republic of Czechoslovakia (1918–45) 923
  - Struggle for independence (1914–18)
  - Establishment of Czechoslovakia (1918–25)
  - Political consolidation (1925–35)
  - Moving toward the abyss (1935–38)
  - From Munich to the disruption of the republic (1938–39)
  - Struggles at home and abroad (1939–45)
- Czechoslovakia since 1945 926
  - Provisional regime (1945–48)
  - People's democracy (1948–60)
  - Reform movement (1960–68)
  - Prague Spring (1968)
  - Husák's policies
  - The collapse of Communism

Bibliography 929


## Physical and human geography

### THE LAND

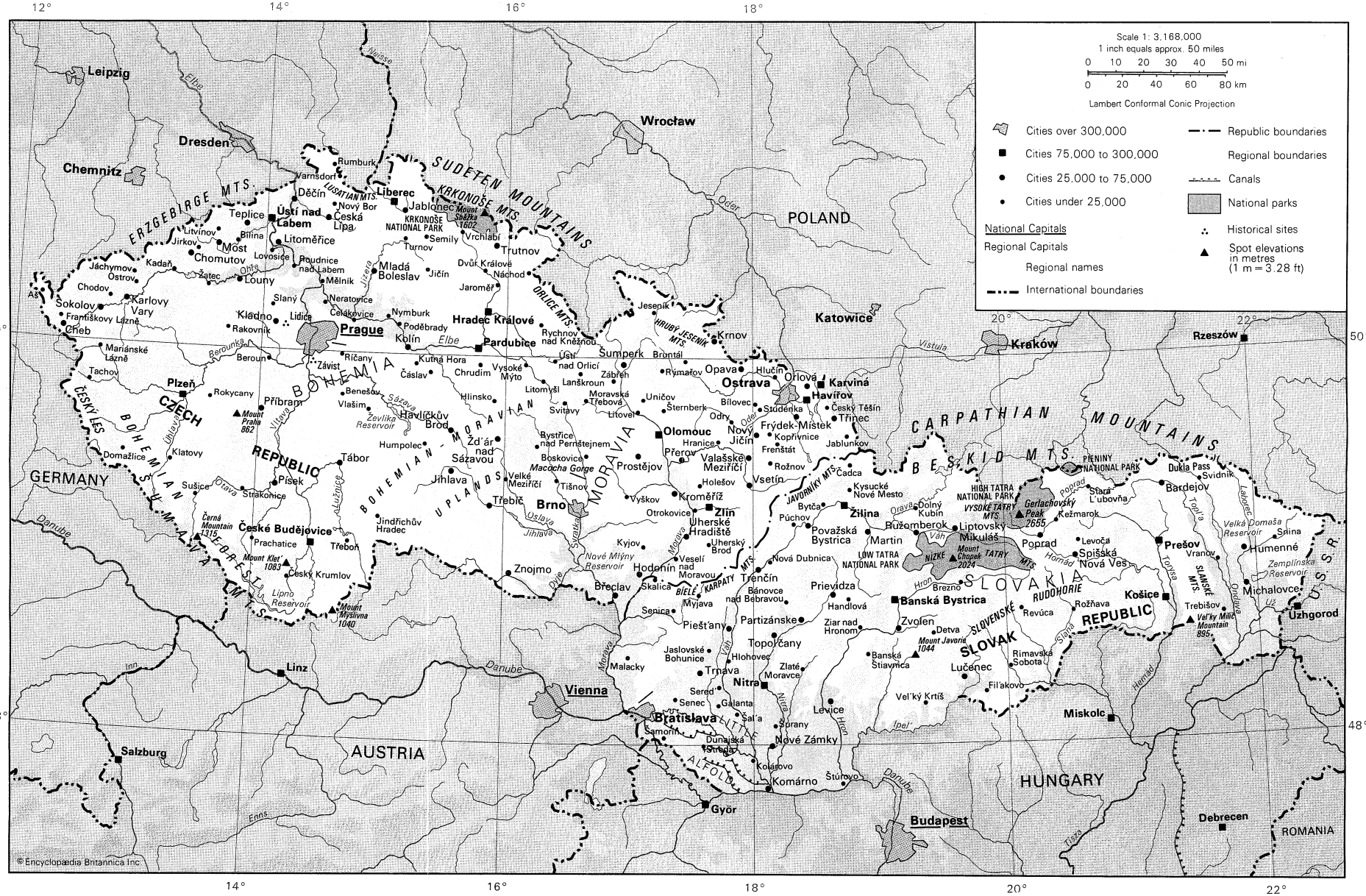
**Relief.** Geographers divide Czechoslovakia into two large provinces, the Bohemian Massif (Český masív) in the west and the Carpathian Mountains in the east. Each is subdivided into several smaller regions.

Bohemian Massif

The Bohemian Massif consists of a large elevated basin (the Bohemian Plateau, or Česká tabule) encircled by mountain ranges that occasionally reach heights exceeding 3,000 feet (900 metres) above sea level. The massif may be divided into seven major units. The first is the centrally located Bohemian Plateau, also called the Labe Plain or Polabí, after the Elbe River, which traverses it from east to west. The Elbe receives the Vltava River on its left bank, about 18 miles (30 kilometres) below Prague.

Southwest of the Bohemian Plateau is the largest unit, the Southern Bohemian Highlands (Jihočeská vysočina), which comprise the older, crystalline core of the massif. The most important parts of the Southern Bohemian Highlands are the Bohemian–Moravian Uplands, the Central Bohemian Highlands (Středočeská pahorkatina), the Southern Bohemian Basin (Jihočeské pánve; centred on the lake district to the east of the city of České Budějovice), and, in the southwest, the Šumava Subsystem (Podsoustava Šumavy; comprising the Český Les as well as Šumava portions of the Bohemian Forest).

West of the Bohemian Plateau are the Berounka River Highlands (Poberounská vrchovina). The city of Plzeň lies at the centre of this region. To the northwest of the Bohemian Plateau are the Erzgebirge (Krušné hory; Ore Mountains), which form the frontier with Germany, comprising two elongated subsystems paralleling the border. The lowest point in the Bohemian Massif (384 feet [117 metres]) is found there, where the Elbe breaches these ranges as it enters Germany.


Scale 1: 3,168,000
1 inch equals approx. 50 miles
0 10 20 30 40 50 mi
0 20 40 60 80 km
Lambert Conformal Conic Projection
Cities over 300,000
Cities 75,000 to 300,000
Cities 25,000 to 75,000
Cities under 25,000
National Capitals
Regional Capitals
Regional names
International boundaries
Republic boundaries
Regional boundaries
Canals
National parks
Historical sites
Spot elevations in metres (1 m = 3.28 ft)
CZECH REPUBLIC
SLOVAK REPUBLIC
BOHEMIA
MORAVIA
SILESIA
SUDETEN MOUNTAINS
ERZGEBIRGE MTS.
KRKONOŠE MTS.
ORLICE MTS.
HRUBÝ JESENÍK MTS.
JAVORNÍKY MTS.
BESKID MTS.
CARPATHIAN MOUNTAINS
BOHEMIAN-MORAVIAN UPLANDS
ČESKÝ LES
BOHEMIAN FOREST
ŠUMAVA MTS.
KARPATY BIELE
LOW TATRA NATIONAL PARK
NÍZKE TATRY MTS.
HIGH TATRA NATIONAL PARK
VYSOKÉ TATRY MTS.
PIENINY NATIONAL PARK
SLOVENSKÉ RUDOHORIE
SLANSKÉ MTS.
ZEMPLÍNSKA ŠIRAVA
LITTLE ALFOLD
Prague
Brno
Ostrava
Plzeň
Hradec Králové
Pardubice
Olomouc
Zlín
Žilina
Banská Bystrica
Košice
Prešov
Nitra
Trnava
Trenčín
Martin
Poprad
Bratislava
Vienna
Linz
Salzburg
Budapest
Győr
Miskolc
Debrecen
Uzhgorod
Michalovce
Humenné
Wrocław
Katowice
Kraków
Rzeszów
Leipzig
Dresden
Chemnitz
POLAND
AUSTRIA
HUNGARY
ROMANIA
GERMANY
Elbe
Danube
Vistula
Oder
Morava
Váh
Hron
Ipeľ
Nitra
Hornád
Torysa
Ondava
Laborec
Poprad
Orava
Slaná
Ostrá Hole 1572
Mount Snežka 1602
Mount Praha 862
Mount Klet' 1083
Černá Mountain 1315
Mount Mysliva 1040
Gerlachovský Peak 2655
Mount Chopok 2024
Mount Javorie 1044
Veľký Milič Mountain 895
Veľký Krtíš
Fiľakovo
Lučenec
Rimavská Sobota
Banská Štiavnica
Zvolen
Detva
Partizánske
Prievidza
Bánovce nad Bebravou
Považská Bystrica
Púchov
Bytča
Kysucké Nové Mesto
Dolný Kubín
Ružomberok
Liptovský Mikuláš
Kežmarok
Stará Ľubovňa
Bardejov
Svidník
Dukla Pass
Vranov
Spišská Nová Ves
Levoča
Rožňava
Revúca
Trebišov
Veľká Domaša Reservoir
Zemplínska Reservoir
Orlová
Havířov
Karviná
Frýdek-Místek
Nový Jičín
Kopřivnice
Frenštát
Jablunkov
Třinec
Český Těšín
Bílovec
Studénka
Odry
Hlučín
Opava
Rýmařov
Bruntál
Šumperk
Jeseník
Krnov
Zábřeh
Lanškroun
Litomyšl
Svitavy
Litovel
Moravská Třebová
Uničov
Šternberk
Olomouc
Přerov
Prostějov
Hranice
Valašské Meziříčí
Rožnov
Vsetín
Čadca
Kysucké Nové Mesto
Holešov
Kroměříž
Otrokovice
Uherské Hradiště
Uherský Brod
Veselí nad Moravou
Nová Dubnica
Skalica
Myjava
Senica
Piešťany
Topoľčany
Jaslovské Bohunice
Hlohovec
Malacky
Trnava
Nitra
Zlaté Moravce
Šaľa
Šurany
Nové Zámky
Komárno
Štúrovo
Levice
Senec
Galanta
Dunajská Streda
Šamorín
Kolárovo
Handlová
Ziar nad Hronom
Banská Bystrica
Brezno
Senica
Myjava
Skalica
Hodonín
Uherské Hradiště
Kyjov
Znojmo
Břeclav
Třebíč
Jihlava
Velké Meziříčí
Tišnov
Boskovice
Macocha Gorge
Blansko
Svitavy
Žďár nad Sázavou
Bystřice nad Pernštejnem
Humpolec
Jihlava
Havlíčkův Brod
Vlašim
Benešov
Příbram
Rokycany
Klatovy
Domažlice
Tachov
Cheb
Sokolov
Karlovy Vary
Františkovy Lázně
Mariánské Lázně
Aš
Chodov
Jáchymov
Ostrov
Kadaň
Žatec
Louny
Most
Litvínov
Jirkov
Chomutov
Teplice
Bílina
Ústí nad Labem
Děčín
Varnsdorf
Rumburk
Česká Lípa
Litoměřice
Lovosice
Roudnice nad Labem
Mělník
Slaný
Kladno
Rakovník
Lidice
Neratovice
Čelákovice
Nymburk
Poděbrady
Kolín
Prague
Beroun
Kutná Hora
Čáslav
Chrudim
Vysoké Mýto
Ústí nad Orlicí
Rychnov nad Kněžnou
Náchod
Trutnov
Jaroměř
Dvůr Králové
Jičín
Mladá Boleslav
Turnov
Semily
Vrchlabí
Jablonec
Liberec
Nový Bor
Česká Lípa
Lusatian Mts.
Jizera
Ohře
Berounka
Sázava
Vltava
Elbe
Orlice
Dyje
Svratka
Jihlava
Úslava
Otava
Lužnice
Lipno Reservoir
Orlík Reservoir
Želivka Reservoir
Slapy Reservoir
Nové Mlýny Reservoir
Macocha Gorge
Inn
Enns
Danube
Tisza
Ipeľ
Hron
Nitra
Morava
Váh
Hornád
Torysa
Ondava
Laborec
Uh
Poprad
Orava
Slaná
Ostrá Hole 1572
Mount Snežka 1602
Mount Praha 862
Mount Klet' 1083
Černá Mountain 1315
Mount Mysliva 1040
Gerlachovský Peak 2655
Mount Chopok 2024
Mount Javorie 1044
Veľký Milič Mountain 895
Veľký Krtíš
Fiľakovo
Lučenec
Rimavská Sobota
Banská Štiavnica
Zvolen
Detva
Partizánske
Prievidza
Bánovce nad Bebravou
Považská Bystrica
Púchov
Bytča
Kysucké Nové Mesto
Dolný Kubín
Ružomberok
Liptovský Mikuláš
Kežmarok
Stará Ľubovňa
Bardejov
Svidník
Dukla Pass
Vranov
Spišská Nová Ves
Levoča
Rožňava
Revúca
Trebišov
Veľká Domaša Reservoir
Zemplínska Reservoir
Orlová
Havířov
Karviná
Frýdek-Místek
Nový Jičín
Kopřivnice
Frenštát
Jablunkov
Třinec
Český Těšín
Bílovec
Studénka
Odry
Hlučín
Opava
Rýmařov
Bruntál
Šumperk
Jeseník
Krnov
Zábřeh
Lanškroun
Litomyšl
Svitavy
Litovel
Moravská Třebová
Uničov
Šternberk
Olomouc
Přerov
Prostějov
Hranice
Valašské Meziříčí
Rožnov
Vsetín
Čadca
Kysucké Nové Mesto
Holešov
Kroměříž
Otrokovice
Uherské Hradiště
Uherský Brod
Veselí nad Moravou
Nová Dubnica
Skalica
Myjava
Senica
Piešťany
Topoľčany
Jaslovské Bohunice
Hlohovec
Malacky
Trnava
Nitra
Zlaté Moravce
Šaľa
Šurany
Nové Zámky
Komárno
Štúrovo
Levice
Senec
Galanta
Dunajská Streda
Šamorín
Kolárovo
Handlová
Ziar nad Hronom
Banská Bystrica
Brezno
Senica
Myjava
Skalica
Hodonín
Uherské Hradiště
Kyjov
Znojmo
Břeclav
Třebíč
Jihlava
Velké Meziříčí
Tišnov
Boskovice
Macocha Gorge
Blansko
Svitavy
Žďár nad Sázavou
Bystřice nad Pernštejnem
Humpolec
Jihlava
Havlíčkův Brod
Vlašim
Benešov
Příbram
Rokycany
Klatovy
Domažlice
Tachov
Cheb
Sokolov
Karlovy Vary
Františkovy Lázně
Mariánské Lázně
Aš
Chodov
Jáchymov
Ostrov
Kadaň
Žatec
Louny
Most
Litvínov
Jirkov
Chomutov
Teplice
Bílina
Ústí nad Labem
Děčín
Varnsdorf
Rumburk
Česká Lípa
Litoměřice
Lovosice
Roudnice nad Labem
Mělník
Slaný
Kladno
Rakovník
Lidice
Neratovice
Čelákovice
Nymburk
Poděbrady
Kolín
Prague
Beroun
Kutná Hora
Čáslav
Chrudim
Vysoké Mýto
Ústí nad Orlicí
Rychnov nad Kněžnou
Náchod
Trutnov
Jaroměř
Dvůr Králové
Jičín
Mladá Boleslav
Turnov
Semily
Vrchlabí
Jablonec
Liberec
Nový Bor
Česká Lípa
Lusatian Mts.
Jizera
Ohře
Berounka
Sázava
Vltava
Elbe
Orlice
Dyje
Svratka
Jihlava
Úslava
Otava
Lužnice
Lipno Reservoir
Orlík Reservoir
Želivka Reservoir
Slapy Reservoir
Nové Mlýny Reservoir
Macocha Gorge
Inn
Enns
Danube
Tisza
Ipeľ
Hron
Nitra
Morava
Váh
Hornád
Torysa
Ondava
Laborec
Uh
Poprad
Orava
Slaná
Ostrá Hole 1572
Mount Snežka 1602
Mount Praha 862
Mount Klet' 1083
Černá Mountain 1315
Mount Mysliva 1040
Gerlachovský Peak 2655
Mount Chopok 2024
Mount Javorie 1044
Veľký Milič Mountain 895
Veľký Krtíš
Fiľakovo
Lučenec
Rimavská Sobota
Banská Štiavnica
Zvolen
Detva
Partizánske
Prievidza
Bánovce nad Bebravou
Považská Bystrica
Púchov
Bytča
Kysucké Nové Mesto
Dolný Kubín
Ružomberok
Liptovský Mikuláš
Kežmarok
Stará Ľubovňa
Bardejov
Svidník
Dukla Pass
Vranov
Spišská Nová Ves
Levoča
Rožňava
Revúca
Trebišov
Veľká Domaša Reservoir
Zemplínska Reservoir
12°
14°
16°
18°
20°
22°
50°
48°
© Encyclopædia Britannica Inc.

**MAP INDEX**

**Political subdivisions**

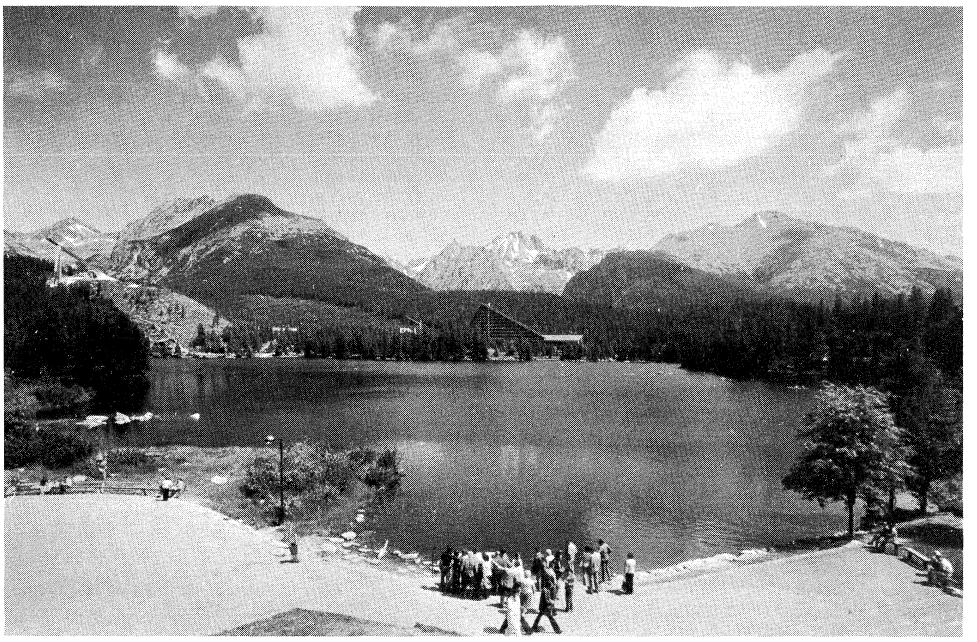

Bratislava ..... 48 09 N 17 07 E
Czech Republic .. 49 45 N 15 00 E
Jihočeský ..... 49 05 N 14 40 E
Jihomoravský ... 49 15 N 16 45 E
Prague ..... 50 05 N 14 28 E
Severočeský .... 50 35 N 14 15 E
Severomoravský . 49 45 N 17 40 E
Slovak Republic . 48 40 N 19 30 E
Středočeský .... 50 00 N 14 25 E
Středoslovenský . 48 40 N 19 10 E
Východočeský ... 50 10 N 16 00 E
Východoslovenský ..... 48 55 N 21 10 E
Západočeský ... 49 40 N 13 00 E
Západoslovenský . 48 20 N 18 00 E


**Cities and towns**


Aš ..... 50 13 N 12 11 E
Bánovce nad Bebravou ..... 48 43 N 18 16 E
Banská Bystrica . 48 44 N 19 09 E
Banská Štiavnica . 48 27 N 18 54 E
Bardejov ..... 49 17 N 21 17 E
Benešov ..... 49 47 N 14 41 E
Beroun ..... 49 57 N 14 05 E
Bílina ..... 50 32 N 13 48 E
Bílovec ..... 49 46 N 18 02 E
Boskovice ..... 49 29 N 16 40 E
Bratislava ..... 48 09 N 17 07 E
Břeclav ..... 48 46 N 16 53 E
Brezno ..... 48 49 N 19 39 E
Brno ..... 49 12 N 16 38 E
Bruntál ..... 49 59 N 17 28 E
Bystřice nad Pernštejnem ... 49 31 N 16 16 E
Bytča ..... 49 13 N 18 34 E
Čadca ..... 49 26 N 18 47 E
Čáslav ..... 49 55 N 15 24 E
Čelákovice ..... 50 10 N 14 46 E
Česká Lípa ..... 50 41 N 14 33 E
České Budějovice .... 48 59 N 14 28 E
Český Krumlov .. 48 49 N 14 19 E
Český Těšín .... 49 45 N 18 37 E
Cheb ..... 50 04 N 12 22 E
Chodov ..... 50 15 N 12 45 E
Chomutov ..... 50 27 N 13 26 E
Chrudim ..... 49 57 N 15 48 E
Děčín ..... 50 47 N 14 13 E
Detva ..... 48 34 N 19 25 E
Dolný Kubín .... 49 12 N 19 18 E
Domažlice ..... 49 26 N 12 56 E
Dunajská Streda . 47 59 N 17 37 E
Dvůr Králové .... 50 26 N 15 49 E
Fil'akovo ..... 48 16 N 19 50 E
Františkovy Lázně ..... 50 07 N 12 22 E
Frenštát ..... 49 33 N 18 13 E
Frýdek-Místek ... 49 41 N 18 21 E
Galanta ..... 48 12 N 17 43 E
Handlová ..... 48 44 N 18 46 E
Havířov ..... 49 47 N 18 22 E
Havlíčkův Brod .. 49 37 N 15 35 E
Hlinsko ..... 49 46 N 15 54 E
Hlohovec ..... 48 26 N 17 48 E
Hlučín ..... 49 54 N 18 11 E
Hodonín ..... 48 52 N 17 08 E
Holešov ..... 49 20 N 17 35 E
Hradec Králové .. 50 13 N 15 50 E
Hranice ..... 49 33 N 17 44 E
Humenné ..... 48 56 N 21 55 E
Humpolec ..... 49 33 N 15 22 E
Jablonec ..... 50 43 N 15 11 E
Jablunkov ..... 49 35 N 18 46 E
Jáchymov ..... 50 23 N 12 56 E
Jaroměř ..... 50 22 N 15 55 E
Jaslovské Bohunice ..... 48 28 N 17 39 E
Jeseník ..... 50 14 N 17 12 E
Jičín ..... 50 26 N 15 22 E
Jihlava ..... 49 24 N 15 35 E
Jindřichův Hradec ..... 49 09 N 15 00 E
Jirkov ..... 50 30 N 13 27 E
Kadaň ..... 50 23 N 13 16 E
Karlovy Vary .... 50 13 N 12 54 E
Karviná ..... 49 52 N 18 33 E
Kežmarok ..... 49 08 N 20 26 E
Kladno ..... 50 09 N 14 06 E
Klatovy ..... 49 24 N 13 18 E
Kolárovo ..... 47 55 N 18 00 E
Kolín ..... 50 02 N 15 12 E
Komárno ..... 47 46 N 18 08 E
Kopřivnice ..... 49 36 N 18 09 E
Košice ..... 48 42 N 21 15 E
Krnov ..... 50 06 N 17 43 E
Kroměříž ..... 49 18 N 17 24 E
Kutná Hora ..... 49 57 N 15 16 E
Kyjov ..... 49 01 N 17 07 E
Kysucké Nové Mesto ..... 49 18 N 18 47 E
Lanškroun ..... 49 55 N 16 37 E
Levice ..... 48 13 N 18 36 E
Levoča ..... 49 02 N 20 36 E
Liberec ..... 50 47 N 15 03 E
Liptovský Mikuláš . 49 05 N 19 37 E
Litoměřice ..... 50 32 N 14 08 E
Litomyšl ..... 49 52 N 16 19 E
Litovel ..... 49 43 N 17 05 E
Litvínov ..... 50 36 N 13 37 E
Louny ..... 50 21 N 13 48 E
Lovosice ..... 50 31 N 14 04 E
Lučenec ..... 48 20 N 19 40 E
Malacky ..... 48 26 N 17 01 E
Mariánské Lázně . 49 58 N 12 42 E
Martin ..... 49 04 N 18 56 E
Mělník ..... 50 21 N 14 29 E
Michalovce ..... 48 45 N 21 56 E
Mladá Boleslav .. 50 25 N 14 54 E
Moravská Třebová ..... 49 45 N 16 40 E
Most ..... 50 32 N 13 39 E
Myjava ..... 48 45 N 17 34 E
Náchod ..... 50 25 N 16 10 E
Neratovice ..... 50 16 N 14 31 E
Nitra ..... 48 19 N 18 05 E
Nová Dubnica ... 48 47 N 18 10 E
Nové Zámky .... 47 59 N 18 10 E
Nový Bor ..... 50 46 N 14 35 E
Nový Jičín ..... 49 36 N 18 01 E
Nymburk ..... 50 11 N 15 03 E
Odry ..... 49 40 N 17 50 E
Olomouc ..... 49 35 N 17 15 E
Opava ..... 49 57 N 17 55 E
Orlová ..... 49 51 N 18 25 E
Ostrava ..... 49 50 N 18 17 E
Ostrov ..... 50 18 N 12 57 E
Otrokovice ..... 49 12 N 17 32 E
Pardubice ..... 50 02 N 15 47 E
Partizánske ..... 48 38 N 18 23 E
Piešťany ..... 48 36 N 17 50 E
Písek ..... 49 18 N 14 09 E
Plzeň ..... 49 45 N 13 22 E
Poděbrady ..... 50 09 N 15 08 E
Poprad ..... 49 03 N 20 18 E
Považská Bystrica ..... 49 07 N 18 27 E
Prachatice ..... 49 01 N 14 00 E
Prague (Praha) .. 50 05 N 14 28 E
Přerov ..... 49 27 N 17 27 E
Prešov ..... 49 00 N 21 15 E
Příbram ..... 49 42 N 14 01 E
Prievidza ..... 48 46 N 18 38 E
Prostějov ..... 49 28 N 17 07 E
Púchov ..... 49 08 N 18 20 E
Rakovník ..... 50 06 N 13 45 E
Revúca ..... 48 41 N 20 07 E
Říčany ..... 49 59 N 14 39 E
Rimavská Sobota . 48 23 N 20 02 E
Rokycany ..... 49 44 N 13 36 E
Roudnice nad Labem ..... 50 25 N 14 15 E
Rožňava ..... 48 40 N 20 32 E
Rožnov ..... 49 28 N 18 08 E
Rumburk ..... 50 57 N 14 34 E
Ružomberok .... 49 05 N 19 19 E
Rychnov nad Kněžnou ..... 50 10 N 16 17 E
Rýmařov ..... 49 56 N 17 16 E
Šaľa ..... 48 09 N 17 53 E
Šamorín ..... 48 02 N 17 19 E
Semily ..... 50 36 N 15 20 E
Senec ..... 48 13 N 17 24 E
Senica ..... 48 41 N 17 22 E
Sered' ..... 48 17 N 17 44 E
Skalica ..... 48 51 N 17 14 E
Slaný ..... 50 14 N 14 06 E
Snina ..... 48 59 N 22 09 E
Sokolov ..... 50 11 N 12 38 E
Spišská Nová Ves ..... 48 57 N 20 34 E
Stará Ľubovňa .. 49 18 N 20 42 E
Šternberk ..... 49 44 N 17 18 E
Strakonice ..... 49 16 N 13 54 E
Studénka ..... 49 44 N 18 05 E
Štúrovo ..... 47 48 N 18 44 E
Šumperk ..... 49 58 N 16 58 E
Šurany ..... 48 05 N 18 11 E
Sušice ..... 49 14 N 13 31 E
Svidník ..... 49 18 N 21 34 E
Svitavy ..... 49 45 N 16 28 E
Tábor ..... 49 25 N 14 40 E
Tachov ..... 49 48 N 12 38 E
Teplice ..... 50 38 N 13 50 E
Tišnov ..... 49 21 N 16 26 E
Topoľčany ..... 48 34 N 18 11 E
Třebíč ..... 49 13 N 15 53 E
Trebišov ..... 48 38 N 21 43 E
Třeboň ..... 49 00 N 14 45 E
Trenčín ..... 48 54 N 18 02 E
Třinec ..... 49 41 N 18 39 E
Trnava ..... 48 22 N 17 36 E
Trutnov ..... 50 34 N 15 54 E
Turnov ..... 50 35 N 15 10 E
Uherské Hradiště . 49 04 N 17 27 E
Uherský Brod ... 49 02 N 17 39 E
Uničov ..... 49 46 N 17 08 E
Ústí nad Labem .. 50 40 N 14 02 E
Ústí nad Orlicí ... 49 59 N 16 24 E
Valašské Meziříčí . 49 28 N 17 58 E
Varnsdorf ..... 50 55 N 14 37 E
Velké Meziříčí ... 49 21 N 16 01 E
Vel'ký Krtíš ..... 48 12 N 19 21 E
Veselí nad Moravou ..... 48 57 N 17 24 E
Vlašim ..... 49 42 N 14 54 E
Vranov ..... 48 54 N 21 41 E
Vrchlabí ..... 50 38 N 15 36 E
Vsetín ..... 49 20 N 18 00 E
Vyškov ..... 49 17 N 17 00 E
Vysoké Mýto .... 49 57 N 16 10 E
Zábřeh ..... 49 53 N 16 52 E
Žatec ..... 50 20 N 13 33 E
Žďár nad Sázavou ..... 49 35 N 15 56 E
Žiar nad Hronom . 48 35 N 18 52 E
Žilina ..... 49 13 N 18 44 E
Zlaté Moravce ... 48 23 N 18 24 E
Zlín ..... 49 13 N 17 40 E
Znojmo ..... 48 51 N 16 03 E
Zvolen ..... 48 35 N 19 08 E


**Physical features and points of interest**


Berounka, *river* .. 50 00 N 14 24 E
Beskid Mountains ..... 49 25 N 20 15 E
Bielé Karpaty Mountains ..... 48 55 N 17 40 E
Bohemia (Čechy) . 50 00 N 14 30 E
Bohemian Forest . 49 00 N 13 00 E
Bohemian-Moravian Uplands ..... 49 25 N 15 40 E
Carpathian Mountains ..... 49 30 N 21 00 E
Čechy, see Bohemia
Černá Mountain .. 48 59 N 13 34 E
Český Les, *mountains* ..... 49 40 N 12 30 E
Chopek, Mount .. 48 57 N 19 35 E
Danube (Dunaj), *river* ..... 47 48 N 18 50 E
Dukla Pass (Dukelský Průsmyk) ..... 49 25 N 21 42 E
Dunaj, see Danube
Dyje, *river* ..... 48 37 N 16 56 E
Elbe (Labe), *river* ..... 50 20 N 14 20 E
Erzgebirge (Krušné hory), *mountains* ..... 50 30 N 13 10 E
Gerlachovský Peak ..... 49 10 N 20 08 E
High Tatra Mountains, see Vysoké Tatry Mountains
High Tatra National Park ... 49 10 N 20 10 E
Hornád, *river* .... 48 00 N 20 56 E
Hron, *river* ..... 47 49 N 18 45 E
Hrubý Jeseník Mountains ..... 50 05 N 17 10 E
Ipeľ, *river* ..... 47 49 N 18 51 E
Javorie, Mount .. 48 26 N 19 16 E
Javorníky Mountains ..... 49 20 N 18 27 E
Jihlava, *river* .... 48 55 N 16 36 E
Jizera, *river* ..... 50 11 N 14 43 E
Klet', Mount .... 48 52 N 14 17 E
Krkonoše Mountains ..... 50 45 N 15 40 E
Krkonoše National Park ... 50 40 N 15 40 E
Krušné hory, see Erzgebirge
Labe, see Elbe
Laborec, *river* ... 48 36 N 22 00 E
Lidice, *historical site* ..... 50 08 N 14 12 E
Lipno Reservoir .. 48 45 N 14 01 E
Little Alfold (Little Hungarian Plain) ..... 47 30 N 17 00 E
Low Tatra Mountains, see Nízke Tatry Mountains
Low Tatra National Park ... 49 00 N 19 40 E
Lusatian Mountains ..... 50 50 N 14 35 E
Lužická Nisa, see Neisse
Lužnice, *river* .... 49 14 N 14 23 E
Macocha Gorge .. 49 23 N 16 38 E
Morava, see Moravia
Morava, *river* .... 48 10 N 16 59 E
Moravia (Morava) . 49 30 N 17 00 E
Myslivna, Mount . 48 38 N 14 41 E
Neisse (Lužická Nisa), *river* ..... 51 30 N 15 00 E
Nitra, *river* ..... 47 48 N 18 08 E
Nízke Tatry (Low Tatra) Mountains ..... 48 55 N 19 45 E
Nové Mlýny Reservoir ..... 48 53 N 16 40 E
Oder (Odra), *river* ..... 49 45 N 18 00 E
Ohře, *river* ..... 50 32 N 14 08 E
Ondava, *river* ... 48 27 N 21 49 E
Orava, *river* ..... 49 09 N 19 09 E
Orlice Mountains . 50 16 N 16 15 E
Oslava, *river* .... 49 06 N 16 22 E
Otava, *river* ..... 49 27 N 14 12 E
Pieniny National Park ..... 49 25 N 20 30 E
Podunajská rovina, see Little Alfold
Poprad, *river* .... 49 35 N 20 39 E
Praha, Mount ... 49 40 N 13 49 E
Sázava, *river* .... 49 48 N 14 24 E
Slaná, *river* ..... 48 15 N 20 20 E
Slanské Mountains ..... 48 45 N 21 30 E
Slovakia (Slovensko) .... 48 40 N 19 30 E
Slovenské Rudohorie, *mountains* ..... 48 45 N 20 20 E
Slovensko, see Slovakia
Sněžka, Mount .. 50 43 N 15 49 E
Sudeten Mountains ..... 50 30 N 16 00 E
Šumava Mountains ..... 48 55 N 13 35 E
Svratka, *river* .... 48 55 N 16 36 E
Topľa, *river* ..... 48 44 N 21 45 E
Torysa, *river* .... 48 39 N 21 22 E
Úhlava, *river* .... 49 45 N 13 24 E
Už, *river* ..... 48 36 N 22 00 E
Váh, *river* ..... 47 55 N 18 01 E
Veľká Domaša Reservoir ..... 49 02 N 21 41 E
Vel'ký Milič Mountain ..... 48 35 N 21 28 E
Vltava, *river* ..... 50 20 N 14 29 E
Vysoké Tatry (High Tatra) Mountains ..... 49 10 N 20 00 E
Závist, *historical site* ..... 49 57 N 14 20 E
Zemplínska Reservoir ..... 48 46 N 22 00 E
Ževlika Reservoir . 49 40 N 15 08 E



Štrbské Pleso, a morainic lake in the High Tatras (Vysoké Tatry), Tatra National Park.
Frantisek Malecek

To the east of the Erzgebirge, and north of the central Bohemian Plateau, are the Sudeten (Czech and Polish, Sudety) Mountains, which form most of the border between Czechoslovakia and Poland west of the city of Ostrava. In the Sudeten is found the greatest altitude in the Bohemian Massif, 5,256 feet (1,602 metres), at Sněžka in the Krkonoše Mountains ("Giant Mountains"). Ostrava lies in the smallest unit of the Bohemian Massif, called the Oder Lowland. It is a small fringe along the Polish border that may be considered to be a foreland of the eastern Sudeten.

The Brno System (Brněnská soustava), south of the central Sudeten Mountains and east of the Bohemian Plateau, is the last of the uplands that enclose the plateau. The most spectacular feature of the Brno System is the Moravian Karst, north of the city of Brno.

'arpathian
lountains

The Carpathian Mountains dominate the eastern part of the country. They comprise a system of east–west ranges separated by valleys and intermontane basins that can be divided into four orographic regions: the Outer Carpathian Depressions (Vnékarpatské sníženiny), the Inner Carpathian Depressions (Vnitrokarpatské sníženiny), the Outer Carpathians, and the Inner Carpathians.

The Outer Carpathian Depressions, known to geographers as the Moravian Corridor, are the narrow lowlands that separate the Bohemian Massif to the west from the easternmost Carpathians; these include the valleys of the upper Oder and Morava rivers and the headstreams of the Dyje. The southern courses of the Morava and Dyje belong to the Inner Carpathian Depressions. Two other large lowland areas north of the Hungarian border comprise the remainder of the Inner Carpathian Depressions region; these are the Danubian Lowland (Podunajská nížina) extending north of the Danube River and the Eastern Slovakian Lowland (Východoslovenská nížina) along the Bodrog and Ondava rivers. Both lowlands are probably better classified as outliers of the Great Alfold, the vast plain to the south, than as intermontane basins.

The Carpathian Mountains proper, which extend into southern Poland, consist of two regions, the Outer Carpathians, to the north, composed of flysch, with a well-developed nappe (horizontally overthrust and folded sedimentary beds) structure; to the south, the Inner Carpathians, composed of a crystalline core mantled by Mesozoic sediments and, farther south, by Neogene volcanic series. The highest ranges are the Low Tatras (Nízke Tatry), of the Inner Carpathians, reaching about 6,000 feet, and the High Tatras (Vysoké Tatry), containing the highest point in the republic, Gerlachovsky Peak (Gerlachovský štít), 8,711 feet (2,655 metres).

According to the definitions of Czechoslovak geographers, mountains and high mountains occupy 23 percent of the country, uplands 67 percent, and lowlands only a scant 9 percent.

**Drainage.** Czechoslovakia lies in the headwater area of the main European watershed, with the Elbe draining more than a third of the country, the Oder and Vistula nearly a tenth, and the Danube more than half. There are many small rivers of little economic importance. The greatest flows generally occur in spring, when the rivers are swollen by melting snow, and the lowest in summer. The country is rich in mineral springs, and good use is made of the groundwater reserves.

**Soils.** Rich, black chernozem soils (about one-tenth of the total) and good-quality brown soils (about one-fifth) occupy the drier and lower portions of the country. The soils known as podzols, occurring in regions with more than 23 inches (600 millimetres) of rainfall annually, constitute half of all of the soils in the country but only a third of the agricultural land. Stony mountain soils in the highest regions and rich alluvial soils alongside river courses are characteristic. This soil variety is again further complicated by topographic variations, with loamy-sandy soils most common in the Bohemian Massif and heavy clay soils predominant in peripheral zones of the Carpathians.

**Climate.** The Czechoslovak climate is mixed, its continental tendencies being masked by large fluctuations in both temperature and precipitation and oceanic influences diminishing from west to east. Relief variations show up in temperature figures: the mean annual temperature drops to 25.4° F (−3.7° C) in the High Tatras, rising to 50.9° F (10.5° C) in the Danube lowlands. Average July temperatures exceed 68° F (20° C) in the Danube lowlands, and average January temperatures can be as low as 23° F (−5° C) in mountain basins. The growing season is about 200 days in the south and less than half that in the mountains. Annual precipitation ranges from 18 inches (450 millimetres) in the central Bohemian basins, to more than 60 inches (1,500 millimetres) on windward mountain slopes in the Krkonoše Mountains. Maximum precipitation falls in July, minimum in February. There are no recognizable climatic zones but rather a succession of small and varied districts; climate thus follows the topography in contributing to the diversity of the natural environment.

Tempera-
ture and
precipita-
tion

**Plant and animal life.** Czechoslovakia's vegetation may be divided into three major groups: Hercynian (central and western European), Pannonian (steppe), and Carpathian (eastern European and Siberian). Each of these groups is modified by altitudinal and other climatic influences. Al-

though large areas of original forest have been cleared for cultivation and for timber, forests still cover about one-third of the country. The major forest types include the oak (*Quercus robur*), steppe oak (*Q. pubescentis*), and oak-grove assemblages of the lowlands and uplands adjoining the Danubian Plain; the beech forests of the Carpathians (*Fagus sylvatica*); and the spruce (*Picea abies*) forests of the Carpathians and adjoining areas of Bohemia and Moravia. In the very highest elevations can be found taiga and tundra assemblages characteristic of more northerly or more elevated regions elsewhere in Europe. Since cultivation is mostly restricted to areas of level terrain, the proportion of original forest cover is highest in mountainous areas.

The timberline runs at about 4,500 feet above sea level in the west, rising to about 5,000 feet in the east. At these higher elevations, particularly in the Tatras, the tree cover below the timberline consists of little more than dwarf pine (*Pinus mugo*). The alpine zone contains grass and bushes (*Vaccinium, Calluna, Salix retusa,* and others); above 7,500 feet, lichens predominate.

Wildlife

Czechoslovakia's wildlife is extensive and varied; the High Tatra National Park probably shelters the most exceptional collection of wild animals, including bears, wolves, lynx, wild cats (*Felis silvestris*), marmots, otters, martens, and minks. Most of these may be hunted, although the parks themselves are fully protected. Certain other animals, however, are protected everywhere; the chamois, for example, very near extinction for many years, now appears to be increasing in numbers. Game birds, especially pheasants, partridges, wild geese, and ducks, are common in forests and marshes. All may be hunted, but larger species such as eagles, vultures, ospreys, storks, eagle owls, bustards, and capercaillies are much rarer and in many cases are protected.

The preservation of the natural heritage is an important goal of the Czechoslovak government. Rare or endangered species such as the mouflon (a mountain sheep) are bred in game reserves, and nature reserves have been created to preserve especially important landscapes: the Šumava Forest, Moravian Karst, and Jizera Mountains (Jizerské hory) are among the most notable. Tourists are given controlled access to the reserve areas.

The two largest and most important areas, however, are the High Tatra and Krkonoše national parks. The former, established in 1948, is an international area, being contiguous with, and jointly administered with, the Polish Tatras National Park. It preserves most of the species of wildlife mentioned above, together with a glacial landscape, alpine flora and fauna, and relict species from the Pleistocene glaciations. Krkonoše National Park, established in 1963, protects glacial and alpine landscapes and vegetation and some relict boreo-Arctic species, such as the alpine shrew (*Sorex alpinus*); it is extensively developed, however, as a ski resort.

**Settlement patterns.** Industrialization and urbanization have changed the face of many of the traditional regions of Czechoslovakia, although Bohemia, Slovakia, and to a lesser extent Moravia are still recognizable entities, reflecting differing national and cultural heritages. Local traditions are preserved in the mountainous parts of Slovakia in the wearing of folk costumes on holidays and in the preparation of local dishes, the making of cheese, and the preserving of fruits and vegetables. Similar traditions have been retained in southern Bohemia and in southeastern Moravia.

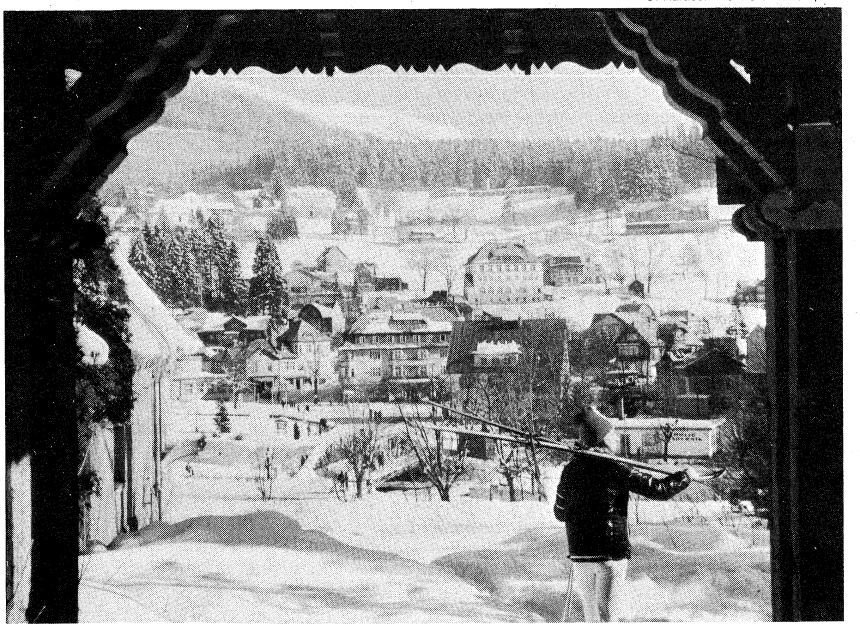
The country has a high density of settlement, with communities lying, on average, only a few miles apart. Rural settlements are characteristically compact, but in the mountainous regions, colonized during the 13th and 14th centuries, straggling row villages are common. Dispersed rural settlements tend to be exceptional, occurring mostly on the Moravia–Slovakia boundary and in central Slovakia and reflecting the later colonization of the 17th and 18th centuries. Rural settlements with up to several hundred inhabitants tend to prevail except in southern Moravia and southwestern Slovakia. The collectivization of farmland that took place in the decades following World War II has resulted in a pattern of large, regularly shaped fields, replacing the centuries-old division of land into small, irregular, privately owned plots.

Villages

Urbanization in Czechoslovakia is not particularly high for an industrialized country. Even the smallest urban centres, however, usually contain some manufacturing industry. Urbanization is greater in the Czech Lands than in Slovakia, although postwar industrialization in the latter has reduced the disparity. Prague, the national capital, historically occupies a predominant role. Bratislava is the capital of Slovakia and Brno the chief city of Moravia. The other large cities are Ostrava, the leading Czech coal-mining and steel centre; Košice, with its large steel complex; and Plzeň, with old established engineering and brewing industries.

New towns were founded both before and after World War II. Prewar settlements include Gottwaldov (formerly Zlín) and Partizánske (in western Slovakia), founded in 1923 and 1935, respectively. The towns of Havířov, Nová Dubnica (near Trenčín), Ostrov (near Karlovy Vary), and

O. Karasek—© Peter Arnold, Inc.



Ski resort in the Krkonoše Mountains ("Giant Mountains"), a range in the Sudeten Mountains in northeastern Bohemia.

Chodov (near Sokolov) have been built since World War II, and all have substantial populations.

## THE PEOPLE

**Ethnic composition.** Anthropologically, the population of Czechoslovakia is very mixed. About half the people exhibit the physical traits of the Alpine local race and thus are reminiscent of western Europeans, while eastern European Baltic elements predominate farther east and Dinaric elements (related to the mountainous region of Yugoslavia) manifest themselves in Slovakia. Cultural exchanges resulting from dependence on the Austro-Hungarian monarchy from 1526 until 1918 resulted in an influx of German and western European influences, whereas the linguistic affinity, together with political developments since World War II, has worked to strengthen ties with eastern Europe.

Czechs and Slovaks form the great majority of the population, in contrast to the situation before World War II, when minorities made up more than a third. The postwar expulsion of Germans, an exchange of populations with Hungary, and the cession of Ruthenia (with a Ukrainian population) to the Soviet Union brought about the present homogeneous state. Czechs are about twice as numerous as Slovaks, and minorities include a large number of Hungarians in southern Slovakia, small groups of Ukrainians near the Soviet frontier, and a few Poles in northeastern Moravia. Some Germans still live in northwestern Bohemia. Gypsies—who have no legal status as an ethnic group—constitute, according to some sources, the largest such population in any one country.

Postwar changes

**Linguistic composition.** The majority of the population speak Czech as their first language, while Slovak is the first language of about one-third. These very similar languages—some linguists consider Slovak to be a dialect of Czech, while others consider it a language in its own right—are part of the western group of Slavic languages that uses the Roman rather than the Cyrillic alphabet. Czech as a literary language dates to the late 13th century, and Slovak as a literary language dates to the 19th century. Hungarian, Polish, German, Ukrainian, and Russian are among the other languages that are spoken in Czechoslovakia.

**Religions.** During the Communist era, no official statistics were kept on religion, though the activities of the churches were financed by the government following the nationalization of all church property in 1949. Atheism was the official government philosophy, and the churches' role was largely restricted to religious rites. Clergy were civil servants and were required to take an oath of loyalty to the government. Religious freedom was restored in 1989. A visit to the country by Pope John Paul II in April 1990 celebrated the resurgence of Roman Catholicism, the predominant religious sect. There is an Evangelical element in Slovakia.

**Demographic trends.** The changes in way of life associated with urbanization and employment of women in the labour force help to explain Czechoslovakia's rather low rate of population increase. Internal movements of population are of increasing significance with the further modernization of the economy. Migration is directed from the country to the towns and from the smaller towns to the larger cities. Although the lower natural growth rate in the western portion of the country has been supplemented by immigration from Slovakia, this movement has declined with the growth of urbanization in Slovakia itself.

In the past, population growth was slowed by emigration to the urban centres of the Austro-Hungarian Empire and farther afield, especially to the United States. Eastern and central Slovakia had the heaviest losses.

The most complicated demographic problems in contemporary Czechoslovakia stem from the fact that, with only moderate birthrates and a prolongation of average life span, the population is aging. A high percentage of the population in central and eastern Bohemia is elderly; Slovakia continues to have a higher birthrate, and thus a younger population, than the Czech Lands.

Aging population

Geographic conditions make possible the settlement of almost the whole of Czechoslovakia, but the population tends to be concentrated at lower altitudes, more than half the population living below about 1,300 feet above sea level. Population density declines abruptly with increasing altitude. A notable feature is the low density in some frontier areas, partly reflecting the induced emigration of minorities after World War II. (M.Bl./R.H.O./F.W.C.)

## THE ECONOMY

Czechoslovakia is in transition from a centrally planned to a capital market economy. Under the planned economy, most of the agricultural land belonged to the state, and industry, banking, and trade were under state control. Only very small-scale private enterprises were tolerated. Economic planning was based on five-year plans.

The prerequisites for the introduction of an efficient market include institutional support for the development of the private sector, financial and capital markets, a more diversified banking system, attractive conditions for foreign capital investment, and demonopolization of the economy. Laws to be amended under the democratic government include those on state enterprises, joint stock companies and commercial codes, joint ventures, economic activities for foreigners, and private enterprise activity. The main points of the country's reform program include private enterprise, land reform, and enterprises with foreign participation.

Between the end of World War II and 1975 the country enjoyed extensive economic growth, but thereafter a dramatic slowdown took place in all economic performance. Czechoslovakia, like its Socialist neighbours, began to feel the impact of the economic recession in the West, with a subsequent decline in markets for Czechoslovakian products. Moreover, the country suffered from a deterioration in its terms of trade with the Soviet Union, particularly in 1975 when the Soviet Union increased the price of oil exported to its satellite nations.

De-escalation in the two major branches of the economy, industry and agriculture, in turn caused downturns in the construction industry and in transport, communications, and other services. The economy slowed in 1989, with the net material product growing by less than 2 percent. Related economic problems at the time of the change of government included declining exports, a surplus of outdated machinery and technology, and a shortage of high-quality consumer goods.

*Taxation.* Because the country under the former system derived substantial revenue from the profits of state enterprises, taxes played a relatively small role in the government's budget. The principal tax was the turnover tax on consumer goods that was levied on producer, wholesaler, retailer, and ultimate buyer. This was followed by the payroll tax levied on enterprises; the remainder was made up of a large number of different levies, the yield of which was insignificant. Reform of the tax system was expected to be given high priority on the economic agenda of the new government.

*Labour force.* About half of the total population is in the labour force. The female employment rate is among the highest in the world, almost half of the employed people being women. Industry predominates in urban areas and, although the number of persons engaged in agriculture increases eastward, only a quarter of the total rural population works in agriculture.

*Trade unions.* Most employees are members of trade unions, which prior to the events of 1989–90 were controlled by the Communist Party. The unions are now in a transitional period awaiting the outcome of economic reform.

**Resources.** Conversion of domestic heating systems from coal to gas is far behind schedule. Progress has been made in the construction of nuclear power facilities, with the last increase in capacity coming from the Dukovany power station in 1987. A new plant at Mochovce was scheduled for completion in 1993–94, and one at Temelín was expected to begin generating in late 1992. Nuclear power generation increased by 10 percent in 1989, accounting for 28 percent of total electricity generation, with a saving of 120 million tons of brown coal.

Although reserves are limited, black and brown coals are Coals

produced in significant quantities. Most of the black coal is derived from the Ostrava-Karviná coalfield, though it is also mined near Kladno, in the Plzeň basin, near Trutnov, and near Brno. A high proportion of the black coal is of coking quality. Production of brown coal increased rapidly up to the mid-20th century and has remained fairly static since then, although demand for it was maintained by the delay in conversion to gas heating; open-pit methods are used. The main mining areas are around Chomutov, Most, Teplice, and Sokolov. The brown coal is used extensively in thermal power stations and as fuel in the home, and large quantities are also utilized as raw material in the chemical industry. Small quantities of petroleum and natural gas are produced near Hodonín. Pipelines import Soviet oil and natural gas, the latter supplementing existing coal gas supplies. There has been a slight drop in energy use. Brown coal output declined in 1989, owing to inclement weather, a decrease in military production after Warsaw Pact arms cuts, and the failure in growth of new industrial sectors.

Czechoslovakia has a limited endowment of metallic ores. Iron ore is mined chiefly in the Slovak Ore Mountains (Slovenské Rudohorie), while areas between Prague and Plzeň are historically important. Copper and manganese ores are also mined in the Slovak Ore Mountains. Lead and zinc ores are mined near Kutná Hora, Příbram, and Banská Štiavnica and in the Jeseník Mountains. Uranium is mined near Příbram and around Hamr in northern Bohemia. A significant gold deposit verified at Mokrsko, in central Bohemia south of Prague, is estimated to contain 100 tons of the metal. The Krušné Mountains of Bohemia yield small quantities of tin. Imported bauxite and nickel ore are refined at Žiar nad Hronom and Sered', respectively. Other mineral resources include salt in eastern Slovakia, graphite near České Budějovice, and kaolin near Plzeň and Karlovy Vary.

**Agriculture and forestry.** Czechoslovak agriculture is among the most advanced in eastern Europe, with better than average yields. Agriculture, along with the food industry as a whole, will, under current economic reforms, be subject to an injection of market discipline, which is expected to render it more responsive to consumer demand and business productivity. Land is to be restored to the individuals from whom it was confiscated after World War II. This is expected to result in the reestablishment of private farming in some cases and in the continuation of some cooperative ventures on a share basis. The country does not suffer from a shortage of agricultural land, but such land in Czechoslovakia is used far less efficiently than that in western Europe.

Agricultural reform

On the fertile lowlands, wheat, barley, sugar beets, corn (maize), and fodder crops are the most important, but on the relatively poor soils of the mountains the principal crops are rye, oats, and potatoes. Cereals lead in total production, with wheat, barley, rye, and oats as the most important crops. The country also raises a sizable amount of pigs and cattle.

In 1984 a decision was made to expand forest areas through reforestation, but by the late 1980s forests were being cut more rapidly than planned because trees were dying from the effects of pollution. By 1982, 26 percent of Czech Lands forests and 14 percent of those in Slovakia were destroyed or seriously damaged; by 1989 the figures were 57 percent and 30 percent, respectively, with a forecast of 70 percent and 40 percent by the year 2000. The most endangered areas are in the western Beskid Mountains, on the Moravia–Slovakia border, and in central Slovakia near Žiar nad Hronom.

**Industry.** Although overall industry in Czechoslovakia can be characterized as obsolete, some sectors, notably electronics, are modern and efficient and could be competitive in international markets, given an infusion of investment and management skill and the curtailment of state interference. Engineering is the largest branch of industry, constituting more than a quarter of total industrial production. Also important are the food industry, followed by electrotechnics, the chemical, rubber, and asbestos industries, ferrous metallurgy, and fuel mining and processing. Czechoslovakia's iron and steel industries are among the largest in eastern Europe, based largely on imported ores (especially on ores from the Ukrainian S.S.R.). Steel production is centred on the plants of the Ostrava area (in Moravia), with lesser amounts produced at Kladno, Plzeň, and Chomutov (all in Bohemia); the major centre in Slovakia is Košice.

Czechoslovakia has also been foremost in the production of a number of engineering products, including automobiles and electric and diesel railroad locomotives.

In the Communist period, tourism was supervised by Čedok, the official state travel bureau. Most foreign visitors were from other Communist countries, notably East Germany, Hungary, and Poland. Democratic Czechoslovakia is becoming a popular destination for Western tourists. Winter and summer sports, national parks and nature reserves, and touring in the mountains are among the major attractions. Monuments and castles are visited throughout the country, and Prague, with its important cultural and architectural heritage, is a popular attraction. Although there has been an increase in foreign visitors, the number of nights spent in Czechoslovakia has declined. This is partly explained by the growth of "shopping" tourism, especially by visitors from the Soviet Union, Hungary, Yugoslavia, and Poland who purchase goods that are in short supply in their own countries.

**Finance.** The traditional monobank, the State Bank of Czechoslovakia, has been split into a central bank and a limited number of commercial banks. There are to be two universal commercial banks (the Czech Commercial Bank and the Slovak Commercial Bank) and a new Investment Bank (which will be another commercial bank). The activities of the two existing specialized banks (the Czechoslovak Foreign Trade Bank and Živnostenska Bank) will be broadened. This means a considerable enhancement of the status of banking in the country.

Banking reforms

**Trade.** The country is highly dependent on foreign trade; in per capita terms it is among the largest foreign traders in eastern Europe. Exports to eastern European countries have declined but still show a surplus. There is also a clear surplus with countries of the Organisation for Economic Co-operation and Development, resulting from high demand for raw materials, especially steel. Higher world prices account for more than half the increase in export value. The engineering sector has seen an export decline. Outside of eastern Europe, West Germany was (before reunification) the major partner for both exports and imports. (E.I.U./R.H.O./F.W.C.)

**Transportation.** Czechoslovakia has a modernized transport system. Transit functions, reflecting the country's role as a link in trans-European trade, lost much of their importance in the new political and economic conditions that prevailed on the Continent after 1945. A new transportation role for Czechoslovakia was expected to result from the regionwide changes of 1989.

*Railways.* The most important element of the transport system is the railways, which are especially important in freight transport, notably of coal, ores, metals, and building materials. Density of the railway network is fairly high, but the system is not always effective, in part because of low operating efficiency. The railways developed in the second half of the 19th century and mostly ran from north to south, to Vienna and to Budapest, whereas the main direction of present-day transportation, developed after World War II, is from east to west. Rail links with the main trading partner, the Soviet Union, have been facilitated by an extensive program of reconstruction and electrification. Most freight moves along main-line routes, but shorter routes connecting the larger towns are also used heavily for passenger traffic.

Railway network

*Road, water, and air transport.* The density of roads and the number of scheduled bus routes are also high, and the transport of passengers and freight by state-owned vehicles predominates. Again, the urban centres are the focus of the most intensive road networks. Although Czechoslovakia's roads are of good quality, there are few of the throughways and expressways familiar in the West, and the Western visitor is surprised by the small number of private cars seen on the Prague–Brno–Bratislava highway. Only in Prague and its surroundings and on weekends in

other large cities does traffic flow approach the western European level: the number of private automobiles is relatively low and gasoline is rationed.

Czechoslovakia's position in the headwater area of the main European watershed means that the potential of water transport is small; of the rivers flowing through the country, including the Danube, only small portions actually in the country are navigable. Freight is transported on the Elbe to Hamburg, which functions as a port for mixed cargo, but Danube river transport, with Komárno the largest Czechoslovak port, now predominates, carrying raw materials imported from the Black Sea ports of the Soviet Union and its neighbours. The Oder gives access to the Polish port of Szczecin. Major changes in freight transportation have been introduced by the Friendship (Druzhba) oil pipeline, coming into the country from the Volga River region of the Soviet Union, and also by a gas line from Siberia, which extends beyond Czechoslovakia to western Europe.

Air connections between Prague, Bratislava, and other regional centres are of considerable importance, and Prague as well as Bratislava and Poprad–Tatry have international airports.

## GOVERNMENT AND SOCIAL CONDITIONS

**Government.** *The constitutional framework.* Following the collapse of the Socialist government in 1989–90, Czechoslovakia reestablished itself as a federated republic. The bicameral Federal Assembly comprises a Chamber of Nations with 150 delegates (75 Czech and 75 Slovak) and a Chamber of the People with 150 deputies (101 Czech and 49 Slovak). The electoral system is one of universal direct suffrage with a president as head of the federation. The national legislatures include a Czech National Council of 200 deputies and a Slovak National Council of 150 deputies. The federal government consists of a coalition.

Political parties

The main political parties are Civic Forum (a Czech party), Public Against Violence (its Slovak equivalent), Communist Party of Czechoslovakia, Christian Democratic Party, Green Party, Czechoslovak Farmers' Party, Czechoslovak Social Democratic Party, and Czechoslovak Democratic Initiative.

*Local government.* Local administration is controlled by "national" committees that operate on three levels. Czechoslovakia is divided into seven Czech and three Slovak *kraje* (regions), the capitals, Prague and Bratislava, each occupying independent positions. The regions are subdivided into *okresy* (districts); and these, in turn, are subdivided into numerous *obce* (local communities, or communes).

*Justice.* The major components of the judicial system are the federal Supreme Court, separate supreme courts for the Czech and Slovak republics, regional and district courts, and military tribunals. In the civil division, the lawcourts are most concerned with family affairs, including divorce cases; the criminal courts deal with crime and traffic accidents.

*Armed forces.* The four military forces are integrated as the National People's Army, including ground forces, air and air defense forces, frontier guards, and interior guards. A period of military service is required from men over age 17 who are physically and otherwise fit and who are called. Discharged conscripts usually remain in the reserves until age 50. The Soviet Union agreed to withdraw all troops by 1991.

**Education.** State kindergartens are available for children three to six years old. Compulsory education lasts nine years, from age six to 15. Instruction in the Russian language begins in the fourth year of school and is followed later by another major language, usually German or English. General secondary schools are attended by pupils from 15 to 18 years of age and offer courses mainly directed to preparation for university study. Vocational secondary schools offer four-year courses that prepare students to compete for jobs in various branches of the government.

Most students in universities or colleges study teaching, civil engineering, economics, electrical engineering, earth sciences, or medicine. The leading educational institutions, providing four to five years of intensive study, have long-standing traditions. Charles University (Univerzita Karlova, founded 1348) and the Czech Technical University (České vysoké učení technické, founded 1707), both in Prague, are the oldest in Central Europe. After World War II institutions of higher education increased in number, especially in Slovakia. Research work is organized not only in the universities but also in special research institutions. Basic research is carried out by the Czechoslovak Academy of Sciences, which has institutes in Prague, Brno, Bratislava, and elsewhere. The largest student populations are at Prague and Bratislava.

**Health and welfare.** Czechoslovakia is traditionally among the very best-equipped countries medically, although Slovakia is somewhat less well served than the Czech Lands. A majority of doctors work in specialized medical establishments and in district and regional hospitals. For most patients, treatment is free. Many of the spas are in northwestern Bohemia, in such internationally known resorts as Karlovy Vary, Mariánské Lázně, and Františkovy Lázně.

Cooperative housing projects are increasingly popular, and private ownership has remained important. The housing shortage, however, continues to be one of the most severe problems afflicting the country. Almost all housing units are supplied with electricity, water, and bathrooms, and where electricity is not available gas connections are supplied.

Czechoslovakia has a high standard of living as compared to other eastern European countries. The mean monthly wage shows some variation, industrial workers as a whole getting above-average wages, though lower wages can be found, for example, in the clothing industry and in some services; coal miners receive the highest wages. Farmers' incomes are about average. All workers participate in pension insurance, and the retirement age is 60 years for men and 55 for women. Old-age pensions, however, remain low, approaching only 50 percent of mean wages. Regional variations in income may be considerable, as, for example, between mining districts and cities, on the one hand, and predominantly agricultural districts, with families with many children, on the other; some peasants in remote areas of Slovakia live at the subsistence level.

Wages and pensions

The greater part of the average family income is provided by two working spouses. Sick benefits, old-age and disability pensions, and child allowances represent most of the remainder. In addition, medical care, medicines, and education are free, and there are reduced fares for transportation to and from work. Vacations range between two and four weeks per year, and there is generous paid maternity leave.

Following democratization, there was a need to raise the prices of subsidized consumer goods. The situation was expected to be improved by Czechoslovakia's joining the International Monetary Fund in September 1990. At present most consumer subsidies remain untouched. Nevertheless, there were no plans to immediately privatize railways or water distribution. Between 30 and 40 percent of the economy was to remain in state hands. The first businesses to be freed from state ownership were to be small shops, services, and workshops. Certain unspecified sectors of the economy were to be opened to foreign ownership.

(M.Bl./R.H.O./F.W.C.)

## CULTURAL LIFE

The rich cultural tradition of contemporary Czechoslovakia derives from many diverse sources. The position of Bohemia and Moravia on the border between Germans and Slavs, and of Slovakia between Slavs and Magyars, has strongly marked the development of Czech and Slovak cultures. Influences from farther afield have also been strong at various times. Visually the most striking are Italian influences in architecture of the Renaissance and the Baroque periods, while literature, music, the visual arts, and popular culture are indebted to a variety of external influences, many of them from the West. Because of their historical circumstances, both Czech and Slovak traditions are informed, in many of their manifestations, by a strong patriotic sentiment.

**Literature.** The earliest writings were in Latin and, for a time, Old Church Slavonic. (For a full discussion of Czech and Slovak literature, see CZECHOSLOVAK LITERATURE.) Writing in Czech has a continuous tradition from the late 13th century on. There are 14th- and 15th-century chivalrous romances and chronicles, legends of the saints, love lyrics, satires, Bible translations, and religious prose. In the period of the Counter-Reformation there was a serious decline in the official and administrative use of Czech, though the Baroque style brought fresh impulses to poetry and influenced equally the Roman Catholic and non-Catholic writers. There was a renewed flowering of Czech literature in the 19th century, which drew on work done toward the end of the 18th century. Indeed, the national revival of the Czechs and the Slovaks started as a cultural enterprise before it acquired, during 1848, distinctly political overtones.

Revival of literature in the 19th century

For the Czechs and Slovaks to become full-fledged members of the 19th-century community of European nations, their languages had to be codified and developed for modern use and their histories constructed. Josef Dobrovský, a Jesuit priest and scholar who wrote in German, published an outstanding systematic grammar of the Czech language. František Palacký, a historian who turned to politics, had the first volume of an ambitious history of the Czech nation published in German in 1836. The volumes that were published after 1848 appeared both in German and in Czech. The Slovaks, who had survived as a separate ethnic group for about a millennium within the Magyar state, developed their awareness of themselves as a nation in the 19th century, also largely through the means of literature. They opted, from the 1840s and under the influence of L'udovít Štúr, for the development of their own literary language, rather than regarding their speech as a dialect of Czech.

Ján Kollár, a poet of Slovak origin, completed in 1832 his collection of sonnets, *Slávy dcera* ("The Daughter of Sláva"), celebrating the common past of the Slavs. The Czech Romantic school of poetry, early in the 19th century, is best represented by Karel Hynek Mácha and by Karel Jaromír Erben; the Slovaks are represented by Janko Král' among others. In Bohemia, the Romantic movement gave way, in the 1840s, to a more descriptive and pragmatic approach to literature. Božena Němcová's novel *Babička* ("The Grandmother") became a lasting favourite with Czech readers; the journalist and poet Karel Havlíček Borovský tried to acquaint the Czechs with some of the stark facts of political life. Jan Neruda, in his poetry and short stories, domesticated literary sophistication within a familiar Prague framework. Toward the end of the 19th century the historical novels of Alois Jirásek began to claim a wide readership, while poetry moved through Parnassian, Symbolist, and Decadent phases.

The making, and breaking, of the Czechoslovak state between the two world wars was reflected in its literature. Jaroslav Hašek's novel *Osudy dobrého vojáka Švejka za světové války* (1921–23; *The Good Soldier Švejk and His Fortunes in the World War*) made a mockery of authority, especially that of the former Austro-Hungarian regime. Karel Čapek wrote popular plays, novels, and travel books, many of which have been translated into English. Vítězslav Nezval, František Halas, Vladimir Holan, Josef Hora, and Jaroslav Seifert were among other writers who came to prominence during the first half of the 20th century. As World War II and German-imposed censorship closed in, poetry became even more popular than in peacetime: the brief life and work of Jiří Orten is an outstanding example of his tragic generation.

Works of Hašek and Čapek

Before the destruction of Czechoslovak Jewry during the war and then the expulsion of the German minority at the end of the war, Bohemia and Moravia had a strong German literary tradition. In about the middle of the 19th century Adalbert Stifter's descriptions of nature and the common people inspired local followers in southern Bohemia. The German Jewish group of writers in Prague in the first half of the 20th century—Franz Kafka, Rainer Maria Rilke, Franz Werfel, and Max Brod—achieved international recognition.

Among the postwar generation of writers, Bohumír Hrabal has become well known for assembling bizarre patterns of memory, usually in haunting short stories (he also wrote the story of the highly acclaimed film *Closely Watched Trains*). The Slovak novelist Ladislav Mňačko was among the first eastern European writers to criticize Stalin and his system of rule in his popular novel *Ako chutí moc* (1968; *The Taste of Power*). Czechoslovak writers have paid a high price for their political commitment: Mňačko, as well as Milan Kundera and Josef Škvorecký, were forced to live and work abroad. Ludvík Vaculík, a writer of the same generation and similar conviction, was among those whose novels were circulated in Prague as unofficial publications.

**Music.** Opera and symphonic music have long had a high place in Czechoslovak cultural life. Such composers as Bedřich Smetana, Antonín Dvořák, Leoš Janáček, and Bohuslav Martinů, who drew heavily on folk music for their inspiration, achieved international reputations, and their works are often played in the annual spring music festival held in Prague. In addition, jazz and related musical forms, and their performers, are very popular, as is folk music.

Prague festival

**Film.** Despite the limitations originally imposed on them by a small market, Czechoslovak films and film directors have made their international mark, especially since the 1960s. Most Czechoslovak films are conceived on a small scale, with a sharp focus on the everyday, common life of the people and its many eccentricities. Jiří Trnka's drawn and animated films have been popular with audiences of all ages and have established a strong tradition in this medium.

**Fine, applied, and folk arts.** Czechoslovakia has a strong tradition in the graphic arts. This includes many forms of caricature, and Josef Čapek, the brother of Karel Čapek, is remembered for a series of drawings, entitled *The Dictator's Boots,* from the time when Hitler was in the ascendant. Much of Czech graphic art derived its inspiration from popular, narrative art: the happy marriage between Hašek's texts and Josef Lada's illustrations makes this point. Czech and Slovak painters and graphic artists have on the whole followed the broad European movements in the last century or so, but realism generally prevails. One of the best-known painters before independence was Josef Mánes, and after his time Alphonse (Czech, Alfons) Mucha, who captured the elusive fin-de-siècle mood in his paintings and posters, gained world renown. The glass designer and sculptor René Roubíček is also world renowned in his field. In the applied arts, manufactured glass ornaments, the traditional north Bohemian costume jewelry, and toys are probably best known. Popular art has been preserved above all in useful objects in ceramics and wood; embroideries and traditional costumes are now of less importance.

**Libraries and museums.** The largest library is the State Library (Státní knihovna) in Prague, created in 1958 by merger of several older libraries. Other major collections are in the National Museum Library (Knihovna Národního musea), also in Prague and founded in 1818, the Slovak National Library (Matica slovenská) in Turčiansky Sväty Martin, the Slovak Technical Library (Slovenská technická knižnica) of Bratislava, the University Library (Univerzitná knižnica) in Bratislava, and the State Scientific Library (Státní vědecká knihovna) in Brno. Public lending libraries are found in every community.

State Library

There are many museums in Czechoslovakia, both in Prague and in the provinces. Three museums in Prague may be especially mentioned: the National Museum (founded 1818), the National Gallery (1796), and the Museum of Decorative Arts (1885), the last housing one of the world's largest collections of glass.

**Press and broadcasting.** All publishers and news media were, until the political changes in late 1989, subject to approval by the government's Office for Press and Information. The Socialist government owned all telephone, telegraph, television, and radio systems, and news was disseminated by the official Czechoslovak News Agency (Československá tisková kancelář). Freedom of speech and freedom of the press were instituted under the new democratic government. Books are

among the cheapest goods in Czechoslovakia, and translations of foreign literature are popular.

For statistical data on the land and people of Czechoslovakia, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL. (M.Bl./R.H.O./Z.A.B.Z.)

## History

The part of Europe that constitutes the modern state of Czechoslovakia was settled by first Celtic, then Germanic, and finally Slavic tribes over the course of several hundred years. The major political and historical regions that emerged in the area are Bohemia, Moravia, and Slovakia. These regions coexisted, with a constantly changing degree of political interdependence, for more than a millennium before combining to form the modern state of Czechoslovakia in 1918. Each was subject to conquest, each underwent frequent shifts of population and periodic religious upheavals, and at times at least two of the three were governed by rival rulers. Bohemia and Moravia maintained close cultural and political ties and in fact were governed jointly during much of their history. Slovakia, however, which bordered on the Hungarian plain, was ruled by Hungary for almost 1,000 years and was known as Upper Hungary for much of that time. For particular histories from the regional viewpoint, see the articles BOHEMIA, MORAVIA, and SLOVAKIA in the *Micropædia.*

### ORIGINS AND EARLY HISTORY

**Bohemia.** The prehistoric people north of the Middle Danube were of uncertain origin. The Boii, a Celtic people, left distinct marks of a fairly long stay, but its time cannot be firmly established. The name Bohemia is derived through Latin from Celtic origins. The Celtic population was supplanted by Germanic tribes. One of them, the Marcomanni, inhabited Bohemia; others settled in adjacent territories. No outstanding event marked the Marcomanni departure.

Movements of ethnic groups

Archaeological discoveries and incidental references to Bohemia in written sources indicate that the movements of ethnic groups were not always abrupt and turbulent but that the new settlers began to enter the territory before the earlier inhabitants had left it. It can be assumed, therefore, that the Slavic people were coming in groups before the southward migration of the Germanic tribes. In the 6th century Bohemia and the neighbouring territories were inhabited by the Slavs.

While mountains and forests offered protection to Bohemia, the tribes in the lowlands north of the Danube and along its tributaries were hard pressed by the Avars of the Hungarian plains. Attempts to unite the Slavic tribes for resisting the Avars were successful only when directed by such personalities as the Frankish merchant Samo, who gained control of a large territory in which at least part of Bohemia was included. His death in 658 ended the loosely knit state. A more auspicious era dawned after Charlemagne defeated the Avars in the 8th century.

There followed a period of comparative security, in which the concentration of the Slavs into political organizations advanced more promisingly. Soon after 800 three areas emerged as potential centres: the lowlands along the Nitra River; the territory on both sides of the Lower Morava (German, March); and central Bohemia, inhabited by the tribe of the Czechs. In time, the Czechs, protected from foreign intruders, rose to a dominant position. Governed by rulers descended from the legendary plowman Přemysl and his consort Libuše, the Czechs brought much of Bohemia under their control before 800 but failed to defeat the tribes in the east and northeast. Apart from occasional disturbances, such as Charlemagne's invasions (805), the Czech domain was not exposed to war and devastation, and little of the life there came to the notice of clerics who were recording contemporary events in central Europe. (O.O.)

**Unification of Great Moravia.** The earliest known inhabitants of Moravia were the Boii and the Cotini, another Celtic tribe. These were succeeded about 15–10 BC by the Germanic Quadi. The Germanic races were pushed back from the middle Danube by the coming of the Avars in AD 567. The exact date of the arrival of the Slavs in Moravia, as in Bohemia, is uncertain; but by the late 8th century Moravia was settled by the Slavs, who acknowledged no particular tribe but took the general name of Moravians from the Morava River. An important trade route from the Baltic to the Adriatic Sea developed through the Moravian Basin.

When Charlemagne destroyed the Avar Empire in about 796 he rewarded the Moravians for their help by giving them a part of it, which they held as a fief from him. They thus became loosely tributary to him for all their lands, but their princes enjoyed independence and often made war on him and on his successors, Louis I the Pious and Louis the German. By the first half of the 9th century, Moravia had become a united kingdom, under Prince Mojmír I (ruling *c.* 818–*c.* 846). (E.Wn.)

Around 833 Mojmír attached the Nitra region (the western part of modern Slovakia) to his domain. His successor (after 846), Rostislav, consolidated the country and defended it successfully. His relations with the Eastern Frankish Empire (since 843 under Louis the German) were determined by political considerations and by the advance of Christianity into the Slavic areas. The bishoprics of Regensburg, Passau, and Salzburg competed in trying to convert the central European Slavs but achieved only limited success. The Archbishop of Salzburg consecrated a church at Nitra around 828. In 845 Regensburg baptized 14 chieftains from Bohemia. Mojmír's Moravia had apparently more frequent contacts with Passau than with Salzburg. Recent archaeological discoveries indicate that missionaries made noticeable progress before 860; stone churches were built as places of Christian worship at Mikulčice and elsewhere.

Competition of bishoprics

But Rostislav was dissatisfied with the Latin-speaking Frankish clergy and asked the Byzantine emperor Michael III for Slavic-speaking preachers. A group of clerics headed by two brothers of Greek origin, Constantine and Methodius, arrived from Constantinople in 863. They not only preached in Slavic but also translated the sacred books into that language and used them in divine services. To Constantine is attributed the creation of the first Slavic alphabet. After some two and a half years the two brothers journeyed to Rome to ask for the Pope's support of their work and their use of the Slavic language. Constantine entered a convent in Rome, taking the name of Cyril; he died in 869. Methodius received the Pope's sanction for his work in Moravia and in Pannonia, Moravia's southern neighbour. The two territories were organized as a province and connected with the ancient archbishopric of Sirmium, restored by the Pope. Methodius' elevation to archbishop angered the Frankish clergy, who regarded his archdiocese as their missionary field. He was captured and imprisoned until 873; he then returned to Moravia and put himself under the protection of Rostislav's successor Svatopluk. But relations between the ruler and the Archbishop were not harmonious. After Svatopluk's conciliation with the Franks at Forchheim (874), clerics of the Latin rite appeared again in Moravia, interfering with the Archbishop's work. In 880 Methodius obtained from Pope John VIII a formal sanction of his work, including the Slavic liturgy.

Religious conflicts

Svatopluk distinguished himself in the conduct of political affairs. After the death of Louis the German (876), he acquired large territories with Slavic population. The Great Moravia that he created included all Bohemia, the southern part of modern Poland, and the western part of modern Hungary. He annexed some territories and left local princes who recognized his suzerainty in others. Such was apparently the case of the Czech prince Bořivoj I. Propagation of Christianity followed Svatopluk's advances. According to legends, Bořivoj was baptized by Methodius and then admitted clerics of the Slavic rite to his principality. But, while the Archbishop was engaged in missionary work in the annexed territories, advocates of the Latin rite, headed by a Frankish cleric, Wiching, bishop of Nitra, strengthened their position in Moravia. During Methodius' lifetime the Slavic clergy had the upper hand; but after his death (885) Wiching banned Methodius' disciples from Moravia and most of them moved to

Political expansion

Bulgaria. Pope Stephen V reversed his predecessor's policy and forbade the Slavic liturgy.

Svatopluk continued his policy of expansion for several more years. But soon after 890 he made the East Frankish king Arnulf his enemy. Arnulf's expedition into Moravia in 892 opened a period of troubles, which increased when Arnulf made an alliance with the Magyars of Hungary. Svatopluk's successor, Mojmír II, tried unsuccessfully to protect his patrimony; sometime in 905–908 Great Moravia ceased to exist as an independent country.

(O.O.)

**Slovakia.** The country was inhabited in the first centuries AD by Illyrian, Celtic, and then Germanic tribes. The Slovaks—Slavs closely akin to, but possibly distinct from, the Czechs—probably entered it from Silesia in the 6th or 7th century. For a time they were subject to the Avars, but in the 9th century the area between the Morava River and the central highlands formed part of Great Moravia, when the Slovak population accepted Christianity from Cyril and Methodius. In the 890s, however, the German king Arnulf called in the Magyars to help him against Moravia, and Slovakia lay in their path. The Moravian state was destroyed in the first decade of the 10th century, and after a period of disorder Slovakia became one of the lands of the Hungarian crown in the 11th century.

Magyar influence

The main ethnic frontier between Magyars and Slovaks ran along the line where the foothills merge into the plain, though there were also Magyars settled in the larger valleys; later, the landlord class and much of the urban population in the whole area was Magyar. On the other hand, as the country suffered from chronic overpopulation, a constant stream of Slovak peasants moved down into the plains, where they usually were Magyarized in two or three generations.

(E.Wn.)

THE PŘEMYSL RULERS OF BOHEMIA (895–1306)

The Prince of Bohemia made an accord with Arnulf (895) and warded off the danger of invasion. The domain over which the descendants of Přemysl ruled from the Prague castle was, in the early 10th century, the largest unit in Bohemia. The tribal chieftains who opposed centralistic tendencies exercised control over the eastern and northeastern districts, but the extent of their power is not known. The most powerful of them, the Slavníks residing at Libice, remained defiant until the end of the 10th century.

Bohemia maintained close relations with neighbouring Bavaria. Both countries were threatened for several decades by the Magyars, and other developments in their vicinity also affected political and social life. The most important of these was the rise in Germany of the Saxon dynasty, which began with Henry I the Fowler; the imperial coronation of Otto I, in Rome (962), marked the restitution of the Holy Roman Empire, with which Bohemia was linked, thereafter, for many centuries. Bohemia's reorientation toward the Saxon dynasty began under the grandson of Bořivoj, Prince Wenceslas (Václav, ruled 921–929); it was symbolized by the dedication of a stone church at the Prague castle to a Saxon saint, Vitus. Both Slavic and Latin legends praise Wenceslas as a fervent believer but tell little about his political activities. He was murdered by his younger brother Boleslav. The legends present the murder as an outburst against Wenceslas' devotion to the new faith, but the conspiracy probably had also a political motivation.

Restitution of the Holy Roman Empire

Boleslav I (ruled 929–967) reigned as a Christian prince; his daughter married Prince Mieszko I of Poland and helped spread the gospel in that country. Boleslav attempted, unsuccessfully, to loosen the ties with the Saxon dynasty. Boleslav II (ruled 967–999) used friendly relations with the Pope and the Emperor to enhance his prestige. He attached new territories east of Bohemia to his father's annexations. In 973 a bishopric was founded in Prague and subordinated to the Archbishop of Mainz. The first bishop, Thietmar, was from the Saxon land but knew the Slavic language; he was succeeded in 982 by Adalbert (Vojtěch), a member of the Slavník family. Adalbert's promotion can be viewed as an attempt to harmonize relations between the Prague and Slavník princes, but that result did not materialize. Legends hint that Adalbert encountered considerable opposition when attempting to raise the standards of religious life in his diocese, and tension between the rival dynasties showed no signs of abating. In 895 Boleslav II moved against the Slavníks and broke their power. Adalbert enjoyed some initial success among the heathen Prussians on the shores of the Baltic but then suffered a martyr's death in 997.

Annexation of Moravia

Struggles among the descendants of Boleslav II plagued the country for about 30 years and considerably reduced its power. Most of the territories attached to Bohemia in the 10th century were lost. Prince Břetislav I, a grandson of Boleslav II, led a successful expedition into Moravia; he conquered only a minor portion of the former Great Moravia, but it was large enough to constitute a province, and it was linked from then on indissolubly with Bohemia. But the ambitions of Břetislav, who ascended the throne in 1034, ran higher. He invaded Poland in 1039 with only temporary success; he incurred the indignation of the German king Henry III and was forced to evacuate the conquered territory and to make an oath of fealty (1041). In the latter part of his reign, Břetislav cooperated with Henry III, thus protecting his domain against armed intervention.

The entire territory of Bohemia and Moravia was regarded as a patrimony of the House of Přemysl, and no emperor attempted to put a foreign prince of his own choice on the throne. But the ruling family grew large, and after Břetislav's death (1055) it became entangled in competition for primacy. For about 150 years the course of public life in Bohemia was largely determined by dissensions among the adult princes, some of whom ruled in portions of Moravia under Prague suzerainty. The emperors and the landowning magnates exploited the conflicts to promote their selfish interests. A main problem was the absence of any strict law of succession; the principle of seniority usually conflicted with the reigning prince's desire to secure the throne for his oldest son.

The territory's minor obligations toward the emperors was a handicap under weak princes or when the male members of the ruling family were at odds, but a strong prince could turn friendly relations with the empire to his advantage. Břetislav's second son, Vratislav II (ruled 1061–92), as a compensation for services rendered, obtained from Emperor Henry IV the title of King of Bohemia (1085). Another able ruler, Vladislav I, gained the dignity of a cupbearer to the Emperor (1114), one of the highest court offices; and as its holder the prince of Bohemia became one of the electors who chose the Holy Roman Emperor. Vladislav II (ruled 1140–73) participated in the campaigns of Frederick I Barbarossa in Italy. He was named king and crowned by the Emperor at Milan in 1158.

Active participation in imperial policies and military campaigns reduced markedly the Czechs' isolation, caused by Bohemia's geographical position. Other contacts were made with foreign merchants and by clerics from abroad or traveling from Bohemia to Rome and to famous shrines. By the early 11th century the Latin rite prevailed. Cosmas of Prague, who recorded in his chronicle the history of Bohemia to 1125, was an ardent supporter of the Latin liturgy. Western orientation of the hierarchy and of the monastic orders was documented by the prevalence of Romanesque architecture, of which notable examples could be found in Prague and in the residences of appanage princes (lesser members of the ruling family). In social stratification and in economy the country reached such a degree of consolidation that it withstood, without serious damage, struggles that ravaged it in the late 12th century.

Frederick I Barbarossa helped foment discord among Přemysl's descendants. In 1182 he reduced the dependence of Moravia on the Prague princes and subordinated that province to the imperial authority. In 1187 he exempted the Prague bishop, a member of the Přemysl family, from the jurisdiction of the ruling prince and made the bishopric an imperial fief. These decisions had no lasting significance, however, and the Přemysl patrimony survived. The period of trials closed with Frederick's death (1190). Frequent subsequent changes on the imperial throne lessened the danger of intervention. During the same period

the Přemysl family was reduced to one branch, so that the problem of succession lost its pressing importance. In 1198 Přemysl Otakar I received the royal title for himself and his descendants from one of the competitors for the imperial crown. A solemn confirmation occurred in 1212, when Frederick II issued a charter known as the Golden Bull of Sicily, which regulated the relationship between Bohemia and the empire. The king's obligations were reduced to a minimum, but as elector he was able to exercise perceptible influence, ranking first among the temporal members of the college of electors.

The Golden Bull of Frederick I

Under Otakar I and his successors, Bohemia moved from depression to political prominence and economic prosperity. The original socioeconomic structure was giving way to a more developed stratification. The clergy gained independence from temporal lords in 1221. The landowning class, made up of wealthy lords and less-propertied squires, claimed freedom in administering their domains and a more active role in public affairs. In the early 13th century the population of Bohemia and Moravia increased noticeably by immigration from overpopulated areas in Germany. Many of the German-speaking newcomers were encouraged by the kings to found urban communities or to develop mining, especially of silver. The landowning magnates and the ecclesiastical institutions followed the royal example and settled the immigrants on their estates. There thus came into existence an urban middle class that enjoyed valuable privileges, especially the use of German law, and that became only slowly amalgamated with the native population. Apart from the townsfolk, German farmers moved into Bohemia and Moravia and transformed the less attractive border districts into prosperous areas. German immigration continued under Otakar I's successor, Wenceslas I (ruled 1230–53), and reached its peak under Přemysl Otakar II (ruled 1253–78). Bishop Bruno of Olomouc, in cooperation with the King, promoted colonization of large tracts of land in northern Moravia.

Otakar II was a strong and capable ruler who obtained possession of the Austrian lands through marriage and, in 1260, was invited by the nobility of Styria to become their lord. Personal bravery and financial resources facilitated his penetration into other Alpine provinces. Before his opponents could combine forces to check his advance, Přemysl Otakar II had exercised influence in Carinthia as well as in some territories along the Adriatic coast. His expansion aroused the hostility of the kings of Hungary; but even more dangerous was Rudolf, count of Habsburg, who was elected king of the Romans in 1273 and, to get a foothold in central Europe, claimed the Austrian lands as vacant fiefs of the empire. War followed and ended in Otakar's defeat in 1276. Otakar was unwilling to accept the loss of Austria as final, however, and began a new campaign. Not only Rudolf's army but also Hungarian troops moved against the Czech forces, and a group of noblemen, most of them from southern Bohemia, sided with the enemy. Otakar was too weak to resist the unexpected coalition against him, and, on Aug. 26, 1278, he lost, at Dürnkrut on the Marchfeld, both the battle and his life.

Otakar's only son, Wenceslas II (ruled 1278–1305), was too young to take over control immediately. During the period following Otakar's death (remembered as the evil years), Wenceslas was a mere puppet in the hands of ambitious lords; but in 1290 he emancipated himself from the tutelage and ruled with more success than had his father. The country was recovering quickly from both political and economic depression, and it again played an active role in international relations. Instead of resorting to wars, Wenceslas engaged in negotiations and soon achieved success in Upper Silesia. This was a prelude to his penetration into Poland, which culminated in 1300 with his coronation as its king. Diplomatic dexterity and enormous wealth quickly enhanced Wenceslas' prestige. In 1301 he was considered a candidate for the vacant throne of Hungary, but he recommended his son as a candidate instead. Meanwhile, the mines in various parts of the country, particularly at Kutná Hora, yielded so much silver that the King was able to reform the monetary system and issue coins (grossus), which soon circulated within and




Přemyslid expansion under Přemysl Otakar II and his successors (1253–1306).

From *Grosser Historischer Weltatlas*, vol. II, *Mittelalter* (1970); Bayerischer Schulbuch-Verlag, Munich

outside his kingdom. Wenceslas II's acquisitions, however, were lost soon after his death; his son Wenceslas III took over Bohemia but was assassinated on his way to Poland (1306). Thus ended the long rule of the Přemysl family.

The end of the Přemysl dynasty

LATE MIDDLE AGES (1310–1526)

**The Luxembourg dynasty (1310–1437).** After a four-year struggle for the throne, the Bohemian magnates decided for John of Luxembourg, son of Henry VII, the king of the Romans. John, only 14, married Elizabeth (Eliška), the second daughter of Wenceslas II. John confirmed the freedoms that the Bohemian and Moravian nobles had usurped during the interregnum and pledged not to appoint aliens to high offices. But a group of advisers, headed by Archbishop Petr of Aspelt, followed John to Prague and endeavoured to uphold the royal authority. In the resulting conflict a powerful aristocratic faction scored a decisive victory in 1318. Its leader, Jindřich of Lípa, virtually ruled over Bohemia until his death in 1329. John found satisfaction in tournaments and military expeditions, and he attached to Bohemia some adjacent territories; the extension of suzerainty over the Silesian principalities was his most significant achievement. He was assisted late in his reign by his oldest son, Wenceslas, who was brought up at the French royal court, where he changed his name to Charles. Charles endeavoured to raise the prestige of the monarchy but was hindered by John's jealousy and by lack of cooperation among the nobility. In 1346 both John, then blind, and Charles joined the French in an expedition against the English. John fell at Crécy.

John and Charles benefited from friendly relations with the popes at Avignon. In 1344 Clement VI elevated the See of Prague and made Arnošt of Pardubice its first archbishop. Clement VI also promoted the election, in 1346, of Charles as the king of the Romans. In Bohemia, Charles ruled by hereditary right. To raise the prestige of the monarchy, he cooperated with the nobility and

Reign of Charles I

the hierarchy. He made Bohemia the cornerstone of his power and, by a series of charters (1348), settled relations between Bohemia, Moravia, and other portions of his patrimony. He acquired several territories in the vicinity at opportune times by purchase or other peaceful means. At the end of his reign, four incorporated provinces existed in union with Bohemia: Moravia, Silesia, Upper Lusatia, and Lower Lusatia. Charles also confirmed earlier documents defining the position of Bohemia in relation to the empire. In 1355 he was crowned emperor in Rome as Charles IV. After consultation with the electors, Charles issued the Golden Bull, which readjusted the problems of the empire, especially the election of the emperor.

Under Charles, Prague became headquarters of the imperial administration. By the foundation of a new district (*nové město*), Charles facilitated expansion of the city as well as a rapid increase in its population; about 30,000 persons lived there by the latter part of his reign. In 1348 he founded in Prague a university with four traditional divisions (theology, law, medicine, and liberal arts); its members were grouped into four nations (Bohemian, Bavarian, Saxon, and Silesian-Polish). Prague attracted scholars, architects, sculptors, and painters from France, Italy, and German lands; the most distinguished among them was the architect Petr Parléř (died 1399), a native of Swabia. The flourishing of the late Gothic architectural style left a deep mark on both the royal residence and the countryside. Under Charles, Bohemia was spared entanglements in war and reached a high level of prosperity, shared by the upper classes and the peasantry. Charles was anxious to save the power and possessions accumulated since 1346. He succeeded in getting his son Wenceslas crowned king of the Romans in 1376. He also made provisions for dividing the Luxembourg patrimony, with the understanding that its male members would respect Wenceslas as their head. After Charles's death (1378) a smooth transition to Wenceslas' reign appeared to be assured. The country mourned Charles as "the father of the country."

His heir began to rule, without opposition, as Wenceslas IV. Though not without talents, he lacked his father's tenacity and skill in arranging compromise, and in less than a decade the delicate balance between the throne, the nobility, and hierarchy was upset. In a conflict with the church, represented by Archbishop Jan of Jenštein, the King achieved temporary success; the Archbishop resigned and died in Rome (1400). The nobility's dissatisfaction with Wenceslas' regime was serious, mainly over the selection of candidates for high offices, which wealthy families regarded as their domain and to which Wenceslas preferred to appoint gentry or even commoners. The struggle was complicated by the participation of other Luxembourg princes, especially Wenceslas' younger brother Sigismund. The nobles twice captured the King and released him after promises of concessions. But Wenceslas never took his pledges seriously and the conflict continued. Simultaneously with the troubles in Bohemia, discontent with Wenceslas was growing in Germany. In 1400 the opposition closed ranks; the Electors deposed Wenceslas and elected Rupert of the Palatinate as emperor.

Opposition to Wenceslas IV

The turn of the century was a watershed in reform endeavours in Bohemia. The movement arose around 1360 from various causes, one of which was the uneven distribution of the enormous wealth accumulated by the church in a comparatively short time. Moral corruption infected a large percentage of the clergy and spread also among the laity. Prague, with its large number of clerics, suffered more than the countryside. Both the King and the Archbishop showed favours to zealous preachers like Conrad Waldhauser and Jan Milíč of Kroměříž, but exhortations from the pulpit failed to turn the tide. The Great Schism in Western Christendom after 1378 weakened the central authority. Disharmony between Wenceslas and Jan of Jenštein, archbishop of Prague, hindered the application of effective remedies. In the late 14th century the reform movement was centred at Bethlehem Chapel (Betlémská Kaple in Prague); its benefactors stressed preaching in Czech as the main duty of its rector.

Religious reform movement

The second, more dramatic, period of the reform movement began with the appointment in 1402 of Jan Hus to the pulpit at Bethlehem Chapel. A scholar, he combined preaching with academic activities and thus was able to reach the Czech-speaking masses and to group around himself scholars and students dedicated to the idea of reform. The university was split, because foreign members followed the conservative line. Another cause of division was the popularity of John Wycliffe, an English ecclesiastical reformer, among the Czech masters and students. Hus did not follow Wycliffe slavishly but shared with him the conviction that the Western Church had deviated from its original course and was in urgent need of reform. Hus enjoyed the goodwill of Archbishop Zbyněk Zajíc of Hazmburk. But the atmosphere in Prague deteriorated rapidly; the German members of the university allied with Czech conservative prelates, led by Jan Železný ("the Iron"), bishop of Litomyšl. Because Wenceslas favoured the reform party, its opponents pinned hopes on Sigismund, king of Hungary; Wenceslas was childless, and Sigismund had a fair chance of inheriting Bohemia.

In the winter of 1408–09, a strong group of cardinals convened a general council at Pisa, which deposed the two rival popes and elected Alexander V to fill the vacancy. Wenceslas sympathized with the cardinals and invited the university to join him. When the Germans did not respond favourably, he issued, in January 1409 at Kutná Hora (Kuttenberg), a decree reversing the traditional distribution of votes. Thereafter, the three "foreign" nations had one vote and the Bohemian nation had three. The Germans rejected the decree and moved to Leipzig, where some of them unleashed a polemical campaign attributing to Hus more influence on the King than he actually had and depicting him as the chief champion of Wycliffe's ideas. Meanwhile, Alexander issued a bull virtually outlawing Hus's sermons in Bethlehem Chapel and authorizing rigid measures against discussion of Wycliffe's ideas. Hus and his collaborators continued their activities. Neither Wenceslas nor any of the Czech prelates was experienced enough to achieve reconciliation between the church authorities and the reform party, and Bohemia was drawn into a sharp conflict. In 1412 the antipope John XXIII became involved in a war with the King of Naples and offered indulgences for contributions to the papal treasury. When Hus and his friends attacked the questionable practices of papal collectors in Prague, John put Prague under interdict. Hus, hit by the sentence of excommunication, left Prague and moved to the countryside under the protection of benevolent lords.

In 1414 John, acting in harmony with Sigismund, since 1411 the king of the Romans, called a general council to Constance. Hus went there hoping to defend himself against accusations of heresy and disobedience. A safe conduct from Sigismund did not protect him in Constance. Late in November he was imprisoned and was kept there even after John, who had lost control of the council, had fled and been condemned by the cardinals. In the spring of 1415, Hus was called three times before the council to hear charges, supported by depositions of the witnesses and by excerpts from his own writing. The council paid no attention to Hus's protests that many of the charges were exaggerated or false. Hus refused to sign a formula of abjuration; he was then condemned and handed over to temporal authorities for execution. He was burned at the stake on July 6.

Execution of Hus

Some scholars reduced to a small number the points on which Hus had deviated from the official doctrine. But his followers, not interested in doctrinal subtleties, reacted emotionally against the council, Sigismund, and the conservative clergy. A wave of indignations swept over Bohemia and Moravia, and this movement, taking the name Hussite from the martyred leader, grew rapidly. A letter of protest, signed by 452 members of the nobility, was dispatched to Constance in September 1415. The condemnation and burning of Hus's friend Jerome of Prague (May 1416) increased the discontent.

Hus had not evolved a system of doctrine nor had he designated his successor. The most faithful of his disciples, Jakoubek of Stříbro, was not strong enough to keep the movement under his control. Ideological differentiation set

in and resulted in divisions and polemics. The moderate Utraquists were entrenched in Prague; the radicals came mostly from smaller boroughs and the countryside. The Germans in Bohemia and in the incorporated provinces remained faithful to the church, and, thus, the deep-seated ethnic antagonism was accentuated.

Struggle between Sigismund and the Hussites

After the death of Wenceslas IV (1419), political issues gained in importance. The Hussites were resolutely opposed to Sigismund, but the Czech Catholics and the Germans were willing to recognize him. Sigismund, determined to break the Hussite opposition, initiated a period of bitter struggles that lasted more than 10 years. Sigismund had the support of opponents of Hussitism within the kingdom, of many German princes, and of the papacy. Invasions of Bohemia assumed the character of crusades but were pushed back by the Hussites, who pulled together in time of danger.

The moderate Utraquists and the radicals reached agreement on the fundamental articles of their faith. The radicals built themselves a centre, given the biblical name Tábor. The accord, concluded in 1420 in the nation's capital, became known as the Four Articles of Prague; it stressed that: (1) the word of God should be preached freely; (2) the communion should be administered in both kinds to clerics and laymen; (3) worldly possessions of the clergy should be abolished; (4) public sins should be exposed and punished. A wide range of disagreements between Prague and Tábor was left open and often resulted in mutual accusations and embitterment. A third party arose in northeastern Bohemia, around a newly founded centre at Oreb, but it had a much smaller following than either Prague or Tábor.

Meetings were held at which attempts were made to give the country a national government; the most significant was an assembly at Čáslav (June 1421). A regency council was set up, but it lacked sufficient authority; and the virtual master of the country was the leader of the "warriors of God," Jan Žižka. He was originally attached to Tábor, but he became disgusted with the endless disputes of its theologians and left the radical stronghold to organize a military brotherhood in northeastern Bohemia (1422); its members became so devoted to Žižka that after his death (1424) they called themselves Orphans.

Žižka strove tenaciously for two goals—the protection of Bohemia from Sigismund and the suppression of the enemies of the law of God within Bohemia and Moravia. He scored brilliant victories in battles against Sigismund's forces but could not unite the country under his banner. A Catholic minority, stronger in Moravia than in Bohemia, resisted the overtures of the Hussite theologians and Žižka's attacks. After Žižka's death, his heirs, headed by Prokop Holý the Shaven, lost interest in protracted warfare with Catholic lords and undertook instead foraging raids into the German territories bordering on Bohemia. But, whenever a crusade menaced Bohemia, the radical military brotherhoods joined the conservative forces to push back the invader. The last encounter at Domažlice was bloodless; the crusaders fled in panic upon hearing of the Hussite strength.

Meanwhile, a general council of the church met at Basel in 1431 and determined to find a peaceful settlement. At a conference at Cheb (1432), the delegates from Basel and Hussite spokesmen resolved that in controversial matters "the law of God, the practice of Christ, of the apostles and of the primitive church" would be used to determine which party holds the truth. The Hussite envoys reached Basel and opened debate on the cardinal points of their doctrine. It soon became clear, however, that the council was unwilling to abide by the Cheb agreement and that theologians representing the Tábor and Orphan brotherhoods would not acquiesce to a lean compromise. A drastic change occurred in Bohemia in 1434. In a fratricidal battle at Lipany in May, combined Catholic and Utraquist forces defeated the radicals and took over the control of the country. Under the leadership of Jan Rokycana, dealings with the Council of Basel advanced markedly. The final agreement came to be known as the Compacts (Compactata); it followed the Four Articles of Prague but weakened them with subtle clauses (*e.g.*, the council granted the Czechs the communion in both kinds but under vaguely defined conditions). After the promulgation of the Compacts (July 1436), an agreement followed with Sigismund. But he died in 1437, and Bohemia was neither united in religion nor consolidated politically.

The Compacts

Various forces hindered religious pacification. The Catholic clergy refused to respect the Compacts, because they were not sanctioned by the Pope, and would not accept Rokycana as archbishop. The radical parties, although gravely weakened at Lipany, existed as an uncompromising opposition in Rokycana's rear. His bid for recognition was also defied by the right Utraquist wing, which had seized the key positions during Sigismund's brief reign.

**The Hussite preponderance (1437–71).** Sigismund had no son, and the problem of succession caused a split among the nobility, which had been enriched during the revolutionary era by the secularization of church properties and had grown accustomed to the absence of monarchy. The conservatives accepted Sigismund's son-in-law Albert II of Austria, but the more resolute Hussites favoured a Polish candidate. Albert's death in 1439 ushered in another interregnum. In January 1440 an assembly was held to set up provincial administration for Bohemia; its composition demonstrated clearly the steady rise in the importance of the wealthy lords, functioning as the first estates. The lesser nobility, recognized as the second estate, was numerous; although the percentage of Catholics among the lords was rather high, the second estate was predominantly Hussite and conscious of its contributions to the Hussite defense. The upper classes recognized the royal boroughs as the third estate but were more and more reluctant to share power with them. In the January assembly the political alignments were not identical with religious divisions. Some moderate Catholics cooperated with the Utraquist majority, headed by Hynek Ptáček of Pirkštejn; a group of conservative Utraquists joined the Catholic lords, among whom Oldřich of Rožmberk held the primacy. The actual leader of the conservative bloc was Menhart of Hradec, nominally a Utraquist. No one was elected governor of Bohemia. Instead, in the counties into which Bohemia was subdivided, leagues were organized to promote cooperation of local lords, knights, and royal boroughs, irrespective of religious orientation.

The problem of succession became urgent when Albert's widow, Elizabeth, gave birth to a boy, baptized Ladislas and called Ladislas Posthumus. Several foreign princes showed an interest in the throne, but in 1443 the estates recognized Ladislas' claims. As he resided at the court of his guardian the German king Frederick III, the interregnum was extended. Ptáček, who headed the majority, died in 1444, and the party acclaimed George of Poděbrady as its leader. For several years the destinies of Bohemia were determined by efforts of Oldřich of Rožmberk and his allies to obstruct George's endeavours. Apart from political and economic stabilization, George strove for a papal sanction of the Compacts and for the confirmation of Rokycana as archbishop. George realized that Menhart's domination of Prague was a more serious obstacle than Rožmberk's intrigues; in 1448 George attacked and took Prague without bloodshed. Rokycana also entered the city and took over from the archconservatives, the Utraquist (or Lower) consistory. Although Frederick III was of the same religion as Rožmberk, he realized that an alliance with George would improve Ladislas' chances; in 1451 Frederick designated George as governor of Bohemia. From that position of strength, George moved energetically against both the Rožmberk coterie and the remnants of the radicals, entrenched at Tábor.

In October 1453 Ladislas was crowned king in St. Vitus' Cathedral, and George served as his chief adviser. Ladislas had been brought up as a Catholic, and German was his mother tongue. George hoped the King could reestablish Bohemia's connection with the incorporated provinces, especially the populous and rich Silesia. Ladislas died suddenly in November 1457. Several foreign princes competed for the throne, but the estates of Bohemia reaffirmed the elective principle and decided unanimously for George (March 1458).

Election of George of Poděbrady

Although attached to the Utraquist party, George endeav-

oured to rule as a king of "two peoples": the Utraquists and the Catholics; the Czechs and the Germans. He was anxious to be crowned according to the rites prescribed by Charles IV. George's son-in-law, King Matthias of Hungary, sent two bishops to Prague; George took a secret oath in their presence, by which he obliged himself to defend the true faith and to lead his people from errors, sects, and heresies. Because the Compacts were not mentioned, George did not hesitate to make his pledge; since the agreement with the Council of Basel, the Utraquists considered the communion in both kinds as a lawful concession and not a heresy. Because both the election and coronation took place in Prague, George's principal concern was to have his title recognized by the estates of the incorporated provinces. He was mostly successful, but he had to accept the friendly help of papal envoys to get at least a provisional recognition by the Catholic and predominantly German city of Breslau (Wrocław) in Silesia (1459). During the next three years George enhanced his prestige both at home and abroad. Feeling that no lasting peace could be achieved without the speedy settlement of religious issues, George attempted in 1462 to have the Compacts sanctioned by Pope Pius II. Instead of approving the Compacts, the Pope declared them null and void. When informed of the Pope's action, George held a solemn assembly in Prague in August and affirmed his devotion to the communion in both kinds. Although neither the Pope nor the King showed any intention of retreating from his position, armed conflict was not inevitable, and several princes, including Frederick III, were willing to use their influence to arrange a compromise. But the new pope, Paul II, was elected in 1464 and soon adopted an aggressive policy that encouraged George's foes, especially the city of Breslau. A group of Catholic noblemen from Bohemia, headed by Zdeněk of Šternberk, formed a hostile league at Zelená Hora (1465) and entered into negotiations with Breslau and other Catholic centres. Shortly before Christmas 1466 the Pope excommunicated George and released his Catholic subjects from their oath of allegiance. In spring 1467 George's troops attacked the rebel forces. George was, on the whole, successful in desultory campaigns against the castles of the insurgents, but his position became more awkward in the spring of 1468 when Matthias of Hungary brought support to the Czech rebels. The Hungarians invaded Moravia and, by tying down a considerable portion of the royal army, they facilitated rebel successes in other parts of the kingdom. In May 1469 the opposition proclaimed Matthias king of Bohemia. In 1470 George achieved some successes over his rivals, but he was unable to consolidate them because of deteriorating health. He died on March 22, 1471, mourned by both the Utraquists and loyal Catholics.

Rejection of the Compacts by the Pope

**The Jagiellonian kings (1471–1526).** Emperor Frederick III had observed benevolent neutrality. And George had also derived comfort from the friendly disposition of Casimir IV, the Jagiellonian king of Poland. Contacts with the Polish court continued after George's death and resulted, in May 1471, in the election of Casimir's son, known in Bohemia as Vladislas II, as king of Bohemia. Vladislas was supported by George's partisans irrespective of religious affiliation. George's foes adhered to Matthias, who possessed Moravia, Silesia, and the Lusatias. Vladislas' forces were not strong enough to defeat the rival, and an agreement concluded in 1478 enabled Vladislas to consolidate his position in Bohemia but left Matthias in temporary possession of the incorporated provinces. After Matthias' death (1490) Vladislas was elected king of Hungary (as Ulászló II); thus, he was able to reunite the incorporated provinces with Bohemia. Vladislas' successor was his only son, Louis, a sickly boy nine years old at his father's death.

Decline of royal authority

The reign of the two Jagiellonians was marked by a decline of royal authority. Vladislas II had been brought up as a Catholic and made no secret of his dislike of the Utraquist rites. But by his coronation oath he obligated himself to respect the Compacts. As long as Matthias was alive, Vladislas was supported chiefly by the Utraquists. After 1490 he spent more time in Hungary than in Bohemia, as did Louis. In this latter period the Catholic lords attached themselves to the royal court and exercised strong influence on public affairs of Bohemia.

The Jagiellonian era at first appears to have been an unbroken chain of aristocratic feuds and rivalries in which personal ambitions triumphed over patriotic sentiments, but a closer examination reveals brighter spots and concrete examples of constructive cooperation. The king stood aloof, and the Catholic and Utraquist factions of the estates concluded an agreement at Kutná Hora (March 1485) that reaffirmed the Compacts, recognized the existing divisions in Bohemia, and forbade attempts by either party to extend its sphere of influence at the expense of the other. The accord lasted until 1516 but was renewed in 1512 as "of perpetual duration." The Unity of the Czech Brethren, which had come into existence in 1457–58 as a new expression of Hussite rigorism, was not granted legal protection. In 1508 Vladislas II issued a stern decree, ordering persecution of the Unity, but it was not applied too rigidly.

The provincial diet rather than the royal court held primacy under the Jagiellonians, especially when the kings resided at Buda. The lords dominated the diet and were supported by the lesser nobility when attempting to limit royal power or when introducing restrictive measures against the lower classes. Both the mighty lords and the less propertied knights viewed with displeasure the political aspirations of the royal boroughs, their competitors in commerce and production. The diets passed several resolutions to remove the third estate from the positions acquired during the Hussite revolution. Because the boroughs obtained little help from the sovereign and his officers, the nobility encountered little resistance. A land ordinance adopted by the diet in 1500 limited considerably participation of the boroughs in the diet. The boroughs were also hit by several decrees, approved by the diet (especially those of 1487 and 1497), by which the landowners attached the peasantry to their estates. They thus reduced the possibility of migration into the towns and deprived the towns of cheap labour.

The boroughs, prosperous and self-confident, resisted the limitations and sought allies wherever they could be found. They obtained some concessions under Vladislas II, but a general compromise was made by the diet held in 1517 by which the boroughs joined concessions in political and administrative matters and surrendered some of the earlier privileges on which their economic prosperity was based. The higher estates tacitly recognized the right of the royal boroughs to participate in the diet as the third estate but reserved for themselves the positions on the board of provincial officers, including that of the vice chamberlain, who, in the king's name, supervised municipal administration. Although the boroughs gained some reasonable satisfaction, the landowning nobility was permitted to engage in production of articles that were previously the monopoly of the royal boroughs.

Compromise between the nobles and the boroughs

The agreement of 1517 did not end feuds and conflicts among the aristocratic factions and their partisans in the lower classes. In 1522 King Louis II left Hungary for Prague, intending to heighten the royal authority. With the help of loyal lords, in February 1523, he relieved Zdeněk Lev of Rožmitál of the office of supreme burgrave and appointed Karel of Minstrberk, a grandson of George of Poděbrady, to that key position in provincial administration. But, soon after the King's departure, Rožmitál resumed political activity and searched for allies. Religious controversies that flared up soon after Luther's attack on indulgences (October 1517) increased tensions in Bohemia. Rožmitál, posing as a staunch supporter of the old faith, ingratiated himself with the King and regained his office. Louis, fully occupied with Hungarian affairs, was preparing for a campaign against the Turks. Meeting the Sultan's army with inadequate forces, Louis was defeated; he drowned in the marshes near Mohács while retreating from the battlefield (Aug. 29, 1526).

## HABSBURG RULE (1526–1914)

**The Habsburgs to 1848.** Ferdinand I of Habsburg, the husband of Louis's sister Anne, presented his claims to the vacant throne. He made substantial concessions to

the Bohemian magnates and was elected king in October 1526; the coronation took place in February 1527. Ferdinand ruled also in other countries and, beginning in 1531, he assisted his brother, the emperor Charles V, in imperial affairs. After Charles's resignation (1558) Ferdinand was elected emperor. He considered Bohemia his most precious possession.

Early in his reign, Ferdinand was frequently absent; but, when in Bohemia, he endeavoured to dilute his precoronation pledges and curtail the privileges of the estates. He was obliged by the coronation oath to observe the Compacts and to treat the Utraquists as equal with the Catholics. But since 1517 Bohemia had been open to ideas emanating from Wittenberg and other Reformation centres. Lutheranism had adherents among the Utraquists and among the German-speaking inhabitants of Bohemia and Moravia. The Unity of the Czech Brethren resisted successfully repeated attempts at its extermination; although not protected by the Compacts, the Unity increased in numbers and was shielded by sympathetic landowners, some of whom became members. The teachings of radical reformers also had echoes in Ferdinand's domains.

An opportunity to settle controversial problems arose in 1547. During the Schmalkaldic War (1546–47), between the Habsburgs and the Protestant Schmalkaldic League, the estates of Bohemia pursued an inconsistent policy, and, after the Habsburg victory at Mühlberg (April 1547), Ferdinand moved quickly against them. The high nobility and the knights suffered comparatively mild losses. But the royal boroughs virtually lost their political power and were subordinated more rigidly to the royal chamber. Another target of the King's wrath was the Unity; significantly, Ferdinand's vindictive policy did not apply to Moravia, the estates of which were more cooperative during the Schmalkaldic War than those of Bohemia. After 1547 the Unity flourished in Moravia; its members, driven from Bohemia, moved to Moravia or emigrated to Poland.

Ferdinand's moves against the Bohemian estates

The Diet of 1549 approved Ferdinand's request that his firstborn son, Maximilian, be accepted as the future king. Ferdinand also resumed his scheme of religious reunion on the basis of the Compacts, but he soon realized that few Utraquists adhered to that outdated document. The majority, called Neo-Utraquists by modern historians, professed Lutheran tenets as formulated by Martin Luther's associate Philipp Melanchthon. Disheartened by the meagre results of his policy, Ferdinand turned toward the Catholic party to consolidate its organization. He introduced the newly founded and militant Society of Jesus into Bohemia (1556) and obtained from Rome consecration of Antonín Brus of Mohelnice as archbishop (1561). Shortly before his death, Ferdinand succeeded in getting from Pius IV a sanction of the communion in both kinds, but the Pope insisted on so many restrictions that his bull satisfied only the Utraquist extreme right.

Maximilian II (ruled 1564–76) was reluctant to grant free exercise of the Lutheran faith, which the majority of the estates requested in 1571. After several years of futile efforts, the estates adopted a more flexible policy. Both the Czech Neo-Utraquists and the German-speaking Lutherans came together and prepared a summary of their faith, known as the Bohemian Confession; it agreed in the main points with the Augsburg Confession. The Brethren cooperated with the adherents of the Bohemian Confession but preserved both their doctrine and their organization. In 1575 Maximilian II approved the Bohemian Confession, but only orally; it was commonly assumed that his oldest son, Rudolf, who was present in the session, would respect his father's pledge.

The early stage of Rudolf II's long reign (1576–1612) was simply an extension of Maximilian's regime. But in 1583 Rudolf transferred his court from Vienna to Prague, bringing with him the high offices and foreign envoys; the Bohemian capital became once more an imperial residence and a lively political and cultural centre. Rudolf, brought up in Spain, had sympathy only for the Roman Catholic faith. Because the crown possessions were too small to yield adequate income, he depended mostly on the estates, whose majority was Protestant; only the provincial diets had the power to approve increased taxation and to grant subsidies for interminable wars against the Turks. The Catholic party, stronger among the lords than among the lesser nobility and burghers, came under the influence of militant elements, trained in Jesuit schools, and listened attentively to the papal nuncios and Spanish ambassadors. Because of its long antipapal tradition and its political prominence, Bohemia had an important place in the strategy of the Counter-Reformation. The Catholics singled out the Unity as their first target. Although numerically weak, the Brethren exercised a strong influence on Czech religious life and developed lively literary activities (in Rudolf's reign they produced a translation of the Bible from the original languages, which was printed in a hamlet of Kralice on the domains of the lords of Žerotín and which came to be known as the Kralice Bible). The Catholics sought to create a breach between the majority party of the Bohemian Confession and the Unity.

The Counter-Reformation in Bohemia

By a succession of new appointments, Catholic radicals around 1600 occupied the key positions in the provincial administration of Bohemia; their head, Zdeněk Vojtěch of Lobkovice, served as the supreme chancellor and enjoyed Rudolf's confidence. In 1602 Rudolf issued a rigid decree against the Unity, which was enforced not only in the royal boroughs but also on the domains of fervent Catholic lords. The Brethren and also the more resolute adherents of the Bohemian Confession realized that the days of peaceful coexistence were gone. They closed ranks under the leadership of Václav Budovec of Budov, a prominent member of the Unity. Dissatisfaction with Rudolf's regime was growing rapidly in other Habsburg domains. His younger brother, Matthias, made contacts with the Austrian and Hungarian opposition; the Moravian estates, headed by Karel the Elder of Žerotín, joined Matthias. In 1608 rebel forces advanced to Bohemia; Rudolf was unable to resist them, and he made peace and transferred to Matthias the dissatisfied provinces. The Protestant estates of Bohemia used Rudolf's weakness for their purposes. Although reluctant, Rudolf issued, in July 1609, a charter, known as the *Majestát* (Letter of Majesty), that granted freedom of worship to the Catholics and to the party of the Bohemian Confession, with which the Brethren closely cooperated. Some passages of the charter were vague, and so the Protestant and Catholic estates concluded an agreement, stipulating that future conflicts should be settled by negotiation. The Catholic radicals, too weak to upset the agreement, were unwilling to accept the *Majestát* as the final word in religious controversies.

In 1611 Rudolf was deposed, and Matthias was crowned king of Bohemia. Because he was childless, the question of succession was debated both in the court circles and among the estates. In 1617 Matthias presented his nephew Ferdinand of Styria to the Diet of Bohemia as his successor. The resolute faction among the Protestant nobility was caught unprepared and acquiesced in Ferdinand's candidacy; he was accepted and crowned in St. Vitus' Cathedral. Opposition grew quickly to Ferdinand, who was suspected of cooperation with the irreconcilable opponents of the *Majestát.* In the spring of 1618 the Protestant estates decided on an action. Two governors of Bohemia, William Slavata and Jaroslav Martinic, were accused of violation of the *Majestát;* after an improvised trial they were thrown, together with the secretary of the royal council, from the windows of the Royal Chancellery in Hradčany Castle (May 23, 1618) but escaped with minor injuries. The act of violence, usually referred to as the Defenestration of Prague, sparked a rebellion in Bohemia. The estates replaced the board, or royal governors, with 30 directors, who assembled troops for defensive purposes and gained allies in the predominantly Lutheran Silesia and in the Lusatias; the estates of Moravia were reluctant to join. The death of Matthias (March 1619) changed the situation profoundly. The directors refused to admit Ferdinand II into Bohemia. In Moravia the militant Protestant party overthrew the provincial government, elected 30 directors, and made an accord with Bohemia. At a general assembly of representatives of all five provinces, a decision was made to form a federal system. Ferdinand II was deposed, and Frederick V, elector of the Rhine Palatinate, a son-in-law of James I, king of England and Scotland, was offered

The Defenestration of Prague

the crown. He accepted and early in November 1619 was crowned king according to an improvised Protestant rite. Frederick's chances for success were slight; the population of Bohemia, especially the peasantry, was unenthusiastic in its support of the rebellion. Frederick received some financial help from the Netherlands, but German Protestant princes hesitated to become involved in a conflict with the Habsburgs, among whose allies were not only Catholic Bavaria but also Lutheran Saxony. In late summer 1620 Maximilian I of Bavaria coordinated the Catholic forces; although short, the battle on the White Mountain, at the gates of Prague (Nov. 8, 1620), had a decisive effect and delivered Bohemia to Ferdinand II. Frederick and his chief advisers fled from Bohemia. Fighting continued in 1621 at some isolated places and in Moravia, but no one succeeded in pushing back Ferdinand's troops.

In imposing penalties Ferdinand treated Bohemia more harshly than the incorporated provinces. On June 21, 1621, 27 leaders (three lords, seven knights, and 17 burghers) were executed. Landowners who had participated in any manner in the rebellion had much of their property confiscated. The upper estates and the royal boroughs were ruined; they ceased to function as centres of economic and cultural activities. Ferdinand rescinded the *Majestát* and declared his intention to promote the program of re-Catholicization of Bohemia and Moravia. The Society of Jesus, banned in 1618 by the directors, returned triumphantly and acted as the vanguard in the systematic drive against the non-Catholics, including the moderate Utraquists.

Under absolutist rule (1627–1848)

In 1627 Ferdinand II promulgated the Renewed Land Ordinance, a collection of basic laws for Bohemia that remained valid, with some modifications, until 1848; he issued a similar document for Moravia in 1628. Ferdinand settled, in favour of his dynasty, issues that had disturbed Bohemian public life since 1526: the kingdom was declared hereditary in both the male and female branches; the king had the right to appoint supreme officers; in the provincial diet the higher clergy was constituted as the first estate, and all the royal boroughs were represented by one delegate only; the diet lost legislative initiative and could meet only upon the king's authorization to approve his requests for taxes and other financial subsidies; the king could admit foreigners to permanent residence; the use of German besides the traditional Czech was authorized. No faith other than Roman Catholicism was permitted.

Royal decrees pertaining to religion granted the upper classes the right to choose either conversion or emigration. A rather high percentage decided for the latter and settled abroad, mostly in Saxony. Many peasants left the country illegally, especially during the Protestant invasions of Bohemia. The Czechs' most significant representative abroad came to be a scholar, John Amos Comenius (Jan Âmos Komenský). The majority of the population remained in the homeland and gradually converted to the Roman Catholic faith. The Jesuits became the most important force in Czech spiritual life. Their leading school, the Clementinum, was, in 1654, united with the remnants of Charles University. The Jesuits controlled not only higher education but also literary production. With an increasing number of Czech novices, the Society could reach the common people, whose majority spoke only the Czech language.

Conversion of the Czechs to the Roman Catholic faith

The vacated places among the upper social classes were gradually filled by newcomers, most of whom obtained land as a compensation for services rendered to Ferdinand II and his successor, Ferdinand III (ruled 1637–57); some enterprising individuals purchased land in Bohemia either during or after the Thirty Years' War (1618–48). The old families and the newcomers had in common their attachment to the Roman Catholic Church and to the dynasty; they intermarried and became amalgamated over the next several decades. German became the language in which public affairs were transacted. But language was not the only barrier separating the peasantry and lower middle class from the propertied noblemen and burghers. Both the victorious church and the wealthy laymen regarded the Baroque style as the most faithful expression of their religious convictions and their worldly ambitions. For about 100 years, the Baroque dominated in architecture, sculpture, and painting and influenced literature, drama, and music. The external appearance of Prague and the smaller boroughs and towns changed markedly. In the countryside sumptuous aristocratic residences contrasted sharply with the modest dwellings of the peasantry.

Leopold I (ruled 1657–1705) soon became involved in long and costly wars against the Turks and the French. Although Bohemia was not threatened by either of his enemies, its population had to share the financial burdens; the landed nobility was reluctant to accept financial obligation, so the major part of the contributions was expected to come from the burghers and the peasants. The urban communities, which had been impoverished during the Thirty Years' War, made no progress toward social and economic recovery. The lot of the peasantry was so heavy that risings occasionally flared up, though with no chance of success. For the common people the short reign of Joseph I (ruled 1705–11) brought some relief, but under his brother and successor, Charles VI (ruled 1711–40), their plight reached appalling dimensions. The court and the residences of the ranking aristocrats consumed vast sums of money, which had to be squeezed from the depopulated towns and poorly managed domains. At this time, alienation of the masses of people reached its apex.

Retention of autonomy under the Habsburgs

The Habsburgs, ruling over Bohemia from 1620 to 1740, did not insist on its close union with their other domains. The kingdom of Bohemia, though under an absolutist regime, retained its autonomy. The two Lusatias were ceded in 1635 to Saxony; Bohemia, Moravia, and Silesia retained their provincial administration. Members of the local nobility were appointed to high offices. The supreme chancellor of Bohemia served as a link between the kingdom and the sovereign and resided in Vienna to facilitate communication with the court and various central agencies attached to it.

Most of the reforms of Charles's daughter Maria Theresa (ruled 1740–80), although motivated primarily by dynastic interests, improved the living conditions of the population. Soon after her accession Bavaria and Prussia invaded the Habsburg territories. Charles Albert, elector of Bavaria, occupied a major part of Bohemia and was acclaimed king by a fairly strong party among the estates; but he could not establish himself permanently, and in 1742 he pulled his forces back. Three wars fought against Frederick II the Great of Prussia in 1741–63, mostly in Bohemia and Moravia, were more serious and costly; finally, Maria Theresa acquiesced in the loss of the major part of Silesia. Small duchies that she was able to retain were constituted as a crown land of Silesia and remained closely connected with Moravia and Bohemia.

Realizing that the system inherited from Charles VI was the main source of weakness, Maria Theresa launched a program of administrative reforms (1749); its principal point was a closer union of the Bohemian crown land with the Alpine provinces. The Queen's staunchest opponents were members of the landowning nobility who, up to that time, had controlled the provincial administration. In 1763 Maria Theresa made some concessions but would not abandon her centralist policy. The opposition did not remain united. The conservative faction remained unreconciled to the new course, but more flexible individuals accepted high positions in Vienna or in the provincial capitals and helped to build up the system, which Joseph II (ruled 1780–90) inherited from his mother and subordinated more rigidly to the sovereign's will and discretion.

Maria Theresa, partly under the influence of her husband, Francis Stephen of Lorraine, had adopted the idea of curtailing the privileges of the upper social classes so as not to conflict with the interest of the state, of which the ruler was the supreme representative. Joseph II had grown up in this enlightened atmosphere, and, when confronted with conservative opposition as king, he went far beyond his mother's limits. Apart from the administrative reforms, the judicial and fiscal systems were revamped to serve the enlightened monarch more adequately. The state extended its influence in such other fields as education, religion, landowner–tenant relationships, the economic recovery of the royal boroughs, and a more adequate distribution of

the burden of taxes. The reforms did not aim at total abolition of social and economic distinctions, but they generally improved the lot of the lower middle class and of the peasant. Two decrees of 1781 made Joseph popular among the masses: he abolished restrictions on the personal freedom of the peasants, and he granted religious toleration. After the long period of oppression, these were hailed as beacons of light, although they did not go as far as enlightened minds expected. The edict of toleration in Bohemia and Moravia was not followed by a mass defection from the Roman Catholic Church, partly because it did not refer to either the Utraquism or the Unity but authorized worship according to either the Augsburg or Reformed Confession.

Joseph's conservative successors, Leopold II (ruled 1790–92), Francis II (ruled 1792–1835), and Ferdinand V (Ferdinand I of Austria; ruled 1835–48), left intact the centralistic system inherited from Maria Theresa and Joseph II, but they did engineer gradual transition from the manorial system to the full ownership of land by the peasants. They made peace with the landowning nobility, seeing in it their most faithful ally; but the provincial diets of Bohemia and Moravia still had no more than a decorative function. A fairly large number of persons of rank distinguished themselves as patrons of learning, lovers of theatre and music, promoters of new and more profitable methods in agriculture, and, in the early 19th century, as pioneers of industry. In these activities they made contacts with gifted men of middle-class or of peasant origin, gave them financial support, and shielded them from the ubiquitous police and rigid censorship. Provincial loyalties were stronger than ethnic differentiation, which emerged as a new factor in Bohemia around 1800, partly out of opposition to the centralistic tendencies of the Vienna court, partly under the impact of the French Revolution. Institutions destined to play an important role in the Czech national renascence, such as the Royal Bohemia Society of Sciences or the National Museum (1818), were bilingual and drew support both from the propertied German population and from a small fraction of the Czechs who became conscious of their origin, of the brighter periods, and of their kinship with other Slavic peoples.

Czech national renascence

**From absolutism to constitutionalism (1848–1914).** In 1848 the German-speaking population of Bohemia and Moravia had a distinct advantage over the Czechs. The upper classes of these two provinces were almost entirely German and the rural areas in which, after 1620, the Germans gained predominance extended from the mountain ranges deep into the lowlands, once purely Czech. There were, however, limited opportunities for Czechs of middle-class or peasant origin, who prepared for more lucrative occupations through higher studies or who acquired special skills. Some improvement could be observed in the last stage of Habsburg absolutism, from about 1830 onward. The efforts of scholars, writers, clergymen, and schoolmasters, aware of their Czech origin, stirred a national consciousness among the common people. Not only the countryside but also the urban communities witnessed an awakening. The Habsburgs, symbolized by Prince Metternich, tolerated no political activities but did not hinder cultural activities, the printing and distribution of nonpolitical books in Czech, theatrical performances, and gatherings for other than political purposes. The Czechs had their social and intellectual elite, small in numbers but devoted to the national cause, and they were shielded by a group of sympathetic aristocrats.

Similar conditions existed in the Hungarian counties inhabited by the Slovaks. Contacts between these two ethnic groups were hindered by the existence of provincial boundaries, but the groups were close enough to permit cultural exchanges. Up to 1840 the Czech language, regenerated by such eminent linguists as Josef Dobrovský and Josef Jungmann, was used by both Czech and Slovak authors. But the growing national awareness gave rise to endeavours to develop a literary language for the Slovaks in order to reach people with no more than elementary training. The literary Slovak gradually replaced the Czech among Slovak authors. Thus, the mounting wave of nationalism created conditions for differentiation and for the establishment of two closely related but distinct ethnic units.

In opposing Metternich's oppressive regime, the Czech intellectuals were allied with the progressive forces among the Germans. When the revolutionary wave reached Bohemia in March 1848, leaders of the two nationalities worked together in an attempt to shift from absolutism to constitutionalism. Both parties had a vague notion that Bohemia should return to its autonomous status and become a constituent part of the regenerated monarchy, but they could not resolve specific problems. The Germans saw advantages in cooperating with their kinsmen in other Habsburg lands; moreover, they were keenly interested in the idea of German unification, debated in the German constituent assembly at Frankfurt. The Czech voters looked to František Palacký as their leader; he had written several volumes of *A History of the Czech People* and was a respected political thinker. He was assisted by Karel Havlíček Borovský, a journalist, and by František Rieger, a student of political science and economics. The Czech leaders sensed danger in the schemes laid before the Frankfurt assembly and in plans for a modernized but highly centralized Austria. Their primary concern was the Diet of Bohemia, and, at times, they included among their desiderata a general assembly of deputies from Bohemia, Moravia, and Silesia, to stress a continuity of modern political efforts with the ancient kingdom. But, despite some support from the aristocratic circles, the Czechs were unable to change the movement toward centralization.

Revolution of 1848

Movement toward centralization

A good deal of vacillation in and after 1848 was caused by the inability of Palacký and others to harmonize the emphasis of historical rights with genuine devotion to the principle of nationality. In late spring 1848 the idea of an elected diet for Bohemia was obscured by a loftier project, an assembly of spokesmen of the Slavic peoples from all parts of the Habsburg Empire. No matter how sincerely Palacký and other prominent figures professed their loyalty to the ruling house, the Slavic congress was viewed with displeasure by the Germans and the Magyars and was finally dispersed by the archconservative Alfred, Fürst zu Windischgrätz, who ordered that no election for the provincial diet could be held. The Czech leaders recognized that the constituent assembly meeting in July 1848 in Vienna was the only representative body before which they could express their aspirations. They participated in the late summer and early autumn sessions and worked with even more vigour when the assembly reconvened at Kroměříž (Kremsier). They made themselves allies of all factions that attempted to prepare the ground for a constitutional and federal system. Rieger, in particular, rose to the occasion when defending the principle that all power comes from the people. But the draft of a constitution for the Habsburg monarchy ran counter to ideas prevailing among the advisers of the new king Francis Joseph I (ruled 1848–1916). Early in March the Kroměříž assembly was dispersed. The Habsburg government, headed by Felix, Fürst zu Schwarzenberg, ruled for some time in accordance with a constitution drafted by the crown advisers; but, on Dec. 31, 1851, Francis Joseph abolished the last vestiges of constitutionalism and began to rule as absolute master.

The regime, allied with the church and supported by the army, police, and bureaucracy, was rigid and effective but tolerated no opposition. Its weakness was revealed, however, by the poor show of its armies in a war with Sardinia in 1859. In October 1860 Francis Joseph issued a diploma burying the absolutist rule and inaugurating a constitutional era. But it soon became clear that no scheme forwarded by the crown advisers could reconcile the federalist tendencies with the monarch's desire to concentrate as much power as possible in Vienna.

After a war with Prussia and Italy in 1866, Francis Joseph sought a solution that would promise speedy recovery and the stabilization of internal affairs. The monarchy was transformed, in 1867, into a dual system. The Magyars obtained the dominant position in Hungary, where Slovak ethnic identity was suppressed; in the conglomeration of other provinces, which was briefly called Austria, the Germans were the strongest single group, followed by Czechs,

Poles, and other nationalities. The dual system passed through successive crises but remained in existence until 1918.

Like other nationalities the Czechs resumed political activities after the promulgation of the October Diploma. Palacký was recognized as a dominant figure, but the actual leadership passed into Rieger's hands. Two courses were open to the Czechs: to apply the principle of nationality or to emphasize historical continuity. Palacký and Rieger decided for the latter and were supported by their conservative collaborators; clearly they had no chance for success without a close alliance with the conservative landed aristocracy, to which the electoral system granted a strong position in the provincial diets and in the parliament. But this alliance was exploited by Rieger's progressive opponents. Differentiation within the National Party began in 1863 and continued more rapidly after 1867. The Czechs, irrespective of ideological orientation, opposed the dual system and boycotted institutions that Austria received after the promulgation of a new constitution in December 1867. After two stormy years, an attempt was made to devise a solution that would give Bohemia autonomy within the Austrian half of the monarchy. In agreement with the historically minded nobility, Rieger negotiated in 1870 and 1871 with the Vienna cabinet and consented to a compromise. But Francis Joseph, although originally sympathetic, yielded to heavy pressure from many sides in October 1871 and refused to sanction the compromise. No attempt was made after 1871 to revive the project.

Internal differences in National Party

Despite the setback, Rieger was able to retain leadership for some 20 more years. Most official statements in either the Vienna Chamber of Deputies or in the provincial diets of Bohemia and Moravia contained a formal declaration in favour of the state right. The idea of restitution of the kingdom of Bohemia to its former rank, similar to that of Hungary, was never given up; but its chances of realization declined with the consolidation of the dual system, and Francis Joseph showed no intention of going to Prague to be crowned with the ancient crown of St. Wenceslas. After 1871 the Czech political leadership was confronted with a dilemma: whether to boycott the parliament and the diets or to join the government majority for concessions in education and economic life. In 1874 the National Party split; the progressive wing, commonly called the Young Czechs, gained in popularity among the urban middle class and well-to-do peasants. Rieger found it more and more difficult to defend his alliance with the big landowners, because it brought no tangible results and obstructed the flow of progressive ideas. The Young Czech deputies insisted on its dissolution and were applauded by their supporters, to whom progress in education, emancipation from clerical influences, and improvement of living standards were more vital than the continued emphasis on unforfeited state right. The Old Czechs lost ground in the 1880s and suffered a total defeat in the parliamentary election of 1891.

German-Czech rivalry

The most determined opponents of the state right scheme in 1871 and thereafter were the spokesmen of the German-speaking population of Bohemia and Moravia, later known as the Sudeten Germans, who realized the losses they would suffer with any decentralization of Austria. In the Vienna parliament they cooperated with their kinsmen from the Alpine provinces and helped determine the composition of the cabinets. An alliance between Austria-Hungary and Hohenzollern Germany (1879) increased their sense of belonging to one of Europe's strongest ethnic units. But their population was losing in Bohemia and Moravia in proportion to the Czechs. The losses were not spectacular and were largely neutralized by Vienna's reluctance to change the traditional practices of giving preference to German over Czech candidates in civil service and especially in the army. The electoral system for the provincial diet, introduced in 1861, was not changed, although the right to vote in parliamentary elections was extended several times to benefit less propertied voters. The immediate cause of Rieger's fall was dissatisfaction over concessions he was willing to make to the Germans in 1890. Thereafter, no attempt was made to achieve general agreement on problems of coexistence of the two ethnic blocs. The largest and richest crown land, in fact, became a trouble spot second, after 1908, only to the southern Slavic provinces.

But the Young Czech leaders were soon caught in the same dilemma that had plagued Rieger. Solemn declarations of adherence to the state right scheme were followed by bargaining with the prime ministers, who sought potential members of a government coalition and offered tempting concessions, including Cabinet posts. Graf Kazimierz Badeni, who headed the Austrian Cabinet in 1895–97, promised administrative measures that would sanction wider use of Czech in Bohemian civil service and law courts. But he encountered vigorous opposition, organized by German nationalists, in the parliament and lost the Emperor's confidence. He resigned, and his successor recognized the futility of trying to adjust the outdated laws in favour of the Czechs.

Decline of the Young Czechs

The changing social and economic stratification also sped the decline of the Young Czechs. They unsuccessfully courted industrial workers, who were more attracted by the Social Democrats and voted for their candidates. Václav Klofáč, a talented journalist, after several years of cooperation with the Young Czechs, founded the National Socialist Party. The peasants, dissatisfied with the increasing influence of big business and the upper middle class, turned away from the Young Czechs after 1890. An agrarian movement soon became the Young Czechs' most dangerous rival, because the peasants predominated in the Czech-speaking areas of Bohemia and Moravia. The Young Czech political program was pervaded by liberal principles, which included anticlericalism; that made it unpalatable to the conservative groups, which favoured close cooperation with the Roman Catholic Church and which were stronger in Moravia than in Bohemia. Finally, voters led in Moravia by Adolf Stránský and in both provinces by Tomáš Garrigue Masaryk came to feel that the Young Czechs were not seriously carrying out the progressive ideas included in their program. Parties that developed out of ideological opposition were small when compared with the Agrarians, the Socialists, and the Young Czechs; but their ideas reached the noncommitted voters. The grant of universal manhood suffrage in 1906 greatly improved the chances of parties appealing to the less propertied voters; instead of helping to consolidate the parliament, it caused such differences that the prime ministers, following each other in quick succession, found it increasingly difficult to form a solid majority block. Thus, from the election in 1907 to the outbreak of World War I in 1914, the Chamber of Deputies could easily be bypassed by the court and by the ministries of foreign affairs and war, over which Francis Joseph exercised strong control. The dual monarchy was moving toward more dangerous involvements in international affairs and, finally, toward catastrophe.

(O.O./Z.A.B.Z.)

## THE REPUBLIC OF CZECHOSLOVAKIA (1918–45)

**Struggle for independence (1914–18).** World War I increased the estrangement between the Germans and the Czechs within the Czech Lands. The Germans lent full support to the war effort of the Central Powers, but among the Czechs the war was unpopular. Opposition to the war, however, was uncoordinated, because Czech political leaders were unable to agree on a program. In December 1914 Masaryk, a representative in the Vienna parliament, left Prague to organize activities that could not be developed at home because of the suspension of civil rights and political persecution. After staying some months in neutral countries, Masaryk moved to London. In 1915 he had been joined in Switzerland by his former student Edvard Beneš and by Josef Dürich, a member of the conservative Czech Agrarian Party. Masaryk at first had rather vague notions of the tasks ahead of him, but he eventually opted for a program of political union of the Czechs and Slovaks. A young Slovak astronomer, Milan Rastislav Štefánik, offered his support. Masaryk established contacts with the Czechs and Slovaks living in Allied and neutral countries, especially the United States. In 1916 a Czechoslovak National Council was created under Masaryk's chairmanship. Its members were anxious

to maintain contacts with the leaders at home in order to avoid disharmony, and an underground organization called the Maffia served as a liaison between them.

At home the influence of the military increased. The press was heavily censored, public meetings were forbidden, and those suspected of disloyalty were imprisoned. Among those arrested were the pro-Russian Young Czech leader Karel Kramář and the economist Alois Rašín. Dissatisfaction among the Czech soldiers on the Eastern Front became more articulate in 1915, and whole units often went over to the Russian side.

Moderate reform under Charles I

Francis Joseph died in November 1916 and was succeeded by Charles I. The new emperor called the parliament to session in Vienna and granted amnesty to political prisoners. Charles's reforms, although in many respects gratifying, called for more intensive activities abroad in order to convince the Allied leaders that partial concessions to the Czechs were inadequate to the problems of postwar reconstruction. The position of the Slovaks was not getting better, and the Hungarian government showed no inclination to reorganize the kingdom in accordance with the principle of nationality. Two major events coincided with Charles's new course in home affairs and with his discreet exploration of the chances of a separate peace: the Russian Revolution (March 1917) and the U.S. declaration of war on Germany. In May 1917 Masaryk left London for Russia to speed up organization of a Czechoslovak army. While small units of volunteers had been formed in the Allied countries during the early part of the war, thousands of prisoners of war were now released from Russian camps and trained for service on the Allied side. A Czechoslovak brigade participated in the last Russian offensive and distinguished itself at Zborov in July 1917.

U.S. encouragement

From the United States came moral encouragement, but Pres. Woodrow Wilson's early statements pertaining to the peace aims were rather hazy. Several weeks after the U.S. declaration of war on Austria-Hungary, President Wilson promulgated his celebrated Fourteen Points (Jan. 8, 1918), the 10th of which called for "the freest opportunity of the autonomous development" for the peoples of Austria-Hungary.

After the Bolshevik Revolution, Czechoslovak troops became involved in struggles between the Bolsheviks and the Conservative forces for the control of the Siberian railroad. Their achievements, noticed favourably by the Western governments and press, gave the Czechoslovak cause wide publicity and helped its leaders to gain official recognition. Masaryk left Russia for the United States, where, in May 1918, he gained solid support from Czech and Slovak organizations. A declaration favouring political union of the Czechs and Slovaks was issued at Pittsburgh, Pa., on May 31, 1918 (called the Pittsburgh Convention).

In 1918, dealings with the Allies progressed more successfully. Not only the Siberian campaigns but also increased activities at home were used to get the struggle for independence endorsed by the Allied governments. A demand for a sovereign state "within the historic frontiers of the Bohemian lands and of Slovakia" was made in Prague at the Epiphany Convention (Jan. 6, 1918) and repeated later with more vigour. In May 1918 not only the Czechs but also the Slovaks made statements to which Masaryk and his collaborators could point when pressing for an official recognition. The anti-Austria resolution, adopted at the Congress of Oppressed Nationalities at Rome (April 1918), helped in disarming conservative circles in the Allied countries who opposed a total reorganization of the Danubian region. After several encouraging statements came the recognition by France of the Czechoslovak National Council as the supreme body controlling Czechoslovak national interests; the other Allies soon followed the French initiative. On Sept. 28, 1918, Beneš signed a treaty whereby France agreed to support the Czechoslovak program in the postwar peace conference. To preclude a retreat from the earlier Allied declarations, the National Council constituted itself as a provisional government (Oct. 14, 1918). On October 18 Masaryk and Beneš issued a declaration of independence simultaneously in Washington, D.C., and Paris. Events were moving rapidly toward total collapse of the Habsburg monarchy. The last attempt to avert it, the manifesto issued by Emperor Charles on Oct. 16, 1918, brought no positive results. After that, Vienna had no choice but to accept President Wilson's terms. The surrender note, signed by Count Gyula Andrássy, the last foreign minister, was accepted as a sanction of the idea of independence. The Prague National Committee proclaimed a republic on October 28, and, two days later, the Slovak National Council at Turčiansky Svätý Martin acceded to the Prague proclamation.

**Establishment of Czechoslovakia (1918–25).** Despite all efforts to maintain contacts between the leaders abroad and those at home, the early years of the republic were hindered by differences of opinion and occasional frictions. Masaryk returned to Prague on Dec. 21, 1918. Beneš stayed in Paris and was joined by Karel Kramář, prime minister since November 1918. The Slovak leader Štefánik decided to return home but died in an air accident in May 1919. Masaryk and Beneš conducted external relations, and the leaders of five major parties controlled home affairs.

Boundary settlements after war

Of the many tasks facing the new government, negotiations at the peace conference, though complicated by dissensions among the great powers, were the least onerous. The frontiers separating Bohemia and Moravia from Germany and Austria were approved, with minor rectifications, in favour of the republic. The Slovak boundary was also felt to be satisfactory. The dispute over the Duchy of Teschen strained the relations with Poland; the partition of the duchy in 1920 was opposed by powerful Polish groups, and the Polish senate did not ratify the treaty. The northeastern counties of prewar Hungary (Carpathian Ruthenia) were attached to the new state. The area was inhabited by Slavic peoples, the majority of whom were keenly aware of their kinship with the Ukrainians.

Consolidation of internal affairs proceeded slowly. The winter of 1918–19 was critical. The most urgent task of the new government was to replace the wartime economy by a new system. The network of railroads and highways had to be adjusted to the new shape of the republic, which stretched from the Cheb (Eger) region in western Bohemia to the Carpathians in the east. The new country's first minister of finance, Alois Rašín, saved the Czechoslovak currency from catastrophic inflation, and his death in February 1923, after he was shot by a young revolutionary, was a shock to the new republic.

In the chaotic conditions prevailing in central Europe after the armistice, a parliamentary election appeared to be impossible. The Czech and Slovak leaders agreed on the composition of the National Assembly. The Assembly's main function was the drafting of a constitution. The new, democratic constitution was adopted on Feb. 29, 1920, and was modeled largely after that of the French Third Republic. Supreme power was vested in a bicameral National Assembly. The Chamber of Deputies and the Senate had the right to elect, in a joint session, the president of the republic for a term of seven years. The Cabinet was made responsible to the Assembly. Fundamental rights of the citizens, irrespective of ethnic origin, religion, and social status, were defined generously. Some parties, however, saw a contradiction between the constitutional guarantee of equal rights for all citizens and the intention to create a state of the Czechoslovaks.

Opposition from Sudeten Germans

Large segments of the population gave wholehearted support to the republic; the most resolute opposition, however, came from an ethnic minority that soon came to be known as the Sudeten Germans. The age-old antagonism between Germans and Slavs, accentuated during the war, prevented cooperation during the opening stages of the republic. The Germans issued protests against the constitution but participated, nevertheless, in parliamentary and other elections. In 1925 two German parties—the Agrarian and Christian Socialist—joined the government majority, thus breaking the deadlock. Disagreement with the trend toward centralism was the main source of dissatisfaction among the Slovak Populists, a clerical party headed by Msgr. Andrej Hlinka. Calls for Slovak autonomy were counterbalanced by other parties seeking closer contacts with the corresponding Czech groups; the most significant contribution to that effort was made

by the Agrarians under Milan Hodža and by the Social Democrats under Ivan Dérer. The strongest single party in the opening period, the Social Democracy, was split in 1920 by internal struggles; in 1921 its left wing constituted itself as the Czechoslovak section of the Comintern. After the separation of the Communists, the Social Democracy yielded primacy to the Agrarians. The Republicans, as the peasant party was called officially, became the backbone of government coalitions until the disruption of the republic; from its ranks came Antonín Švehla (prime minister 1921–29) and his successors.

**Political consolidation (1925–35).** Foreign relations were largely determined by wartime agreements. Czechoslovakia adhered loyally to the League of Nations. Treaties with Yugoslavia and Romania gave rise to the Little Entente. France was the only major power that concluded an alliance with Czechoslovakia (January 1924). Relations with Italy, originally friendly, deteriorated after Benito Mussolini's advent to power in 1922. Czech anticlerical feeling precluded negotiation of a concordat with the papacy until 1928, when an agreement was worked out providing for settlement of the most serious disputes between church and state. But it was Germany that most strongly influenced the course of Czechoslovak foreign affairs. The relations between the two neighbours, correct but cool, improved slightly in 1925 after the Locarno Pact, a series of agreements among the powers of western Europe to guarantee peace. In the milder climate of the late 1920s, a third party, the Social Democrats, joined the German activists. Attempts to change the attitude of the Slovak Populists met with partial success. Reorganization of public administration in 1927, while marking a retreat from rigid centralism, did not go far enough to meet demands for Slovak autonomy. Monsignor Hlinka and his chief collaborator, Msgr. Josef Tiso, tenaciously pursued the program of decentralization and only at short intervals supported the Prague government.

Impact of the Great Depression

When the impact of the Great Depression reached Czechoslovakia, soon after 1930, the highly industrialized German-speaking districts were hit more severely than the predominantly agricultural lowlands. The ground was thus prepared for the rise of militant nationalism. Parties supported by middle-class German voters and persisting in opposition to Prague gained in popularity and were encouraged by Adolf Hitler's rise to power in Germany. In October 1933 Konrad Henlein, a supporter of Hitler and head of the politically active Sudeten Turnverband gymnastics society, launched his Sudeten German Home Front. Professing loyalty to the democratic system, he asked for recognition of the German minority as an autonomous body. In 1935 Henlein changed the name of his movement to the Sudeten German Party so as to enable its active participation in the parliamentary election (May 1935). The party captured nearly two-thirds of the Sudeten German vote and became a political force second only to the Czech Agrarians.

**Moving toward the abyss (1935–38).** A tense interlude of little more than two years followed the landslide victory of the Sudeten Germans. In December 1935 Masaryk retired from the presidency, and Beneš was elected his successor by an overwhelming majority, including Hlinka's party. A treaty with the Soviet Union in 1935 enhanced the sense of national security. The program of the Communist Party was determined not only by this treaty but also by the general reorientation of the Comintern, which now urged cooperation with anti-Fascist forces in popular fronts. The Czechoslovak Communists did not, however, seek Cabinet posts. The erection of fortifications along the German frontier modeled on France's Maginot Line was commonly interpreted as an unwritten pledge of the French army to aid Czechoslovakia in the event of an unprovoked attack. Their capture would have given (and did give) the Germans the key to the French defense system. In February 1937 Prime Minister Milan Hodža made significant progress toward gaining the cooperation of those segments of the German population that were attached to the principles of democracy. The hope that Czechoslovakia would be able to withstand pressure from the Third Reich seemed, for a while, to be justified.

Hitler and eastward expansion of Germany

But, soon after the death of Masaryk, in September 1937, Hitler embarked on his program of eastward expansion. As early as November 1937, he informed his military chiefs of his intention to move against Austria and Czechoslovakia. After the annexation of Austria in March 1938, the Czechoslovak crisis became acute.

The Czechoslovak leaders divided their energies. Prime Minister Hodža devoted all his talents to a search for a compromise that would satisfy the Sudeten Germans and held long conferences with Henlein's lieutenants. President Beneš, assisted by his foreign minister, Kamil Krofta, maintained contacts with foreign powers. Henlein played his hand so skillfully that the influential circles, especially in London, believed that he was a free agent and not Hitler's stooge. The advocates of "appeasement," then rapidly gaining ground in Great Britain and in France, failed to realize that the Sudeten German negotiators had no intention of compromise and acted on instructions from Berlin. The main task of Henlein's party was to give Hitler a better chance to dislocate the republic without recourse to war. To invalidate critical comments from London and Paris, Beneš consented late in July to the mission of Lord Runciman, whose avowed purpose was to observe and report on conditions within the country.

The crisis culminated in September 1938. Armed with information supplied by Lord Runciman, British Prime Minister Neville Chamberlain visited Hitler at Berchtesgaden and Godesberg. Chamberlain assured Hitler that the German objectives could be achieved without fighting. The French consented to Chamberlain's policy, thus abandoning their former commitments. The Soviet Union was under no treaty obligation to assist Czechoslovakia, since the treaty of 1935 was to be operative only if the French would honour their pledges. Thus, the stage was set for a meeting between Hitler, Mussolini, Chamberlain, and Edouard Daladier, at Munich on Sept. 29, 1938. They agreed on a document enjoining the Prague government to cede to the Third Reich all districts of Bohemia and Moravia with 50 percent or more of Germans; October 10 was set as the deadline for the transfer of these territories. Although presented as a measure to make Czechoslovakia more homogeneous and viable, the pact and its ruthless implementation sealed the fate of the country.

**From Munich to the disruption of the republic (1938–39).** Beneš resigned the presidency rather than agree to the German annexation. After several weeks he left Prague, first for London and then for Chicago. The leaders who took over had to face mounting difficulties. The annexations completed according to the Munich timetable were not Czechoslovakia's only territorial losses. Poland obtained the Duchy of Teschen as a reward for its menacing attitude during the Munich crisis. By the Vienna Award (Nov. 2, 1938) Hungary was granted large portions of Slovak and Ruthenian territories. By all these amputations Czechoslovakia lost about one-third of its population and the country was rendered defenseless.

The chances of recuperation were greatly reduced by the rapid growth of centrifugal tendencies. The Slovak Populists, headed since Hlinka's death by Monsignor Tiso, presented Prague with urgent demands for autonomy, which the government accepted. A similar request came from Carpathian Ruthenia. A cumbersome system composed of three autonomous units (the Czech Lands, Slovakia, and Ruthenia) united by allegiance to the Prague government was introduced late in the fall. On Nov. 30, 1938, Emil Hácha was elected president; an Agrarian leader, Rudolf Beran, formed the federal Cabinet. Under German pressure the complicated party system was changed drastically. The right and centre parties in the Czech Lands formed the Party of National Unity; the Socialists organized the Party of Labour. In Slovakia the Populists absorbed all other political groups. Despite all efforts of the loyal elements, stabilization of political and economic life made little progress. Moreover, the public knew little of the confidential negotiations being conducted in Vienna and Berlin by Tiso's aides, who went along with Hitler's preparation for the final takeover. In early 1939, Tiso's group prepared for the secession of Slovakia, and, on March 14, 1939, the Slovak National Assembly voted for independence. On

the following day, Bohemia and Moravia were occupied and proclaimed a protectorate of the Third Reich.

**Struggles at home and abroad (1939–45).** The basic laws regulating the status of Bohemia and Moravia were drafted hastily, and many loopholes were left in them to facilitate German intervention. Hitler installed a Reich protector, Konstantin von Neurath, in Prague as his personal representative. The Cabinet under President Hácha operated with limited rights and powers. For some two years the protectorate kept the semblance of an autonomous body. But in September 1941 Reinhard Heydrich replaced Neurath as Reich protector and inaugurated a reign of terror. After Heydrich's assassination (May 1942) the Germans virtually took over the country. Hácha stayed on as president, but the Cabinet was reconstructed in such a manner that it served only as a screen behind which the Germans carried out retaliatory measures and exploited the country's economy for their own purposes. Mass executions, consignment of Czech patriots to concentration camps, and recruitment of young people for work in Germany or behind the front continued until the collapse of the Nazi regime.

Reich protector n Prague

Several months after the proclamation of the protectorate, Beneš moved from Chicago to London to resume his political activities. His position was originally rather awkward, as neither French nor British statesmen wanted to deal with him. But, after the fall of France, in spring 1940, the British prime minister, Winston Churchill, granted Beneš recognition; a provisional government, with Msgr. Jan Šrámek as prime minister, began to function in London. In July 1941 Great Britain and the Soviet Union granted Beneš and his government-in-exile full recognition. Beneš's main occupation was with diplomacy. He devoted a good deal of energy to get the Munich Agreement denounced as invalid. While London and Washington were reluctant to make statements that might prejudice the outcome of the future peace conference, Moscow did not hesitate to condemn the past and open bright prospects for cooperation in the war and in the postwar reconstruction. Beneš visited Moscow in December 1943 and signed a treaty of alliance for 20 years, the terms of which far exceeded the pact of 1935. Not only the treaty but also conversations with Klement Gottwald, the leader of the Czechoslovak Communists, from then on determined both the policy of the exiles and the underground movement in the protectorate and in Slovakia.

Commuist role in he underground

The Communist groups gradually took over the leadership from other clandestine organizations. It was of decisive importance that the Red Army penetrated deep into the territory of the republic several months before the Western Allies were able to cross the traditional borderline between Germany and Bohemia. In March 1945 Beneš and other political figures journeyed from London to Moscow to make a final accord with Stalin and Gottwald. A program of postwar reconstruction was worked out under decisive Communist influence; Zdeněk Fierlinger became prime minister.

The new government, set up at Košice in Slovakia on April 3, 1945, exercised jurisdiction in the eastern portion of the republic; in Moravia and Bohemia, fighting continued until early May. Underground activities, guided by the Czech National Committee, were intensified. On May 5 the people of Prague launched an uprising against the German troops concentrated in central Bohemia and fought them bravely for four days. Their appeals for Allied help were largely ignored. The U.S. general George S. Patton, though sympathetic, did not move from Plzeň, complying with instructions from Gen. Dwight D. Eisenhower. On May 9 the forces of the Soviet marshal Ivan Konev entered Prague. (O.O./Z.A.B.Z.)

## CZECHOSLOVAKIA SINCE 1945

**Provisional regime (1945–48).** President Beneš returned to Prague from his seven years' exile on May 18, 1945, with the intention of restoring in Czechoslovakia the liberal democratic regime that he had been forced to abandon in 1938. It would not be an exact replica but an "improved" version adapted to the new circumstances. The problem of minorities was resolved by large-scale expulsions of the Germans and Hungarians from Czechoslovakia. The country was to remain a republic whose president would retain considerable constitutional and executive power; a government based on the electoral performance of the political parties would run the country by means of a professional civil service, while the judiciary would enforce laws passed by parliament (the National Assembly). In his search for improvement, Beneš decided to limit the number of political parties to six; subsequently two additional parties were permitted in Slovakia, but too late for the election in 1946. In the autumn of 1945 Beneš nominated a Provisional National Assembly, which reelected him president and confirmed in office the government that he had appointed in April. Its premier, Fierlinger, was a Social Democrat. The vice premier was Gottwald, and the leaders of all the other political parties also held vice premierships. A general election was to be held to legitimize the provisional regime as well as to test the nation's acceptance of this new order, in compliance with the agreement of the Great Powers at Yalta.

Six political parties

On May 26, 1946, the Communist Party of Czechoslovakia won a great victory in the general election, polling 2,695,293 votes, 38.7 percent of the total. The non-Communist parties were not alarmed, however, since, combined, they had a decisive majority. Gottwald became premier, and the Communists controlled all the key ministries, including interior (Václav Nosek), information (Václav Kopecký), agriculture (Julius Ďuriš), and finance (Julius Dolanský). Foreign affairs were administered by Jan Masaryk, and Gen. Ludvík Svoboda remained minister of defense. Thus the provisional system had been endorsed by an overwhelming majority of the Czechoslovak people; provided the political parties, grouped in a coalition called the National Front, continued to work harmoniously, the provisional regime would be finalized in 1948, when the Constituent Assembly was to produce a constitution and the next general election was to be held.

From the beginning, however, collaboration between the Communists and non-Communists was difficult, and it became more so. While all parties agreed that the program of postwar economic recovery should go on, and a two-year plan was launched to carry it out, they began to differ as to the means to be employed. The non-Communists wanted no further nationalizations or land confiscations, no special taxes, raises in pay for the civil service, and, above all, economic aid from the United States by way of the Marshall Plan. When in 1947 the idea of Marshall Plan aid had to be abandoned because of pressure from the Soviet Union, the Czechoslovak coalition partners realized that long-term cooperation was impossible, and each party thought of how to resolve the conflict in its own favour. Gottwald and the Communists, while mistrusting their coalition partners, still thought in purely Czechoslovak terms when they announced their strategy: gain an absolute majority (51 percent) in the next election. They organized their party in such a way as to achieve this aim and more: they wanted to be able to mobilize their membership at any given time to exert pressure within the system. Their opponents were disunited, with no common tactics or organization; in common they had only a desire to defeat the Communists within the system. After blocking Communist policies within the government throughout 1947, they were anxiously waiting for the coming election to defeat the Communists decisively.

Marshall Plan aid

The crisis between the two factions came over the question of who was to control the police during the elections. On Feb. 13, 1948, a majority at a Cabinet meeting adopted a resolution ordering the minister of the interior (a Communist) to stop the practice of packing the police force with Communists. The minister ignored the instruction and was supported by Gottwald. On February 20 most of the non-Communist ministers resigned, hoping to force Gottwald to resign as well. He did not. Instead, Communists seized the ministries held by resigning ministers and the headquarters of the parties now in opposition. Mass demonstrations of Communist-led workers took place, and columns of workers armed with rifles paraded through the streets of Prague. In the capital and in the provinces, "action committees" of Communists and of men and women

nominated by Communists were set up, and authorities were ordered to cooperate with them.

President Beneš yielded. On February 25 a new government was formed in which the Communists held the key posts, left-wing Social Democrats were well represented, and the other parties were nominally represented by individual members chosen not by the parties themselves but by the Communists. The Provisional National Assembly overwhelmingly endorsed the new government and its program (230 out of 300 voted for Gottwald).

Most of the non-Communist political leaders fled the country; thousands of intellectuals and managers also escaped, in some cases shooting their way out; many ordinary people also went West, to avoid living under Communism. On March 10 the dead body of Jan Masaryk was found beneath a window of the Foreign Ministry.

Death of Jan Masaryk

Overnight the Communist Party had come to be the only organized body left to run the country. More than a million non-Communists joined it to help Gottwald, who, overwhelmed by the power he so suddenly possessed, continued to cherish a dream of the Czechoslovak way to Communism.

**People's democracy (1948–60).** As had happened in the past and was to happen in the future, the Czechs and Slovaks after momentous events became so self-centred that they forgot the world around them. Gottwald also chose to ignore foreign affairs, even the expanding Soviet hegemony in eastern Europe, and plunged into domestic reforms. Industry was now completely nationalized and confiscation of agricultural land further extended. A new constitution was promulgated on May 9; since it was based on the Soviet model, Beneš refused to sanction it and resigned (he died three months later). Under a new electoral law and with a single list of candidates, a general election was held on May 30, and the new National Assembly (Národní shromáždění) elected Gottwald president; his friend Antonín Zápotocký succeeded him as premier. The Communist Party itself was headed by Rudolf Slánský. Throughout the autumn of 1948 the National Assembly, now a pliant tool of the party, passed reform laws, preparing the administrative reorganization and drawing up a five-year plan. But it all proved in vain. Gottwald went on his holiday to the Crimea, where Joseph Stalin told him what should happen in Czechoslovakia.

Stalin's will was subsequently imposed on the country. Czechoslovakia had to adopt the Soviet model of government: the Communist Party substituted itself for the state. Gottwald and his Communists seemed incapable of running the country along Soviet lines and rooting out subversion. Josip Broz Tito's break with Stalin in Yugoslavia prompted Moscow to tighten discipline within the Socialist camp; in autumn 1949 Soviet advisers were sent to Czechoslovakia.

Gottwald had initiated a campaign against the Christian, especially the Roman Catholic, Church in June, interning Catholic archbishops and bishops and isolating the church from Rome. Monasteries and religious orders were dissolved, and a state office for church affairs was set up to bring churches under Communist control. Soviet security advisers helped prepare the trials of the clergy who refused to cooperate with the Communist authorities, and an effort was made to organize a group of collaborationist clergy.

The purges began in 1950, with non-Communists charged with various anti-state activities. In June, Milada Horáková, former member of the National Assembly, and other politicians from the right to the left were tried for espionage, and several, including Horáková, were sentenced to death. But Gottwald was also under pressure to uncover ideological opponents in his own party, whose leaders the Soviet advisers now began to scrutinize. Evidence of "nationalistic deviationism" and "Titoism" was found, and a purge of the Communist Party of Czechoslovakia followed.

In 1950 Vladimír Clementis, the foreign minister, was dismissed from office, as were Gustav Husák, the Slovak regional premier, and several other Slovaks; all were accused of bourgeois nationalism. In February 1951 Clementis, Husák, and several others were arrested, and in December 1952 Clementis was executed. Ten other high party officials, one of whom was the first secretary, Rudolf Slánský, also were killed. The 10—most of whom, including Slánský, were Jewish—were accused of leading an anti-state conspiracy. Altogether, some 180 politicians were killed in these purges, while thousands were held in prisons and concentration camps.

Party purge

In March 1953, a few days after Stalin's death, Gottwald unexpectedly died. Antonín Zápotocký succeeded Gottwald as president, while Viliám Široký became premier. Both wanted to return to a less repressive way of government, but the widespread rioting that followed a monetary reform (which in effect deprived the farmers and better paid workers of all their savings) in May 1953 gave the diehard faction, led by Antonín Novotný, the new first secretary, an excuse to check Zápotocký's and Široký's initiative. Novotný formally became first secretary of the party in September 1953, and in 1954 he and his faction appealed to the Soviet Union to stop the reform attempts. The Czechoslovak leadership was invited to Moscow, and President Zápotocký was told to adhere to "collective leadership," which in practice meant abandoning power to Novotný. Events in Poland and Hungary in 1956 further justified Novotný's caution in the eyes of the Soviet authorities, and in 1957, when Zápotocký died, Novotný was able once again to combine the party secretaryship with the presidency. His faction—mostly mediocre apparatchiks—became supreme and remained so until 1968.

Novotný kept Stalinism alive. Show trials continued until 1955, after which administrative sanctions began to be employed. Terror and administrative sanctions, however, could not solve problems, either in the economy or in cultural life, and the bullying of the Novotný faction resulted only in hopeless "distortions." In 1958 an industrial reform was carried through, but it failed to resolve long-term problems. Under the first three five-year plans, industrial production was much increased, but by the early 1960s stagnation had set in and production began to fall. Cost of production was high, fuel supplies were short, the quality of goods was poor, and absenteeism was widespread. In agriculture the situation was worse: collectivized agriculture produced less in 1960 than had been produced in prewar years. The educational system was reorganized on the Soviet model, and in the arts Socialist Realism became the norm; both were stultified.

Industrial reform

In July 1960, at a conference of the Communist Party of Czechoslovakia, a new constitution was approved, becoming law in the same month. It laid down that "education and all cultural policy are carried out in the spirit of scientific Marxism-Leninism," and it limited personal property to "consumer goods" and "savings acquired through work." The country's official title of "people's republic" was changed to "socialist republic."

**Reform movement (1960–68).** After Novotný had proclaimed his achievements in the new constitution, he experienced nothing but setbacks. His Third Five-Year Plan had to be canceled in the summer of 1962. In agriculture the situation did not improve, and the young generation, raised under the Communist regime, became critical of the low standard of living the system seemed to generate. Novotný, whether he liked it or not, was forced into reforms proposed by new men on the Central Committee.

In September 1964 the government accepted a new set of economic principles put forward by the reformers, prominent among whom was Ota Šik, a professor of economics. The main principle was to move from total planning and centralization to a mixed economy, with managers of enterprises having more control over their management and the efficiency of each enterprise being measured by its "profitability" in terms of the labour and capital invested. Wholesale prices were to be reformed, and in 1966 it was decided to do this in two stages, in 1967 and 1968. Reform in agriculture was approached in 1966, with a cutback in central planning and introduction of marketing principles. To attract Western currency, tourism was to be encouraged, and from June 1964 visitors were offered double the old tourist rate of exchange. Novotný, however, refused to seek credit from the West, and this nullified many reform measures. In agriculture he was cautious about a new plan for cooperatives to be amalgamated into huge

New economic principles

agricultural enterprises, and as a consequence production continued to stagnate. Novotný's timid reforms thus satisfied no one, resolved no serious problems, and brought into existence pressure groups (the "economists") within the party leadership.

Still, the pressure groups would not have been able to overturn him had he not stirred up Slovak nationalism. In 1960 he agreed to the rehabilitation of the Slovaks purged in the 1950s. The new constitution, however, restricted Slovak autonomy further. By 1963, new men moved into power in Slovakia; Karol Bacílek, who was compromised by the purges in the '50s, was replaced as first secretary of the party in Slovakia by Alexander Dubček. When the rehabilitated Slovaks, among whom was Gustav Husák, began to clamour for a federal solution to their problem, Novotný thought of another repression in Slovakia. But the new men, dissatisfied with the loss of autonomy in 1960, turned against him, and they were largely responsible for his downfall. Before this could be achieved, Novotný blundered in the field of international affairs. In October 1964, when the Soviet leaders removed Nikita Khrushchev from power, Novotný protested. Thus when it came to his own deposition, he had no one to appeal to as he had had in 1954 against Zápotocký.

Curiously, the immediate cause of his downfall was unrest in the cultural sphere. Students had been restless throughout the 1960s, and their traditional procession, the Majales, in 1966 was turned into a riot against the regime. Then in 1967, dissatisfied with the conditions in their colleges, they went out into the streets and clashed with the police. In the end, the minister of the interior had to apologize to them (December 15) for police brutality. The intelligentsia in general was more and more insistently questioning why no institutionalized criticism and no official opposition were permitted.

Student riots

It was the writers (who since 1962 had been challenging censorship and the ideology that supported it), however, that were the most immediate cause of Novotný's fall. At their fourth congress, in 1967, many writers refused to conform to the standards of intellectual discipline set by Novotný. He was still able to answer back with sanctions: Jan Beneš was sent to prison for anti-state propaganda; Ludvík Vaculík, Antonín J. Liehm, and Ivan Klíma were expelled from the party; and Jan Procházka was dismissed from the Central Committee, of which he was a candidate member. This repression so disturbed all the opponents of Novotný that they provoked a crisis in the leadership.

During the session of the Central Committee in October 1967 an open clash occurred between Novotný and the Slovaks. Novotný hinted that Dubček and the other Slovak opposition were tainted with bourgeois nationalism and thus sealed his fate as a leader. In addition he spoke to his generals of a *coup de palais* and invited Leonid Brezhnev, first secretary of the Communist Party of the Soviet Union, to Prague to help him. Brezhnev apparently refused to be involved, and Novotný, now deserted, had to face a hostile session on Dec. 19, 1967. After Šik's demand that the presidency be separated from the party office, Novotný offered his resignation as first secretary. This was accepted at the next session, and on Jan. 5, 1968, Novotný himself recommended as his successor his Slovak opponent Alexander Dubček, who was elected unanimously.

**Prague Spring (1968).** Although Alexander Dubček subsequently became the chief reformer, he was not particularly qualified for this role. He was a young Slovak who made his way into politics through the party apparat. People did not expect much from him, and for a month nothing seems to have happened. As the representative of a new generation, Dubček must have felt frustrated with the Novotný type of apparatchiks, and, though he was in power, he did not control the apparat. That is probably why he decided on an experiment rather than the time-honoured purge techniques. His way of getting rid of the old guard was to subject them to the pressure of public opinion. Once he had made this vital decision, many reforms followed. By April the apparat was Dubček's. Several diehards preferred suicide to disgrace, but Novotný and many others resigned only after a hard struggle. There was a new premier, Oldřich Černík, and Šik and Husák became vice premiers in charge, respectively, of reforms of the economy and in Slovakia. Czechoslovakia also had a new president (from March 30), Ludvík Svoboda, who had been purged in the 1950s and since lived in retirement. The Ministry of the Interior was under the control of another purge victim, Josef Pavel. The newly elected Presidium of the Communist Party consisted largely of newcomers, and the Action Program was compiled by young party intellectuals.

Dubček and the apparat

The Action Program, adopted by the Central Committee on April 5, embodied the reform ideas of the several years preceding, and it accepted the connection between economic and political reform. Among its most important points were a new autonomy for the Slovaks (federation); industrial and agricultural reforms, so long overdue; a revised constitution that would guarantee civil rights and liberties; and complete rehabilitation of all citizens whose rights had been infringed in the past. The program also envisaged a strict division of powers: the National Assembly, not the Communist Party, would be in control of the government, which in turn would become a real executive body and not a party body; courts were to become independent and act as arbiters between the legislative and executive branches. The Communist Party would have to justify its leading role by competing freely for supremacy with other forces in elections. This democratization of life in the country would also extend to the party, in which all offices would be elective. Dubček claimed that he was offering "Socialism with a human face."

The effect that all these happenings had on the Czechoslovak public was unprecedented and quite unexpected. With freedom of the press reestablished, there was a revival of interest in alternative forms of political organization. There were even efforts to reestablish the Social Democratic Party, fused forcibly with the Communists in 1948. With the collapse of the official Communist youth movement, youth clubs and the Boy Scout movement were resurrected. The Christian churches became active as unexpectedly as did many long-forgotten societies, national minority associations, and human rights movements.

On June 27 there appeared in *Literární listy* ("Literary Gazette") a document written by Ludvík Vaculík and signed by a large number of people representing all walks of Czechoslovak life. This was the "Two Thousand Words," urging even more rapid progress to real democracy. Dubček—not to mention the Soviet Union and the other Warsaw Pact allies (with the exception of Romania)—became apprehensive, but, though shocked by the proclamation, Dubček remained convinced that he could control the transformation of Czechoslovakia. The Soviet Union began to take a different view. The Czechs and Slovaks failed to comprehend the hostility of the reaction to the "Two Thousand Words," particularly by the Soviet Union, Poland, and East Germany. Dubček declined an invitation to participate in a special meeting of the Warsaw Pact powers, which on July 15 sent him a letter saying that his country was on the verge of counterrevolution and that they considered it their duty to protect it. To the last, Dubček remained confident that he could talk himself out of any difficulties with his Communist neighbours. He accepted an invitation by Brezhnev to a conference at Čierná-nad-Tisou (a small town in Slovakia), where the Soviet Politburo and the Czechoslovak leaders tried to resolve their problems. On Aug. 3, 1968, representatives of the Soviet, East German, Polish, Bulgarian, Hungarian, and Czechoslovak Communist parties met at Bratislava; the communiqué issued after the meeting, while loosely written, gave the impression that pressure would be eased on Czechoslovakia in return for somewhat tighter control over the press.

The "Two Thousand Words"

On the evening of August 20, however, Soviet, East German, Polish, Hungarian, and Bulgarian armed forces invaded the country and occupied it with little opposition. Politically, the invasion was a catastrophe. The Soviet authorities seized Dubček, Černík, and several other leaders and took them to Moscow but failed to produce alternative party and state leaders acceptable to the people. The population reacted against the invasion with passive

Life under occupation

resistance and improvisation (*e.g.*, removing road signs so that the invading troops would get lost). Communications were disrupted, supplies were held up, and the country was almost leaderless, but life went on as if the occupation forces were not there. Even the scheduled 14th Communist Party Congress took place on August 22 and elected a pro-Dubček Central Committee and Presidium—the very thing the invasion had been timed to prevent. The National Assembly (declaring its loyalty to Dubček) continued its plenary sessions. On August 23 President Svoboda, accompanied by Husák, left for Moscow to negotiate a solution. The negotiations were concluded on August 27, and Svoboda, bringing with him Dubček, Černík, and Smrkovský, returned to Prague to tell the Czechs and Slovaks what price they would have to pay for their Socialism with a human face: Soviet troops were going to stay in Czechoslovakia, and the leaders had agreed to tighter controls over political and cultural activities.

The presence of Soviet troops helped the diehards defeat Dubček and the reformers. First of all, the 14th party congress was declared invalid, as required by the Moscow Protocol agreed upon on August 26. Thus the diehards remained in positions of power, and, by using Soviet pressure and divisions among the reformers, they ultimately achieved their victory. In the meantime Czechoslovakia became a federal republic, the Czech Lands (Bohemia and Moravia) and Slovakia becoming the Czech Socialist Republic and the Slovak Socialist Republic, respectively, with national parliaments and governments. The Czech and Slovak people, however, were more impressed by the suicide of Jan Palach—a student who on Jan. 16, 1969, set himself afire in protest at infringements of national independence—than by Dubček's declarations that the revival movement was going on as before the invasion. Gradually Dubček either dismissed his friends and allies or forced them to resign, until he found himself isolated. This slow rise of the diehards and the "realists" (among whom was Husák) culminated on April 17, 1969, in the removal of Dubček as first secretary, after anti-Soviet rioting in Czechoslovakia.

Dubček was replaced by Husák, who promptly declared the Dubček experiments to be finished and proposed a process that he called normalization.

**Husák's policies.** Husák's modest policy of normalization included the dual aim of ending political experiments and concentrating on economic progress. He patiently tried to persuade Soviet leaders that Czechoslovakia was an orthodox member of the Warsaw Pact. He had the constitution amended to embody the newly proclaimed principle of proletarian internationalism and in 1971 went as far as to repudiate the Prague Spring, declaring that "in 1968 Socialism was in danger in Czechoslovakia, and the armed intervention helped to save it." His most cherished aim was to turn the federal arrangements, which came into force on Jan. 1, 1969, into a reality and then concentrate on economic problems. In fact, the implementation of federalism had helped him to get rid of many diehards and supplant them with his own men. In 1970 Premier Černík was finally forced to resign and was succeeded by Husák's Czech rival and probable successor, Lubomír Štrougal. In 1975, when President Svoboda retired because of ill health, Husák once again fused the two most important offices in Czechoslovakia and became president himself, with full Soviet approval.

Federalism

*Problems in the economy.* After the purge of 1969–71, Husák concentrated almost exclusively on economic problems. In the short term, Czechoslovakia did not suffer significantly, even from the disruption caused by the military occupation in 1968. Husák, however, did not permit the industrial and agricultural reforms to be applied and so failed to cure the long-term economic problems. The achievements of the next several five-year economic periods were modest, and by the early 1980s Czechoslovakia was experiencing a serious economic downturn caused by a decline in markets for its products, burdensome terms of trade with several of its supplier nations, and a surplus of outdated machinery and technology.

*Political dissidence.* In the years that he enjoyed power, Husák succeeded in consolidating public life without the bloodletting of his predecessors, but his party purges damaged Czechoslovak cultural and scientific life, since positions in these two areas depended on membership of the party; many writers, composers, journalists, and historians as well as scientists found themselves unemployed and forced to accept menial jobs to earn a living.

Many of these disappointed Communist intellectuals tried to continue the political struggle against Husák, but he eliminated them by proving in court that they had committed criminal acts in pursuance of political objectives. Though these trials could not be compared to the Stalinist trials, they kept discontent among the intellectuals simmering, even if the mass of the population was indifferent. Intellectual discontent erupted again in January 1977, when a group of intellectuals signed a petition, known as Charter 77, in which they aired their grievances against the Husák regime. Many intellectuals and activists who signed the petition were subsequently arrested and detained, but their efforts continued throughout the following decade.

**The collapse of Communism.** The persecuted Charter 77 group played a leading role in the popular upheaval that brought Communist control of Czechoslovakia to an abrupt end in late 1989. On November 17, the authorities allowed a demonstration commemorating the 50th anniversary of the brutal suppression of a student demonstration in German-occupied Prague. The recurrence of police brutality at the anniversary observance set off a protest movement that gained particular strength in the country's industrial centres. Pro-democracy demonstrations and strikes continued nationwide under the leadership of Civic Forum, an opposition group for which dissident playwright and Charter 77 signer Vaclav Havel served as chief spokesman. It was Havel who in late December became Czechoslovakia's first non-Communist president in more than 40 years following the resignation of the Communist government and his election to the office by a parliament still dominated by Communist deputies. In addition, former Party leader Alexander Dubček returned to political life as the new speaker of parliament. In June 1990, in the first free elections held in Czechoslovakia since 1946, the Civic Forum movement and the Slovak counterpart won decisive majorities in both houses of parliament.

Civic Forum

(J.F.N.B./Z.A.B.Z.)

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

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 921, 962, 963, 971, and 972, and the *Index.*

BIBLIOGRAPHY

**Physical and human geography.** *The land and the people:* General descriptive information is available in RICHARD F. NYROP (ed.), *Czechoslovakia: A Country Study,* 2nd ed. (1982); DAVID W. PAUL, *Czechoslovakia: Profile of a Socialist Republic at the Crossroads of Europe* (1981), a brief survey; and two collections of essays: HANS BRISCH and IVAN VOLGYES (eds.), *Czechoslovakia: The Heritage of Ages Past* (1979); and MILOSLAV RECHCIGL, JR. (ed.), *Czechoslovakia Past and Present,* 2 vol. (1968). Works with sections on Czechoslovakia include DEAN S. RUGG, *Eastern Europe* (1985); ANDRÉ BLANC, PIERRE GEORGE, and HENRI SMOTKINE, *Les Républiques socialistes d'Europe centrale,* 2nd ed. (1975); and ROY E.H. MELLOR, *Eastern Europe: A Geography of the COMECON Countries* (1975). Statistical information may be found in *Statistická ročenka Československé socialistické republiky* (annual), including summary sheets of the table of contents in English. Basic geographical information is discussed in JAROMÍR DEMEK *et al., Geography of Czechoslovakia,* trans. from Czech (1971); and VLASTISLAV HÄUFLER, *Ekonomická geografie Československa,* 2nd rev. and enlarged ed. (1984). Useful atlases are JINDŘICH SVOBODA and JOZEF ŠČIPÁK (eds.), *Atlas ČSSR,* 6th ed. (1982); EMIL MAZÚR (ed.), *Atlas Slovenskej socialistickej republiky* (1980); and ČESKOSLOVENSKÁ AKADEMIE VĚD, *Atlas Československé socialistické republiky* (1966). Population information is available in *Demosta* (quarterly), a bulletin of demography and statistics. A comprehensive guide to the country, with city maps and photographs, is CTIBOR RYBÁR, *Československo: Sprievodca* (1980). See also VLADIMÍR HAJKO *et al., Encyklopédia slovenska,* 6 vol. (1977–82), a regional encyclopaedia stressing Slovak and Czech topics, personalities, and events since 1968; and *Příruční slovník naučný,* 4 vol. (1962–67), a concise Czech encyclopaedia. Further bib-

liographic information may be found in DAVID SHORT (comp.), *Czechoslovakia* (1986), a critically annotated bibliography.

*The economy and the government and social conditions:* STÁTNÍ BANKA ČESKOSLOVENSKÁ, *Bulletin* (annual), is a report on the country's yearly economic and monetary development. The history of economic reform proposals from 1948 to 1982 is treated in JOHN N. STEVENS, *Czechoslovakia at the Crossroads: The Economic Dilemmas of Communism in Postwar Czechoslovakia* (1985); while JAROSLAV KREJČI, *Social Change and Stratification in Postwar Czechoslovakia* (1972), is a socioeconomic study of Czechoslovak life in the Communist period. ROBERT CUMMINGS, *A Survey of Czechoslovakia's Agriculture* (1982), is a report from the U.S. Department of Agriculture. See also relevant sections in M.C. KASER (ed.), *The Economic History of Eastern Europe, 1919–1975* (1982– ), 3 vol. of which had appeared by 1987. The history and current practice of education at all levels is outlined in PAVEL JENÍK, *The Czechoslovak Educational System* (1980). KAREL JOSEPH KANSKY, *Urbanization Under Socialism: The Case of Czechoslovakia* (1976), is an urban geography and social history.

*Cultural life:* MILOSLAV RECHCIGL, JR. (ed.), *The Czechoslovak Contribution to World Culture* (1964), is a collection of essays on all aspects of intellectual life, with an extensive bibliography. Specific studies of music and folk art include VLADIMÍR ŠTĚPÁNEK and BOHUMIL KARÁSEK, *An Outline of Czech and Slovak Music*, trans. from Czech, 2 vol. (1960–64); ROSA NEWMARCH, *The Music of Czechoslovakia* (1942, reprinted 1978); and VĚRA HASALOVÁ and JAROSLAV VAJDIŠ, *Folk Art of Czechoslovakia*, trans. from Czech (1974), on the art and architecture of both Slovaks and Czechs.

**History.** Standard works published before the end of the 1960s are found in PAUL L. HORECKY (ed.), *East Central Europe: A Guide to Basic Publications* (1969). No satisfactory full-length history of Czechoslovakia is available in any of the Western languages. The studies by R.W. SETON-WATSON, *A History of the Czechs and Slovaks* (1943, reprinted 1965); S.H. THOMSON, *Czechoslovakia in European History*, 2nd enlarged ed. (1953); and JOZEF LETTRICH, *History of Modern Slovakia* (1955, reissued 1985), remain the standard works on the history up to World War II but are somewhat outdated. A more recent work is WILLIAM V. WALLACE, *Czechoslovakia* (1976).

FRANCIS DVORNIK, *Byzantine Missions Among the Slavs: SS. Constantine-Cyril and Methodius* (1970), illuminates the early medieval period of the region. The kingdom of Bohemia in the 14th and 15th centuries, especially the Hussite movement and its aftermath, is discussed in RUBEN ERNEST WELTSCH, *Archbishop John of Jenstein (1348–1400): Papalism, Humanism and Reform in Pre-Hussite Prague* (1968); HOWARD KAMINSKY, *A History of the Hussite Revolution* (1967); FREDERICK G. HEYMANN, *John Žižka and the Hussite Revolution* (1955, reissued 1969); MATTHEW SPINKA, *John Hus: A Biography* (1968, reprinted 1979); R.R. BETTS, *Essays in Czech History* (1969); OTAKAR ODLOŽILÍK, *The Hussite King: Bohemia in European Affairs, 1440–1471* (1965); FREDERICK G. HEYMANN, *George of Bohemia, King of Heretics* (1965); and JAROLD KNOX ZEMAN, *The Anabaptists and the Czech Brethren in Moravia, 1526–1628: A Study of Origins and Contacts* (1969). PETER BROCK, *The Political and Social Doctrines of the Unity of Czech Brethren in the Fifteenth and Early Sixteenth Centuries* (1957), remains an important work. Subsequent history of Bohemia under Habsburg rule is found in ROBERT J. KERNER, *Bohemia in the Eighteenth Century: A Study in Political, Economic, and Social History, with Special Reference to the Reign of Leopold II, 1790–1792* (1932, reprinted 1969); and R.J.W. EVANS, *The Making of the Habsburg Monarchy, 1550–1700: An Interpretation* (1979, reprinted 1985), and *Rudolf II and His World: A Study in Intellectual History, 1576–1612* (1973, reprinted 1984). See also ROBERT A. KANN and ZDENĚK V. DAVID, *The Peoples of the Eastern Habsburg Lands, 1526–1918* (1984), for an examination of ethnic groups in the Habsburg Empire, including the Slovaks and Czechs.

The development of modern Czech nationalism and of the Czechoslovak state are explored in JOHN F.N. BRADLEY, *Czech Nationalism in the Nineteenth Century* (1984); PETER BROCK and H. GORDON SKILLING (eds.), *The Czech Renascence of the Nineteenth Century* (1970); STANLEY PECH, *The Czech Revolution of 1848* (1969); JOSEPH F. ZACEK, *Palacký: The Historian as Scholar and Nationalist* (1970); BARBARA K. REINFELD, *Karel Havlíček (1821–1856): A National Liberation Leader of the Czech Renascence* (1982); ROMAN SZPORLUK, *The Political Thought of Thomas G. Masaryk* (1981); Z.A.B. ZEMAN, *The Masaryks: The Making of Czechoslovakia* (1976), and *The Break-Up of the Habsburg Empire, 1914–1918: A Study in National and Social Revolution* (1961, reprinted 1977); D. PERMAN, *The Shaping of the Czechoslovak State: Diplomatic History of the Boundaries of Czechoslovakia, 1914–1920* (1962); and VICTOR S. MAMATEY and RADOMÍR LUŽA (eds.), *A History of the Czechoslovak Republic, 1918–1948* (1973).

Interwar Slovakia and Slovak nationalism are covered by R.W. SETON-WATSON (ed.), *Slovakia Then and Now: A Political Survey* (1931), still a valuable classic; PETER BROCK, *The Slovak National Awakening: An Essay in the Intellectual History of East Central Europe* (1976); JOSEPH A. MIKUŠ, *Slovakia: A Political History, 1918–1950*, rev. ed. (1963; originally published in French, 1955); and DOROTHEA H. EL MALLAKH, *The Slovak Autonomy Movement, 1935: A Study in Unrelenting Nationalism* (1979).

The relationship between the Czechs and the Germans is dealt with in ELIZABETH WISKEMANN, *Czechs & Germans: A Study of the Struggle in the Historic Provinces of Bohemia and Moravia* (1938, reissued 1967), a classic work; J.W. BRUEGEL, *Tschechen und Deutsche, 1918–1938* (1967), available also in a substantially altered translation, *Czechoslovakia Before Munich: The German Minority Problem and British Appeasement Policy* (1973); GARY B. COHEN, *The Politics of Ethnic Survival: Germans in Prague, 1861–1914* (1981); and RADOMÍR LUŽA, *The Transfer of the Sudeten Germans: A Study of Czech-German Relations, 1933–1962* (1964). See also HEINRICH RAUCHBERG, *Der nationale Besitzstand in Böhmen*, 3 vol. (1905), which remains the unacknowledged masterpiece on the subject.

Czechoslovakia's fate during World War II is presented in THEODORE PROCHÁZKA, SR., *The Second Republic: The Disintegration of Post-Munich Czechoslovakia, October 1938–March 1939* (1981); VOJTECH MASTNY, *The Czechs Under Nazi Rule: The Failure of National Resistance, 1939–1942* (1971); PETER G. STERCHO, *Diplomacy of Double Morality: Europe's Crossroads in Carpatho-Ukraine, 1919–1939* (1971); and F. NEMEC and V. MOUDRY, *The Soviet Seizure of Subcarpathian Ruthenia* (1955, reprinted 1981).

The period leading up to and including the Soviet invasion of 1968 is covered in Z.A.B. ZEMAN, *Prague Spring* (1969); H. GORDON SKILLING, *Czechoslovakia's Interrupted Revolution* (1976); ZDENEK MLYNÁR, *Nightfrost in Prague: The End of Humane Socialism* (1980; originally published in Czech, 1978); I. WILLIAM ZARTMAN, *Czechoslovakia: Intervention and Impact* (1970); and GALIA GOLAN, *The Czechoslovak Reform Movement: Communism in Crisis, 1962–1968* (1971), and *Reform Rule in Czechoslovakia: The Dubček Era, 1968–1969* (1973). WILLIAM SHAWCROSS, *Dubcek* (1970), is also useful. The important role played by writers and journalists is examined in FRANK L. KAPLAN, *Winter into Spring: The Czechoslovak Press and the Reform Movement, 1963–1968* (1977); and A. FRENCH, *Czech Writers and Politics, 1945–1969* (1982).

Events in the decade after the Soviet intervention are detailed in VLADIMIR V. KUSIN, *From Dubček to Charter 77: A Study of "Normalization" in Czechoslovakia, 1968–1978* (1978). For further study, consult ZDENĚK HEJZLAR and VLADIMIR V. KUSIN, *Czechoslovakia, 1968–1969: Chronology, Bibliography, Annotation* (1975). (F.W.C./Z.A.B.Z.)

# Damascus

Among ancient cities of the world, Damascus is perhaps the oldest continuously inhabited. Its name, Dimashq in Arabic (colloquially ash-Shām, meaning "the northern," as located from Arabia), derives from Dimashka, a word of pre-Semitic etymology, suggesting that the beginnings of Damascus go back to a time before recorded history. Today it is the largest city and capital of the Syrian Arab Republic. Over the centuries, Damascus has been conqueror and conquered, wealthy and destitute, and capital of empire. Its life has been nourished periodically by immigrants from its hinterland and from the Mediterranean Basin and Southwest Asia. Often a focus of contention by powers of East and West, Damascus' fortunes have frequently been linked to those of distant capitals. Now a burgeoning metropolis of the Middle East, it retains, as it has through centuries of triumph and disaster, an indomitable spirit and a not inconsiderable charm.

This article is divided into the following sections:


Physical and human geography 931
- The landscape 931
  - The city site
  - Climate and vegetation
  - The city layout
- The people 931
- The economy 932
  - Industry
  - Commerce
  - Transportation
- Administration and social conditions 932
  - Government
  - Services and health
  - Education
- Cultural life 932

History 932
- Early centuries 932
- The Muslim city 932
- Ottoman period 933
- The modern city 934

Bibliography 934


## Physical and human geography

### THE LANDSCAPE

**The city site.** Water and geography have determined the site and role of Damascus. Early settlers were naturally attracted to a place where a river, the Baradā, rising in the Anti-Lebanon Mountains, al-Jabal ash-Sharqī, watered a large and fertile oasis before vanishing into the desert. This tract, al-Ghūṭah, has supported a substantial population for thousands of years. Damascus itself grew on a terrace, 2,250 feet (690 metres) above sea level, overlooking the Baradā River. The original settlement appears to have been situated in the eastern part of the Old City. City and oasis grew together, and over time Damascus came to dominate the lesser settlements near it.

Advantageous location

The natural endowments of an assured water supply and fertile land made Damascus self-sufficient. Its position on the edge of the desert and at the eastern end of the only easy route through the Anti-Lebanon range also made it a trade centre where caravan routes originated and terminated. Also, since the advent of Islām, Damascus has been the starting point of a pilgrimage road, the Darb al-Hajj, to the Muslim holy cities of Arabia.

**Climate and vegetation.** Some 50 miles (80 kilometres) from the sea, yet separated from it by two mountain ranges, Damascus receives only about seven inches (178 millimetres) of rain annually, most of it from November through February. The Anti-Lebanon range gets far higher amounts of both rain and winter snow, which annually replenish the water table that is a source of the Baradā River. Winter, because of altitude, is rather cold, with average temperatures around 40° to 45° F (5° to 7° C). A short blossoming spring in March and April is followed by six to seven months of hot, dry summer. Temperatures average around 80° F (27° C) in midseason, although occasionally reaching 100° F (38° C) or above. Dust-laden winds blowing in from the desert are somewhat mitigated by small mountain ranges.

Travellers to Damascus have been struck by the sight of aspens and poplars growing along streams, of fruit (particularly apricot) and nut orchards, and of olive groves and vegetable gardens. Ibn Baṭṭūṭah, the Arab travel writer, reaching Damascus in 1326, said that no words could do justice to the city's charm and resorted to quoting his predecessor of more than a century earlier, Ibn Jubayr, that Damascus had "adorned herself with flowers of sweet scented herbs" and "is encircled by gardens as the moon . . . by its halo." A European traveller, Ludolph van Suchem, in 1350 wrote of the city as ". . . begirt with gardens and orchards and watered in and out by waters, rivers, brooks and fountains cunningly arranged to minister to men's luxury . . . ." While the growth of the city since World War II has sharply raised the ratio of buildings to trees and open space, Damascenes still enjoy the parks and gardens of the oasis.

**The city layout.** The heart of the Old City, that part which contains most of the artifacts of its long history, is a rough oblong about 1,640 yards long and 1,100 yards wide, which is defined by historic walls. The long axis of the oblong runs east and west. Many of its most prominent features owe their positions to the city planners of early Hellenistic times and the Roman builders who followed them.

Classical and medieval development

In the 13 centuries following Damascus' capture by Muslim armies, Islāmic urban life and building have largely obscured the classical remains, whose pavements lie some 15 feet below the present street level. Although the population decreased drastically in the early Middle Ages, by the 13th and 14th centuries Damascus had revived and was outgrowing its walls. Two axes of development predominated—one to the northwest linking the city with the suburb of Ṣālḥīyah on the slopes of Jabal Qāsiyūn; the second growing like a long finger to the south.

The modern city follows a plan devised by the French during the mandate period and revised in the 1960s. Along wide boulevards much new housing has developed in the form of concrete blocks of flats. Government buildings are concentrated in an area west of the walled city around Marjah Square and in several districts west of Ṣālḥīyah Street. Stimulated by the appeal of modern housing and amenities, well-to-do families began in the 1930s to move to the area northwest of the Old City and, subsequently, in other directions. As the population grew, more and more of the garden and farm area was converted to residential districts. Farming villages close by were incorporated into the city, administratively and physically. Government efforts to retain green areas and to zone industry have slowed the loss of gardens and orchards.

### THE PEOPLE

Damascus' population of more than 1,200,000 represents a fivefold increase since 1945. It has grown at a rate higher than that of the country as a whole due mainly to migration from rural areas. So heavy has been the influx of migrants drawn by employment and educational opportunity that the average age of the Damascenes has dropped below that of the national level. Among the religious minorities, the ʿAlawites from the Latakia region are notable for their prominence in government. Other groups maintain their identity among the majority Sunnī Muslim

Surging population growth

populace; there are a substantial number of Christians and Palestinians, but the once-flourishing Jewish population has shrunk to a few thousand.

THE ECONOMY

**Industry.** Government is Damascus' most important economic activity. National politics and administration, including a large military establishment, are centred there. Well known over the centuries for luxurious manufactured wares, especially textiles, the growing city with its work force has attracted many new industries since the mid-20th century. All major factories and most industries are state-run. Textile plants, the chemical industry, and cement works are principally distributed to the south, east, and northeast. Traditional artisan crafts are still practiced, and most of the population's requirements for food, clothing, and the like are sold by private businesses.

**Commerce.** The historic role of Damascus as a "desert port" has changed because of political developments and the scale of modern commerce. Most imports come through Syria's own ports of Latakia and Bāniyās instead of through Lebanon, as was the case until about the mid-20th century. Goods are transshipped to countries of the Arabian Peninsula, but trade with Iraq ceased after the borders were closed in 1982. Damascus distributes its own products and imported goods within Syria as well. A large international trade exposition is held there in the autumn.

International trade

**Transportation.** In modern Damascus the internal-combustion engine is superseding horses and donkeys as a means of transport, and camel caravans have vanished from the city scene. Motor vehicles are now the backbone of Damascus' transport system. Buses carry passengers both within the city and to other parts of the country; they are supplemented by the "service"—a car or van that travels an established route for a fare when a full load has been gathered. Major highways fan out in all directions from Damascus, leading to such cities as Beirut, Amman, Baghdad, and Aleppo. A standard-gauge rail line north to Ḥimṣ (Homs), opened in 1983, ties in with the national railroad system; it and the trucking industry bring imported products to the city. Damascus International Airport is about 20 miles east of the city.

ADMINISTRATION AND SOCIAL CONDITIONS

**Government.** The municipality is administered as a *muḥāfaẓah* (governorate), one of 14 in the country. The president of Syria appoints a governor who administers the city with the assistance of a council made up of elected and appointed members. The post of governor of Damascus is an important one that has national implications. Political activity is national, not municipal, Syria being a centralized state with one party dominating public affairs. The outlying portions of the Ghūtah and a vast surrounding district constitute another governorate, Dimashq, of which Damascus city is the capital.

**Services and health.** The growing population has put a strain on the city's services and health facilities. Damascus draws its water from a Baradā River source, receiving it through a centuries-old system that has been enlarged several times. Electricity is generated locally and also is brought from the hydroelectric station at the Euphrates Dam. Health care has been improving and is better than in much of the country. About half of the country's doctors practice in the capital, dividing their services between government hospitals and private clinics. The ratio of hospital beds to population has been rising but is still low compared to more industrialized countries.

Improving health care

**Education.** An extensive public school system provides primary and secondary education for the vast majority of Damascene children. Private schools supplement the public schools, and there is a separate system run by the United Nations for Palestinian refugee children. The University of Damascus, founded in 1923, is the largest and oldest of Syria's four universities, and there are several institutes of technical training.

CULTURAL LIFE

Damascus is Syria's cultural as well as its political centre. Under the Ministry of Culture, which supervises most of the formal aspects of the cultural life of the capital, there is an effort to combine elements of the city's heritage with contemporary developments. The National Museum, charged with preserving the country's past, and the Museum of Popular Arts and Traditions are well attended. An institute for music instructs in both traditional and Western styles; another institute promotes the theatre arts; a third sponsors a performing folklore troupe. The work of Syrian artists and of foreigners is exhibited regularly. Subsidized by the government, however, artistic expression is from time to time impeded by bureaucratic caution. The state dominates publishing, which is centred in Damascus; three national dailies are edited in the city, as are most of the nation's magazines. Damascus also leads the country in book publication, an enterprise that involves the government as the leading publisher.

State regulation of arts

Television enjoys considerable appeal; programming includes locally produced material in addition to imports from other Arab countries and from abroad. Damascus radio broadcasts in Arabic, English, French, Turkish, and other languages. Sports among the Damascenes are growing in popularity. Association football (soccer) especially is becoming a national pastime, and swimming and basketball, along with wrestling, boxing, and tennis, are among other popular recreations. The city's three stadiums draw large crowds for a heavy schedule of events.

## History

EARLY CENTURIES

Excavations in 1950 demonstrated that an urban centre existed in the 4th millennium BC at Tall as-Ṣālḥīyah, southeast of Damascus. Pottery from the 3rd millennium BC has been found in the Old City. Before the 2nd millennium BC an intricate system of irrigation for Damascus and al-Ghūṭah had been developed that was augmented by successive rulers through the centuries. Historically, the first written reference to the city is in the hieroglyphic tablets of Tell el-Amarna in Egypt, where "Dimashqa" is listed among conquered territories in the 15th century BC. Biblical sources refer to it as the capital of the Aramaeans, a Semitic people who have left a legacy in portions of the canal system, place-names, and, in one outlying area, the Aramaic language itself. In succeeding centuries before Christ, it fell like other capitals of the Middle East to foreign conquerors—to Assyrians in the 8th century, Babylonians in the 7th, Persians in the 6th, Greeks in the 4th, and Romans in the 1st.

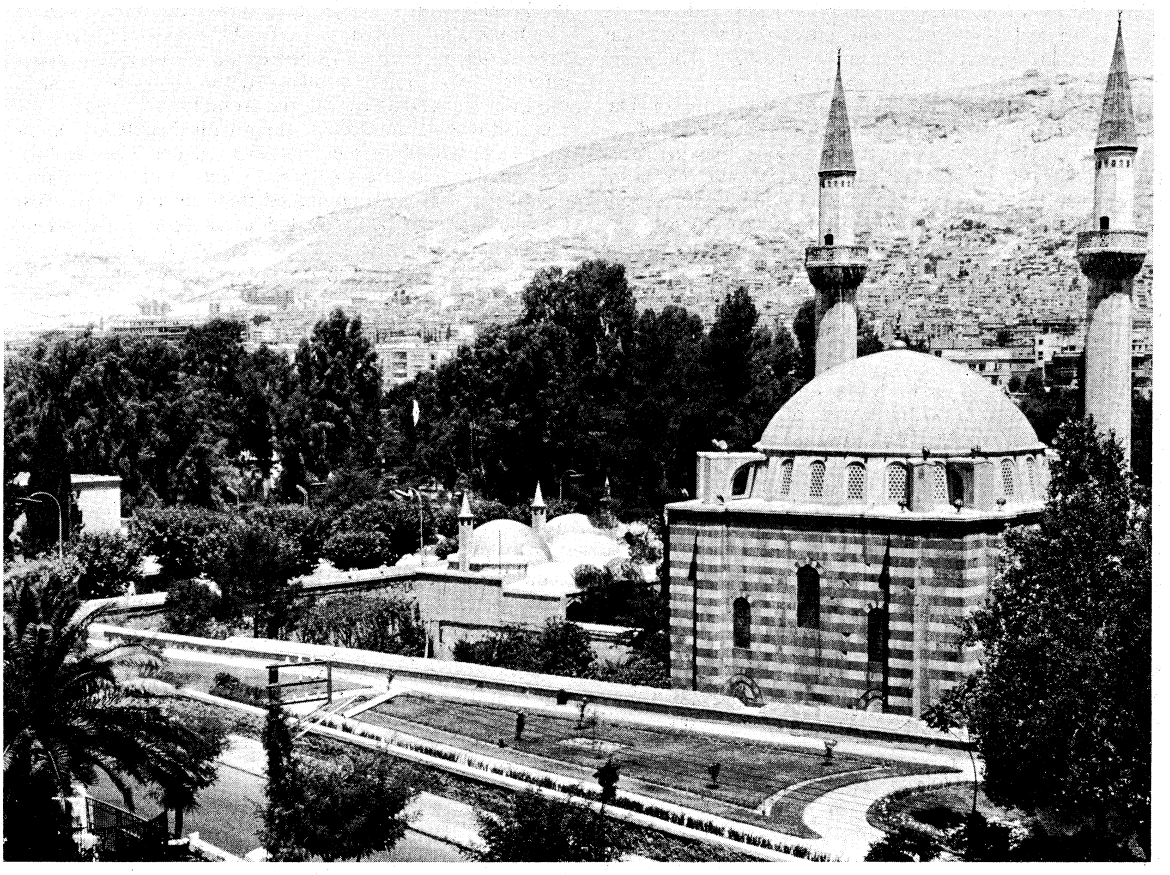
Invasions in ancient times

With Alexander's conquest in 333 BC, Damascus became part of the Hellenistic world for almost a thousand years. The Aramaean quarters coexisted with a new Greek settlement, whose architectural remains may be seen in arcaded streets. Incorporation into the Roman Empire continued the Hellenistic tradition. The citadel in the northwest corner rests on Roman foundations. About 220 yards east of it is the Great Mosque of Damascus, built by the Umayyads on the same site as the Byzantine Church of St. John, the Roman Temple of Jupiter, and the Aramaean sanctuary of Hadad. Still preserved is Ananias (Hanani) Chapel, commemorating the conversion in Damascus of Saul of Tarsus, who became the Apostle St. Paul. It stands near the eastern end of the Street Called Straight (modern Bab Sharqi Street), the classical east–west thoroughfare of the Romans.

With the division of the Roman Empire at the end of the 4th century AD, Damascus became an important military outpost for the Byzantines. Religious and political differences, however, increasingly divided Constantinople from the Syrians. Furthermore, the Persian wars of the 6th century, fought largely on Syrian soil, ruined the economic life of the country. As a result, Damascus opened its gates not unwillingly to the Muslim armies in 635.

THE MUSLIM CITY

Damascus was the first great city of the ancient world that the Muslim Arab forces encountered. In 661, Muʿāwiyah, the first Umayyad caliph, moved out of Arabia and established his court in the Syrian capital. The city was renowned for almost a century thereafter as the capital of



The mosque of Süleyman I, Damascus, with Jabal Qāsiyūn in the background.
Carl Frank

Capital of Islām

a luxurious and extensive empire—the most far-reaching of any achieved by Islām. The principal extant monument of this period is the Great Mosque of Damascus, begun under caliph al-Walīd I in 705. Although it has been damaged, burned, and repaired several times, it is still a glory of Islāmic architecture. On the west wall of the courtyard are the remains of 8th-century mosaics: a golden vision of houses, gardens, streams, and bridges, which has been variously interpreted as a scene of paradise or of Damascus as it then was.

The ʿAbbāsids on coming to power in 750 transferred the capital of the Muslim state east to Baghdad. From that time until the advent of Seljuq Turkish power in the 11th century, Damascus languished in the backwaters of the Muslim world. As ʿAbbāsid power weakened, Damascus and other cities in the region constantly warred with one another. Public order and economic life declined.

During this chaotic period the open plan of the Roman town was modified considerably. With little civic authority city life increasingly centred on quarters where people of common ethnic, religious, or occupational interest lived together: the Muslims in the centre of the city, the Christians to the northeast, and the Jews in the southeast. Behind its barricades, each neighbourhood was thus a minicity under a leader and with its own amenities—mosque, bath, public oven, water supply, and markets. A typical house in the Old City was erected around a courtyard with fountains and trees and showed a blank wall to the street.

Revival under Nureddin

A new era opened when Nureddin (Nūr ad-Dīn ibn Zangī), a powerful Seljuq, captured the city in 1154 and made it once again the capital of a large empire. The town revived; religious buildings were erected, new forms of architecture were introduced, and new quarters for immigrants sprang up. The city continued to flourish under his successors, the Ayyūbids, and their successors, the early Mamlūks, who ruled Egypt and Syria from 1250 until 1382. There were unfortunate interludes when the city was occupied and partially burned by the Mongols. Damascus was the second city—after Cairo—of the Mamlūk Empire, its first line of defense, and the staging point for attacks to drive out crusaders and Mongols. Suqs, or markets, serving the garrisons grew up in the area to the north of the citadel, and suburbs were extended. In the 14th century as many Damascenes lived outside the walls as within. Excess food was exported primarily to the Egyptian capital, and the manufacture of luxury items such as brocaded silks, inlaid metalwork, ceramics, and glass for the conspicuous consumption of the lavish Mamlūk courts and for European markets was encouraged.

For several Islāmic rulers Damascus was a favourite place of residence. Four—Nureddin, Saladin and his brother, al-ʿĀdil, and Baybars I—are interred within the Old City. Their tombs, which are combined with religious colleges, or *madrasah*s, are among the city's most attractive medieval buildings, blending unobtrusively into the urban surroundings. A charming feature is the segmented melon dome that crowns the cupolas. Two of these complexes, the ʿĀdilīyah Madrasah and the aẓ-Ẓāhirīyah Madrasah, face one another across a narrow street and house, respectively, the Arab Academy and the National Library.

Under the later Mamlūks, Damascus suffered from rapacious governors and civil strife among contenders for power. More dire were the pillage of the city in 1401 by Timur (Tamerlane) and his deportation of skilled artisans and workmen to Samarkand. Damascus regained some prosperity in the mid-15th century as corrupt Mamlūk leaders jostled for power. Paradoxically, a lively project of public works continued, but in a more ostentatious fashion than under the Ayyūbids.

Sacking by Timur

## OTTOMAN PERIOD

With the Ottoman conquest of Syria in 1516, Damascus lost its political strength yet retained its commercial importance. Treaties opened the Turkish and Syrian ports of the eastern Mediterranean chiefly to the French and later to other nationalities. For Damascus, Sidon (now in Lebanon) was the chief port. Within the city *khans,* or warehouses, proliferated. These handsome stone structures served as hotels for merchants and places for the storage, exchange, and transshipment of goods. By the

18th century the Damascene merchant had reduced the open courtyard of the *khan* and covered it with cupolas to protect the merchandise from the weather. The As'ad Pash Khan (1732) is a notable example.

A second factor in the continuing prosperity of Damascus was the pilgrimage (*ḥajj*) to Mecca and Medina. Annually a great caravan left Damascus for the Muslim holy cities, under the command of the pasha of Damascus (the Ottoman governor). The pilgrims spent weeks provisioning themselves in Damascus before the caravan set out. The city also profited from trade in the merchandise that the pilgrims brought back from Arabia. The Maydān, the southernmost part of the city, was the headquarters for this traffic, which centred around the 16th-century as-Sinānīyah Mosque with its green-tiled minaret.

Between 1831 and 1840 Syria once more came under Egyptian control with the rise of Muḥammad ʿAlī. Europeans were allowed into the city on more lenient terms; foreign schools and missions were established. The restoration of control of Syria from Constantinople was followed by a violent outbreak of religious fanaticism in the 1850s and early 1860s. After Midhat Paşa, the great Ottoman reformer, became governor in 1878, he made civic improvements, widening streets and improving sanitation. In the early 20th century, the Damascus–Medina rail line, which shortened the pilgrim's trip to five days, was built by German engineers. During World War I, the Syrian capital was the combined headquarters of Ottoman and German forces in their thrust to the Suez Canal and subsequent defensive war against British forces and their Arab allies.

Decline of Turkish rule

Before and during World War I the rise of Arab nationalism found ready ground in Damascus, which became a centre of anti-Ottoman agitation. Fayṣal, son of the grand *sharīf* of Mecca, made secret visits there to enlist support for the Arab cause. In a countermove, Jamal Pasha, the Ottoman commander, hanged 21 Arab nationalists on May 6, 1915, a day that is still commemorated. The Ottomans evacuated the city in September 1918.

## THE MODERN CITY

With the departure of the Ottomans, Damascus entered a new era, during which it has changed in size, physical appearance, and political role. An independent Syrian state was declared in 1919 with Damascus as its capital; Fayṣal was proclaimed king early in 1920. A few months later, the French, with a League of Nations mandate, defeated his army and entered the city. Damascus resisted the French takeover, and an uprising in 1925 was put down only after the French bombarded the city. The years of the French mandate over Syria, from 1920 to 1946, were a period when Damascenes, along with their fellow countrymen, struggled for their nation's independence and for the broader goal of a single Arab state. The Ba'th Party, devoted to that goal, originated there during World War II. The mandate period lasted until April 1946, when, responding to a United Nations resolution, French troops finally left Syria. Once again Damascus was the capital of an independent Syria.

Focus of Syrian politics

Under the republic, Syria's turbulent political life has revolved around Damascus. In this role it has functioned as a pole of attraction for political forces, for economic interests, and for rural people seeking a better life. On several occasions, leadership of the country changed by coup d'état, to the rumble of tanks in the streets. In the 1960s the Ba'th Party came to power in Syria and brought more stability to government.

Contemporary Damascus is a metropolis with many of the features—and problems—found in cities around the world. The physical limits of terrain and finite water sources argue for decentralization to satellite communities some distance away. Were Ibn Baṭṭūṭah, Ibn Jubayr, or other early visitors to return, they would not exclaim so much over a city set in green gardens. They would, however, recognize the spirit and dynamism of the city in the midst of its conversion to a modern metropolis.

**BIBLIOGRAPHY.** General works include ANNE-MARIE BIANQUIS, "Damas et la Ghouta," in *La Syrie d'aujourd'hui,* ed. by ANDRÉ RAYMOND (1980); N. ELISSÉEFF, "Dimashḳ," in *Encyclopaedia of Islam,* new ed., vol. 2, pp. 277–291 (1965); CHRISTINA PHELPS GRANT, *The Syrian Desert: Caravans, Travel, and Exploration* (1937); JEAN SAUVAGET, *Les Monuments historiques de Damas* (1932); and COLIN THUBRON, *Mirror to Damascus* (1967). For climate and geography, see CHARLES COMBIER, *Aperçu sur les climats de la Syrie et du Liban* (1945); RENÉ DUSSAUD, *Topographie historique de la Syrie antique et médiévale* (1927); and RICHARD THOUMIN, *Géographie humaine de la Syrie Centrale* (1936). History is treated in PHILIP K. HITTI, "Damascus: The Imperial Capital," in his *Capital Cities of Arab Islam* (1973); IRA M. LAPIDUS, "Muslim Urban Society in Mamluk Syria," in *The Islamic City: A Colloquium,* ed. by A.H. HOURANI and S.M. STERN (1970); and NICOLA A. ZIADEH, *Damascus Under the Mamlūks* (1964).

(J.F.De./L.C.H.)

# The Art of Dance

Dance is the movement of the body in a rhythmic way, usually to music and within a given space, for the purpose of expressing an idea or emotion, releasing energy, or simply taking delight in the movement itself. Dance is a powerful impulse, but the art of dance is that impulse channeled by skillful performers into something that becomes intensely expressive and that may delight spectators who feel no wish to dance themselves. These two concepts of the art of dance—dance as a powerful impulse and dance as a skillfully choreographed art practiced largely by a professional few—are the two most important connecting ideas running through any consideration of the subject. In dance, the connection between the two concepts is stronger than in some other arts, and neither can exist without the other.

Although the above broad definition covers all forms of the art, philosophers and critics throughout history have suggested different definitions of dance that have amounted to little more than descriptions of the kind of dance with which each writer was most familiar. Thus, Aristotle's statement in the *Poetics* that dance is rhythmic movement whose purpose is "to represent men's characters as well as what they do and suffer" refers to the central role that dance played in classical Greek theatre, where the chorus through its movements reenacted the themes of the drama during lyric interludes.

The English ballet master John Weaver, writing in 1721, argued on the other hand that "Dancing is an elegant, and regular movement, harmoniously composed of beautiful Attitudes, and contrasted graceful Posture of the Body, and parts thereof." Weaver's description reflects very clearly the kind of dignified and courtly movement that characterized the ballet of his time, with its highly formalized aesthetics and lack of forceful emotion. The 19th-century French dance historian Gaston Vuillier also emphasized the qualities of grace, harmony, and beauty, distinguishing "true" dance from the crude and spontaneous movements of early man:

> The choreographic art . . . was probably unknown to the earlier ages of humanity. Savage man, wandering in forests, devouring the quivering flesh of his spoils, can have known nothing of those rhythmic postures which reflect sweet and caressing sensations entirely alien to his moods. The nearest approach to such must have been the leaps and bounds, the incoherent gestures, by which he expressed the joys and furies of his brutal life.

John Martin, the 20th-century dance critic, almost ignored the formal aspect of dance in emphasizing its role as a physical expression of inner emotion. In doing so, he betrayed his own sympathy toward the Expressionist school of modern American dance: "At the root of all these varied manifestations of dancing . . . lies the common impulse to resort to movement to externalise states which we cannot externalise by rational means. This is basic dance."

A truly universal definition of dance must, therefore, return to the fundamental principle that dance is an art form or activity that utilizes the body and the range of movement of which the body is capable. Unlike the movements performed in everyday living, dance movements are not directly related to work, travel, or survival. Dance may, of course, be made up of movements associated with these activities, as in the work dances common to many cultures, and it may even accompany such activities. But even in the most practical dances, movements that make up the dance are not reducible to those of straightforward labour; rather, they involve some extra qualities such as self-expression, aesthetic pleasure, and entertainment.

This article discusses the techniques and components of dance as well as the aesthetic principles behind its appreciation as an art. Various types of dance are discussed with emphasis on their style and choreography. The article DANCE, THE HISTORY OF WESTERN covers the history of dance in the West. For information on the character and history of dance in non-Western cultures, see the sections on dance or the performing arts in AFRICAN ARTS; EAST ASIAN ARTS; CENTRAL ASIAN ARTS; SOUTH ASIAN ARTS; ISLĀMIC ARTS; OCEANIC ARTS; and AMERICAN INDIANS. The interaction between dance and other art forms is discussed in MUSICAL FORMS AND GENRES and FOLK ARTS.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 625.

The article is divided into the following sections:


The aesthetics of dance 935
- Basic motives: self-expression and physical release 935
- Problems in defining dance 936
  - Defining according to function
  - Distinguishing dance from other patterned movement
  - Defining according to intent
- Dance as dramatic expression or abstract form 937
  - The debate in the West
  - Dance as a nonverbal language
- Changes in attitude toward dance 938

Components of the dance 939
- The dancer 939
  - Physical requirements
  - The importance of training
  - Differences among dancers
  - From amateur to professional
- Basic steps and formations 940
  - Ballet and modern dance
  - Folk dance
  - Social dance
- Choreography 941
  - Choreographers' motives and methods
  - The three-phase choreographic process
- Dance notation 943
  - Prominent notation methods
  - Difficulties of notation
- Theatrical elements 944
  - Rhythm and music
  - Set and design
  - Drama

Types of dance 950
- Theatre dance 950
  - Ballet
  - Modern dance
  - Indian classical dance
- Tribal and ethnic dance 954
  - Tribal dance
  - Ethnic dance
- Folk dance 955
- Social dance 956

Bibliography 956


## The aesthetics of dance

### BASIC MOTIVES: SELF-EXPRESSION AND PHYSICAL RELEASE

One of the most basic motives of dance is the expression and communication of emotion. People—and even certain higher animals—often dance as a way of releasing powerful feelings, such as sudden accesses of high spirits, joy, impatience, or anger. These motive forces can be seen not only in the spontaneous skipping, stamping, and jumping movements often performed in moments of

intense emotion, but also in the more formalized movements of "set" dances, such as tribal war dances or festive folk dances. Here the dance helps to generate emotions as well as release them.

Altered perception of the body and environment

People also dance for the pleasure of experiencing the body and the surrounding environment in new and special ways. Dance often involves movement being taken to an extreme, with, for example, the arms being flung or stretched out, the head lifted back, and the body arched or twisted. Also, it often involves a special effort or stylization, such as high kicks, leaps, or measured walks. Dance movements tend to be organized into a spatial or rhythmic pattern, tracing lines or circles on the ground, following a certain order of steps, or conforming to a pattern of regular accents or stresses.

Barbara Morgan



Children dancing in a spontaneous expression of energy and emotion.

All of these characteristics may produce a state of mind and body that is very different from that of everyday experience. The dance requires unaccustomed patterns of muscular exertion and relaxation as well as an unusually intense or sustained expenditure of energy. The dancer may become intensely aware of the force of gravity and of a state of equilibrium or disequilibrium that normal activities do not generate. At the same time, the dance creates a very different perception of time and space for the dancer: time is marked by the rhythmic ordering of movement and by the duration of the dance, and space is organized around the paths along which the dancer travels or around the shapes made by the body.

Dance can, in fact, create a completely self-contained world for dancers, in which they are capable of physical effort, prowess, and endurance far beyond their normal powers. Ṣūfī dervishes, as an extreme example, can whirl ecstatically for long stretches of time without appearing tired or giddy, and certain Indonesian dancers can strike daggers against their naked chests without causing apparent pain or injury.

This transcendence of the everyday may also be experienced by the spectators. Drawn into the rhythms and patterns created by the dancer's movements, they may begin to share in the emotions being expressed through them. They may also experience kinesthetically something similar to the physical sensations of the dancer. Kinesthesia, or the awareness of the body through sensations in the joints, muscles, and tendons, rather than through visual perception, not only defines the dancer's experience of his own body in movement but also the way in which dance exerts its power over the spectators, who not only see it but also feel an echo of the dancer's movements and rhythms in their own nerve endings.

Kinesthesia and the union of dancer and spectator

## PROBLEMS IN DEFINING DANCE

Self-expression and physical release may thus be seen as the two basic motives for dance. Dance itself, however, takes a wide variety of forms, from simple spontaneous activity to formalized art or from a social gathering where everyone participates to a theatrical event with dancers performing before an audience.

**Defining according to function.** Within this broad spectrum of forms, dance fulfills a number of very different functions, including the religious, the military, and the social. Nearly all cultures have had, or still possess, dances that play an important part in religious ritual. There are dances in which the performers and even the spectators work themselves into a trance in order to transcend their ordinary selves and receive the powers of the gods or, as in the case of Indian temple dancers, in which the performers enact the stories of the gods as a way of worshiping them. In some early Christian communities, processions or formal dance patterns formed part of the prayer service.

It is possible to view modern military marches and drilling procedures as descendants of the tribal war and hunting dances that have also been integral to many cultures. War dances, often using weapons and fighting movements, were used throughout history as a way of training soldiers and preparing them emotionally and spiritually for battle. Many hunting tribes performed dances in which the hunters dressed in animal skins and imitated the movements of their prey, thus acquiring the skills of the animal in question and, through sympathetic magic, gaining power over it.

Dance also plays a number of important social roles in all cultures, notably in matters of celebration, courtship, recreation, and entertainment. Courtship dances, for example, allow the dancers to display their vigour and attractiveness and to engage in socially accepted physical contact between the sexes. (The waltz, a relatively modern example of the courtship dance, was banned at certain times because its flagrant contact between the dancers was considered indecent.) Such traditional dances often contain fertility motifs, where mimed (or even actual) motions of sexual intercourse are enacted. One motif in particular, the fertility leap, in which the male dancer lifts the woman as high as he can, is common to many courtship dances, such as the Tyrolean Schuhplattler.

The importance of dance in courtship and social gatherings is probably older than its use as recreation and entertainment. Many scholars have suggested that dance was once an integral part of everyday life, accompanying both practical activities and religious rituals. Only when more complex social and economic structures began to emerge and a leisured class or caste came into existence did people begin to see dance as a source of pleasure, in some way distinct from the most important issues of survival.

As tribal societies gave way to more complex civilizations, many of the earlier ritual forms, such as religious, work, and hunting dances, gradually lost their original significance and developed into recreational folk dances while still retaining many of their original motifs, such as the use of sticks or swords in the English morris dance or the pole in Maypole dances. All kinds of dance in all stages of evolution, however, have retained some importance as means of social cohesion. Dance has also been used as a means of displaying political or social strength and identity. In ancient Greece, for example, citizens were compelled to attend dance dramas partly in order to encourage allegiance to the city-state. An example in the 19th century was Hungary's purposeful revival of its national dances in order to promote a strong sense of national identity.

**Distinguishing dance from other patterned movement.** In all the different dance forms, movement becomes dance through stylization and formal organization, an organization that may be variously determined by an aesthetic idea or by the function of the dance (see below *Choreography*). There are, however, many kinds of activities involving disciplined and patterned movement that do not fit the category of dance—for example, sports or the behaviour of certain animals—because the principles that govern these activities are not the crucial principles of aesthetic pleasure, self-expression, and entertainment.

Distinguishing between a wrestling match and a choreographed fight in a ballet can illustrate the importance of these principles in defining dance. It is easy to distinguish between a real fight and a fight in a ballet because the former occurs in "real life" and the latter takes place in a theatre and because in the latter the antagonists do not

actually want to hurt each other. But in wrestling matches, although the antagonists look as if they are fighting, they are also taking part in a choreographed drama that, like the ballet, is partly appraised on questions of style. In the wrestling match, however, these questions of style are not, as in ballet, central to the event but only incidental. The principle most strongly governing the fighters' movements is the scoring of points rather than aesthetic appeal or self-expression. For this reason, even choreographed wrestling matches do not fit the same category as dance. (The martial arts of Southeast Asia cannot be as easily distinguished from dance, because the movements of the practitioners are expected to be as refined and as graceful as those in dance.)

Competition in sports

Ice skating, particularly in its contemporary form of ice dance competition, is more difficult to distinguish from dance, because both aesthetic and expressive qualities are important. But at the same time, there are certain rules that have to be followed more stringently in ice skating than in dance, and once again the governing principle is the competitive display of skills rather than the enjoyment of movement for its own sake. (Dance competitions in which performers are given points present an even more difficult case of distinguishing art from sport, but, to the extent that it is governed by the principle of scoring points, dance competition cannot be defined as art.)

Marches and processions present another difficulty of classification. Some involve patterned groupings of people and a disciplined, stylized movement such as the military goose step, and the participants may feel and express powerful emotions. Such movements also may be accompanied by highly theatrical elements, such as colourful costumes, props, and music, that often accompany dance. But in a march the movement itself is so subordinate to other considerations—such as the mobilization of large numbers of people or the playing of music—that it cannot be regarded as dance.

**Defining according to intent.** An important factor distinguishing dance from other patterned movement is that of intention. The flight patterns made by swarms of bees or the elaborate courtship rituals of certain birds may be more pleasing to watch and more elaborately organized than the simple, untutored dancing movements of a child. Such patterned movements, however, are not referred to, except analogously, as dances because they are rooted in involuntary genetic behaviour necessary for the survival of the species. In other words, they are not intended as entertainment, aesthetic pleasure, or self-expression. Indeed, it may be argued that for an activity to count as dance, the dancer must be at least capable of distinguishing it as such or must intend it as such. (In a duet by the American choreographer Paul Taylor, two men simply remained motionless on stage for four minutes. Yet the piece was accepted as dance because of its aesthetic context: it was in a theatre and Taylor was known as an experimental choreographer. In addition, the spectators knew that it was intended as a piece that either was dance or was about dance.)

Disagreement over the dominant aesthetic principle

Even when an activity is clearly identified as dancing, there are frequent debates as to whether it is part of the art of dance. Any art form evolves through strong aesthetic principles, and the three main principles governing the art of dance have been discussed above. But of these three principles some may be recognized by one group and not by another. For example, classical ballet reached its zenith in Russia in the late 19th century: Its technique was perfectly developed, and its dancers were acknowledged virtuosos. But a number of choreographers, reacting against the dominant aesthetics of classical ballet, argued that it was simply empty acrobatics and not dance at all because it concentrated on showing the skills of individual dancers and failed to express any significant ideas or emotions. Similarly, when Martha Graham, the pioneer choreographer in American modern dance, first presented her works in the late 1920s, audiences found them so unlike the ballets that they were used to that they refused to acknowledge them as dance (see below *Theatre dance: Modern dance*). The debate goes on over the works of today's avant-garde choreographers, and the same is true for one culture's perceptions of another culture's dance. When Europeans first encountered the highly sophisticated Middle Eastern dance form *raqṣ sharqī,* they perceived it as erotic display and called it the belly dance.

## DANCE AS DRAMATIC EXPRESSION OR ABSTRACT FORM

**The debate in the West.** In Western theatre-dance traditions, notably ballet and modern dance, the most recurrent clash of principles has been over the question of expression. Theatre dance generally falls into two categories: that which is purely formal, or dedicated to the perfection of style and display of skill, and that which is dramatic, or dedicated to the expression of emotion, character, and narrative action. In the early French and Italian ballets of the 16th and 17th centuries, dance was only a part of huge spectacles involving singing, recitation, instrumental music, and elaborate stage design. Although such spectacles were loosely organized around a story or theme, the dance movement itself was largely formal and ornamental, with only a very limited range of mime gestures to convey the action. As dance itself became more virtuosic and ballet began to emerge as a proper theatrical art form, the technical prowess of the dancers became the major focus of interest. Ballet developed into a miscellaneous collection of short pieces inserted, almost at random, into the middle of an opera with no other function than to show off the dancers' skills. In *Lettres sur la danse et sur les ballets* (1760; *Letters on Dancing and Ballets*) Jean-Georges Noverre, the great French choreographer and ballet master, deplored this development. He argued that dance is meaningless unless it has some dramatic and expressive content and that movement should become more natural and accommodate a wider range of expression: "I think . . . this art has remained in its infancy only because its effects have been limited, like those of fireworks designed simply to gratify the eyes. . . . No one has suspected its power of speaking to the heart."

Romantic ballet and dramatic expression

During the great Romantic period of ballet in the first half of the 19th century, Noverre's dream of the ballet d'action was fulfilled as ballet, now a completely independent art form, occupied itself with dramatic themes and emotions. But by the late 19th century the importance attached to virtuosity at the expense of expressiveness had again become an issue. In 1914 the Russian-born choreographer Michel Fokine argued for reform on lines similar to those of Noverre, asserting that "the art of the older ballet turned its back on life and . . . shut itself up in a narrow circle of traditions." Fokine insisted that "dancing and mimetic gesture have no meaning in a ballet unless they serve as an expression of its dramatic action, and they must not be used as a mere divertissement or entertainment, having no connection with the scheme of the whole ballet."

Outside the ballet companies, exponents of modern dance in Europe and the United States were also arguing that ballet expressed nothing of the inner life and emotions, for its stories were childish fantasies and its technique was too artificial to be expressive. Martha Graham, whose commitment to dramatic content was so strong that she often referred to her dance works as dramas, created a new style of movement to express what she saw as the psychological and social condition of modern man: "Life today is nervous, sharp, and zig-zag. It often stops in mid-air. That is what I aim for in my dances. The old balletic forms could not give it voice."

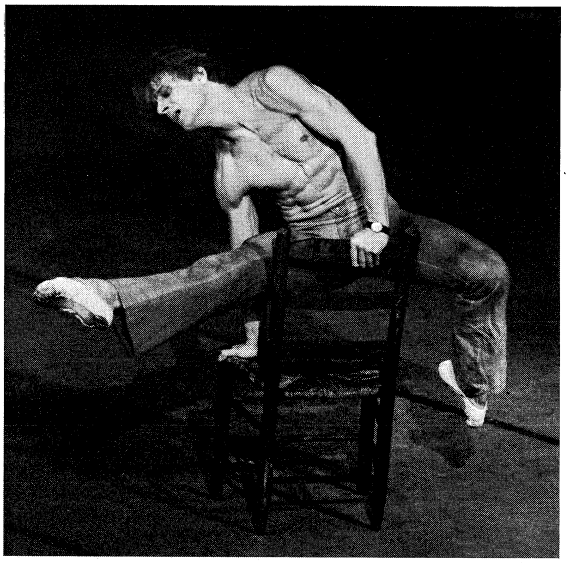
Balanchine and abstract form

In the decades between the world wars, Graham, Mary Wigman, and Doris Humphrey established the school of Expressionist modern dance, which was characterized by serious subject matter and highly dramatic movement. Other choreographers, such as Merce Cunningham and George Balanchine, argued that such close concern with dramatic expression could hamper the development of dance as an art form. Balanchine argued that "the ballet is such a rich art form that it should not be an illustrator of even the most interesting, even the most meaningful literary primary source. The ballet will speak for itself and about itself." The works of these choreographers emphasized formal structure and development of choreography rather than plot, character, or emotion. Partly as a result

of their influence, the "abstract," or plotless, ballet became popular among choreographers during the decades after World War II.

**Dance as a nonverbal language.** At the centre of much debate have been the questions how dance can express emotions and actions in any detailed way and whether it can be thought of as a kind of language. Cultural conventions partly determine the limits of expression. For example, the classical dance of India has more than 4,000 mudras, or gestures through which the dancer portrays complex actions, emotions, and relationships; these gestures are comprehensible to the audience because they have always been at the centre of Indian life and cultural traditions. In classical ballet, however, the vocabulary of mimed gesture is quite small and is comprehensible to only a few informed spectators, thus considerably limiting its expressive range. Referring to the practical impossibility of communicating, through dance, the complex plots and relationships between characters that are common in the spoken theatre, Balanchine once remarked, "There are no mothers-in-law in ballet."

Martha Swope



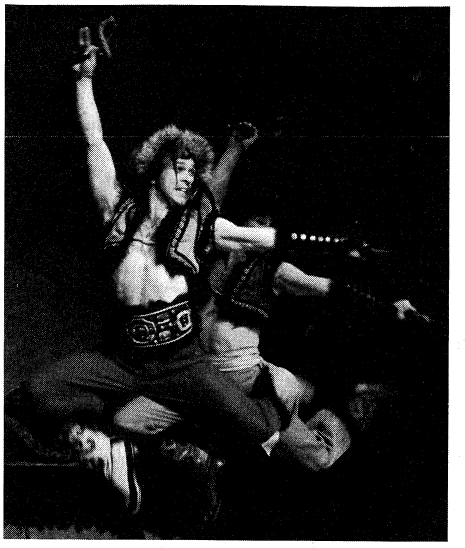
Mikhail Baryshnikov performing *Le Jeune Homme et la mort,* in which a youth (Baryshnikov) expresses his painful attraction to the character Death; choreography by Roland Petit.

While dance cannot communicate specific events or ideas, it is a universal language that can communicate emotions directly and sometimes more powerfully than words. The French poet Stéphane Mallarmé declared that the dancer, "writing with her body, . . . *suggests* things which the written work could *express* only in several paragraphs of dialogue or descriptive prose." Because dance movements are closely related to the gestures of ordinary life, the emotions they express can be immediately understood, partly through a visual appreciation of the gesture and partly through a sympathetic kinesthetic response. Thus, when a dancer leaps, the spectators understand it as a sign of exhilaration, and they feel something of the lifting and tightening sensations that excitement produces in the body. In the same way, if a dancer's body is twisted or contracted, they feel an echo of the knotted sensation of pain.

The immediacy of dance movement

Of course, even the gestures of ordinary life are inherited from cultural conventions. A smile or a wave of the hand can, in certain non-Western cultures, be taken as a sign of aggression rather than welcome. In the same way, how spectators interpret dance movements depends on the context in which those movements occur and on the particular spectator who interprets them. A fall may signify despair in one context, or to one person, and a sinking into ecstasy in another.

The distinction between abstract and expressive dance is also a highly artificial one, becoming a clear distinction in critical theory but certainly not in actual performance. In even the most dramatic and mimetic dance, the movement is highly stylized and subjected to an abstract aesthetic principle. The structure of the piece is determined as much by formal considerations as by the narrative events. On the other hand, even the most abstract work expresses some emotion or character relationship simply because it is performed by people rather than neutral objects, and often the most highly elaborate dance pattern has some representational function.



Vladimir Ivanov in an exuberant leap during the performance of *Polovtsian Dances;* the Moiseyev Dance Company.

MIRA

## CHANGES IN ATTITUDE TOWARD DANCE

Critics have argued the question of abstraction and expression largely in relation to theatre dance and also on the assumption that dance is a serious art form. Within recent history, however, this assumption was not always held. In late 19th-century Europe, outside Russia and Denmark, dance was generally regarded as mere entertainment with little aesthetic value. Attitudes to dance both as an art form and as a social activity have, in fact, varied dramatically throughout history. In cultures where it had, or still possesses, religious significance, it is treated with great respect. The ancient Greeks also took dance very seriously, both as an integral part of their drama—which had strong political and social significance—and as part of education. Plato wrote in the *Laws* that "to sing well and to dance well is to be well educated. Noble dances should confer on the student not only health and agility and beauty, but also goodness of the soul and a well-balanced mind." Aristotle believed that dance was useful for "purging the young soul of unseemly emotions and preparing for the worthy enjoyment of leisure."

The Romans generally looked down on dance as effeminate and decadent. The historian Sallust remarked of a citizen's wife that "she played and danced more gracefully than a respectable woman should." The early Christian leaders took a similar view and tried to repress pagan dance customs wherever they could. This action has been attributed to the Christian belief that the body, being the unworthy vessel of the soul, should not be indulged by any kind of sensual pleasure or display. The attitude was not completely dominant, though, and some leaders felt that sober and decent dances could play an important role in religious worship. In the 4th century St. Basil asked, "Could there be anything more blessed than to imitate on earth the ring-dance of the angels?" Processional, circle, and line dances were included in many church services and can still be seen in some services in Toledo and Seville, Spain.

At the time of the Renaissance, when the hold of the church on secular life loosened, dance became popular at court (the church had never been successful at repressing dance among the peasants). It became an essential part of

every courtier's education to be able to dance and move gracefully, and this was a time, too, when many performed in amateur court ballets. In England dancing was so popular among all classes that foreign ambassadors spoke of the people as the "dancing English."

During the 17th century the Puritans were more effective at stamping out the most exuberant and pagan of English dance customs, though among the upper classes it was still considered proper for young children to learn to dance, in order, as the philosopher John Locke put it, to instill "a becoming confidence" in them. In America the hold of the Puritans was even stronger, and many leaders frowned upon any kind of dance, recreational or otherwise, as idle and lascivious. Others saw it as a necessary part of education, providing that it was sober and serious. The most prominent exception to pious disapproval of dance was the Shaker sect, which, while prospering in the United States during the 18th and 19th centuries, developed choreographed dances as part of its worship service. The dances often represented quite complex religious themes. One figure, the wheel within a wheel, which was made up of circles turning in alternate directions around a central chorus of singers, represented the all-embracing nature of the Gospel; the outer ring of dancers represented the ultimate circle of truth, while the central chorus symbolized the harmony and perfection of God that is at the centre of life.

Gradually, dance as a means of physical education and entertainment became more popular in the United States. Folk dancing and social dancing were encouraged, and by the 20th century theatre dance, too, began to lose its disreputable taint.

Certainly in the Western world, dance as an art form has never been as popular as it is today, with a wide range of choreographic styles and genres attracting large audiences. As a form of recreation it has also undergone a massive revival, for while many of the folk traditions have been lost, some have been carefully revived and are widely enjoyed. In Asia and Africa many traditional dances have been transferred from the community, where they were dying out, to the theatre. This has brought about a rapid growth in their popularity, both in their places of origin and in the West, where they attract large audiences and are also studied.

## Components of the dance

### THE DANCER

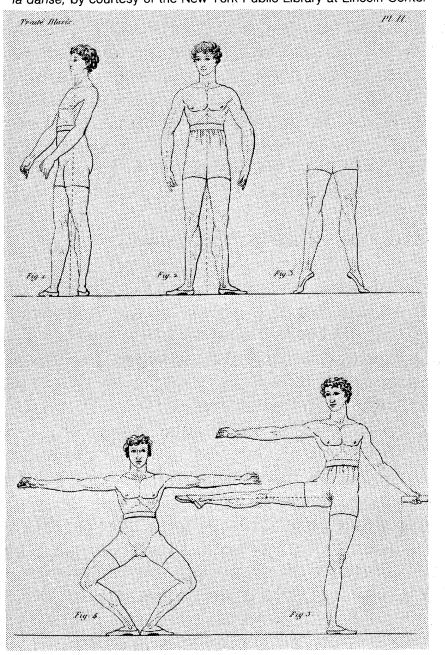
**Physical requirements.** Dancers are not just performing artists; their bodies are also the instruments through which the art is created. The quality of this art, therefore, necessarily depends on the physical qualities and skills that dancers possess. The stronger and more flexible a dancer's body, the more capable it is of a wide range of movement. Nearly all professional dancers start training at a young age in order to shape and develop their bodies correctly. Strength is built up in the right muscles, for example, and the bone-connecting ligaments on which flexibility of the joints is so dependent are lengthened early before they begin to harden.

As well as strength and mobility, a good dancer must also possess great coordination (the ability to work different parts of the body together), a highly developed kinesthetic awareness (in order to know and control the position and state of the body), control over weight and balance in motion, a developed awareness of space, a strong sense of rhythm, and an appreciation of music. Particularly in dramatic dance, the dancer must be able to project movement clearly and make its expressive qualities intelligible to the audience. Grace, fluidity, and harmony of body are also frequently desired in the dancer, as is physical beauty, but these are subjective qualities that differ from one culture to another and change according to fashion. (Today's physical ideal of the ballerina—long-limbed and slender—is quite different from the late 19th-century preference for a more rounded figure.)

Changing standards of beauty

**The importance of training.** Though modern avant-garde choreographers sometimes work with untrained dancers to take advantage of the qualities of natural, untutored movement, most dancers in the West are trained either in a strict technique based on classical ballet or in techniques introduced by the 20th-century modern-dance choreographers Martha Graham and Merce Cunningham. (Other kinds of dance, such as jazz or tap, are usually taught in conjunction with these techniques.) Training generally begins early, between eight and 12 years of age for girls and 14 for boys, although some ballet dancers and many more modern dancers begin later. Ballet training closely follows the rules published in 1828 by the Italian dancing master Carlo Blasis in his *Code of Terpsichore.* Blasis advocated at least three hours of dance classes a day, involving exercises that progressively developed different parts of the body.

From Carlo Blasis, *Traite elementaire, theorique, et pratique de l'art de la danse,* by courtesy of the New York Public Library at Lincoln Center



Engravings of ballet positions from *Traité élémentaire, théorique, et pratique de l'art de la danse* (1820) by Carlo Blasis.

Daily classes are necessary not only to mold the body and develop the necessary physical skills but also to maintain the body in its proper condition and prevent injury. Many dance movements make strenuous and unnatural demands on the joints, muscles, and tendons, and it is easy to strain or damage them if the body is not properly maintained. Some bodies are more suitable for training than others, and in the West many aspiring dancers undergo extensive medical scrutiny to ensure that they have no weaknesses or disabilities, such as a weak or crooked spine, that would make them unfit for dancing.

The exercises involved in a dancer's training depend on the style of the dance. Ballet dancers have to work hard to attain a full turnout (the outward rotation of the legs in the hip socket so that the heels touch back to back and the feet form a 180° angle), which enables them to lift their legs high in the air in jumps or arabesques. While ballet dancers rarely use the torso, African dancers and certain modern dancers have to be extraordinarily supple in the torso and pelvis in order to execute the ripples, twists, and percussive thrusts that their particular dances require. Indian classical dancers, while developing great strength and flexibility in the legs, must also achieve great control over the face and neck muscles and flexibility and control in the joints and muscles of the hands. This is necessary to execute their elaborate mudras, conventional symbolic gestures, with accuracy and grace.

**Differences among dancers.** However rigorous and uniform training may be, each dancer always has a personal style of dancing. Certain skills come more easily to some dancers than to others: one may be an excellent jumper, while another may have exquisite control and balance in

slow, sustained dance passages. The same choreography may also look completely different when executed by two different bodies. Thus, a dancer with very long limbs will make high leg extensions look exaggeratedly long while appearing slightly awkward in fast, intricate footwork. Another dancer may have a great deal of energy and speed but be unable to produce a sustained and beautiful line in held positions.

Finally, dancers vary a great deal in the way they articulate and project movement. Some dancers move in a way that is tense, energetic, and even aggressive in its attack, while others appear soft and fluid. Some phrase their movements so that every detail is sharp and clear; others so that one element flows into another. Some move exactly in time with the phrasing of the music; others phrase their movement slightly independently of it. One dancer may produce movements that are dramatically charged and expressive, while another may be cool and detached, concentrating on technical perfection. Such qualities may vary so markedly that certain dance roles become inextricably connected to the dancers for whom they were created, for example, Anna Pavlova's *Dying Swan,* created by Michel Fokine in 1907, or Rudolf Nureyev and Margot Fonteyn's *Marguerite and Armand,* created by Frederick Ashton in 1963.

In modern dance the dancer may be highly esteemed for individual style and technique but is generally expected to submit his own personality to the demands of the choreography. Some of the works by the American choreographer Alwin Nikolais went so far as to conceal the dancer altogether under a panoply of props, costumes, and lighting projections.

Individuality and conformity

The display of individual style is inevitable in theatre dances such as ballet and modern dance, where trained professionals perform for the pleasure of an audience. Some participatory dances also allow individual dancers to display their talents, as in ballroom or disco dancing, but in many folk dances, particularly those derived from ancient rituals, the sense of unity within the group usually outweighs the importance attached to any one dancer. In primitive religious dance such unity tends to be even more strictly observed. The point of the dance is not the display of the dancer's or choreographer's talents but the perfection of the ritual.

**From amateur to professional.** Exacting standards and rigorous early training are common where dance has become an art performed before an audience. Such scholars as the German musicologist Curt Sachs have pointed out that in very early cultures, where dance was something in which everyone in the tribe participated, dancers were not regarded as specialists to be singled out and trained because of their particular skills or beauty. Once religious worship (the original occasion for dance) developed into ritual, however, it became important for dancers to be as skilled as possible, for if the ritual was not performed well and accurately, the prayers or magic would not succeed. Dancers were thus selected for special training, which may have taken place either through the family or through skilled individuals who lived and taught outside the community. The dancer's performance now became subject to the most rigorous judgments; indeed, Sachs mentions a tribe on the island of Santa Maria in the New Hebrides (present-day Vanuatu) where, it was said, "old men used to stand by with bows and arrows and shoot at every dancer who made a mistake."

Frequently, in religious dances, the dancer is subjected not only to intense physical training but also to spiritual discipline. Such dancers have often formed a special caste set apart from the rest of the community. In the religious hula dances of Hawaii, the dancers observed important taboos and took part in sacred rites in order to be fit to dance. The traditional religious dancers of India also had to remain pure; they were regarded as brides of the gods and were taught by masters of the highest caste. (Frequently such practices became corrupt, and female temple dancers were also paid to perform in the houses of the wealthy, thus acquiring a reputation for sexual license and promiscuity.)

In Europe professional dance was for many centuries restricted to joculators, wandering bands of jugglers, dancers, poets, and musicians, who were generally regarded as social inferiors. The early ballets were performed almost exclusively by amateur dancers at court (though instructed by professional dancing masters) for whom dance was a means of demonstrating their own grace, dignity, and good manners. The comic, or burlesque, parts alone were performed by professionals, who were not so concerned about their dignity. It was Louis XIV of France, himself an enthusiastic amateur dancer when young, who realized that the art of dance could not advance unless dancers were properly trained professionals. To provide standards for this training, he set up the Académie Royale de Danse in 1661, merging it with the Académie Royale de Musique in 1672. (The Académie survives to the present day as the Paris Opéra Ballet.) Through the work of masters such as Pierre Beauchamps, first director of the Académie Royale de Danse, the main principles of dance technique were codified, and dancers rapidly reached much greater heights of virtuosity. Before Louis's innovations, the split between court dancing, with its carefully cultivated style and patterns of movement, and the less refined peasant dances was already marked, but from this point the gap between professional and amateur dance in Europe really came into being.

Académie Royale de Danse

## BASIC STEPS AND FORMATIONS

**Ballet and modern dance.** The style and movement vocabulary of classical ballet is rooted in the five turned-out positions of the feet: (1) heels touching and feet forming a straight line; (2) heels apart and feet forming a straight line; (3) one foot in front of the other with the heel against the instep; (4) feet apart, one in front of the other; and (5) one foot in front of the other with the heel against the joint of the big toe. Each of these ballet positions has a corresponding port de bras, or position of the arms and hands.

The five ballet positions

Movements may be grouped into several broad types. First, there are quick, earthbound, linking steps—for example, the pas de bourrée, a flowing step that may be executed in any direction, and the glissade, a gliding step in which the dancer stretches one foot to the side, front, or back, then stretches the other and brings it in to meet the first.

Second, there are jumps, which may be low and light, with the feet battu ("beaten," or crossed rapidly in front of and behind each other several times in midair). In the entrechat, the dancer takes off from the fifth position into a vertical jump. In an assemblé the dancer brushes one foot out to the side, front, or back while springing off of the other; the two feet then come together in midair (where they may be beaten), and the dancer lands in the fifth position. The pas de chat ("step of the cat") is a jump to the side, with first one foot and then the other drawn up beneath the dancer's body before landing in the fifth position. Higher, more vigorous jumps include the grand jeté, in which the dancer throws one leg forward into the air, hovers with the legs stretched to the front and back, and then lands on the front leg, either holding a position such as arabesque or attitude or else closing the back foot into the fifth position.

Arabesque and attitude are positions in which the dancer stands on one leg. In arabesque the other leg (called the working leg) is stretched straight out to the back; in attitude, it is bent and may be extended to either the front or the back.

Turns include the pirouette, which is executed on one leg and on the spot, with the working leg held in a variety of positions, such as attitude, stretched out to the side (*à la seconde*), or with the foot held just above the ankle or at the knee. In the fouetté en tournant the working leg is whipped straight out to the side and then bent in, the foot being brought back to the knee of the supporting leg at each revolution. The piqué is a traveling turn, the dancer stepping out onto the supporting leg before turning on it.

All of these steps may be performed in numerous enchaînements, or combinations, and with the dancers grouped in many different formations. In classical ballet the formations tend to be symmetrical, with circles or lines framing the main dancers at centre stage. Adagio, or

partner work, is crucial to ballet; the man may support the woman in a series of pirouettes or balances and may lift her in many ways. As a general rule, the pas de deux, solo, and group dance alternate fairly regularly, and in the classical pas de deux the two dancers generally separate for individual variations before coming together in a final coda.

Modern dance variations

Modern dance uses many of the steps and positions of classical dance but often in a very different style. The legs may be turned in and the feet flexed or held loosely rather than pointed (see below *Types of dance: modern dance*). There is much greater use of the torso, which may twist, bend, or crouch, and more rolls and falls, in which the dancer works on or close to the floor. Much postmodern dance uses ordinary movements, such as running or walking, as well as simple swinging, spiraling, or stretching movements that involve the entire body.

**Folk dance.** Many of the steps used in folk dance are like basic versions of ballet steps: small hops and skips; running steps similar to the pas de bourée; and more vigorous steps such as the gallop, in which one leg slides to the front or side and the other leg is brought to meet it in the air with a small spring before the dancer lands on it, ready to slide the original leg forward again. Also common are simple turns, where the dancer pivots on one leg, and lifts, where the man catches his partner around the waist and lifts her into the air. Arm and body movements are usually simple and relaxed, with hands held at the waist or hanging at the sides and the body swaying in rhythm to the movement. In some dances the performers remain separate; in others, they hold hands, link arms, or clasp one another around the waist. Steps are usually repeated in long series, but they often follow quite complex and strictly ordered floor patterns—such as the figure eight, in which the dancers weave around one another. Whether single or in pairs, dancers are usually grouped in circles (often two concentric circles moving at the same time) or lines. Within these groupings there are many specific formations; for example, four or more dancers hold hands and move in a circle, or dancers join hands to form an arch under which the others can pass.

Culver Pictures, Inc.



Traditional Maypole dance from England, with circle formation of dancers interweaving; detail from a 19th-century drawing.

**Social dance.** Except for display, social dances are rarely performed in any strict formation, although dancers may sometimes form themselves spontaneously into lines or circles. Ballroom dances are categorized instead by their step patterns, rhythms, and tempos. Some of the best-known social dances are the waltz, fox-trot, tango, rumba, samba, and cha-cha. The fox-trot is danced in moderate time, with two steps forward and two steps to the side executed in a slow-slow-quick-quick rhythm. The waltz is a three-step dance, and in the cha-cha five steps are executed in a four-beat measure with a slow-slow-quick-quick-slow rhythm.

The basic step patterns are elaborated by different turns, the dancers pivoting on one foot, as in the waltz, or changing direction while they walk, as in the fox-trot. There are also different kinds of walk—for example, the chassé, in which one foot slides to the front, side, or back, the other slides to meet it, and then the first slides forward again.

The ballroom hold

In many ballroom dances the man and woman remain in the same hold throughout, facing each other or turned slightly to the side. The head may turn and the body sway or bend in response to the rhythm and footwork. In the cha-cha the man and woman may also dance separately. With more freedom to move the body and arms, the hips may sway with the steps and the arms swing rhythmically across the torso.

The basic steps of the original rock and roll dances are performed in the traditional ballroom hold. Dancers may then "break" in order to perform different lifts and turns. For example, the man may hold onto the woman's hand and pivot her under his arm, the woman may jump up with her legs straddling the man's waist, or the man may catch hold of the woman's shoulders and slide her between his legs.

In most later rock dances, from the twist to disco, it is much rarer for people to dance as partners. There is also a greater stress on free arm and body movements than on set patterns of steps. Disco enthusiasts may incorporate elements of jazz, modern dance, and gymnastics into their repertoire, executing high kicks, turns, and even backflips.

## CHOREOGRAPHY

Choreography is the art of making dances, the gathering and organization of movement into order and pattern. Most recent works of Western theatre dance have been created by single choreographers, who have been regarded as the authors and owners of their works in a way comparable to writers, composers, and painters. Most social and recreational dances, on the other hand, are products of long evolution, involving innovations that groups of people or anonymous individuals have brought to traditional forms. This evolutionary process is also typical of much non-Western choreography, where both the form and steps of dances are handed down from one generation to another and subject only to gradual and partial change. Even in cultures where it is common for dancers and dancing masters to create their own variations on existing dances, as among the Hopi Indians in northeastern Arizona, it may not be traditional to honour an individual as a particular dance's creator.

**Choreographers' motives and methods.** When choreographers set out to create new works, or possibly rework traditional dances, their impulses or motivations for doing so vary widely. It may be that a particular dance has a function to fulfill, such as marking a celebration, embellishing an opera, or praying for rain. It may be that the piece has no specific function and that the choreographer is simply responding to an outside stimulus—a piece of music that has suggested a structure or movement, perhaps, or a painting, or a theme from literature, or possibly a particular dancer that the choreographer is interested in working with. Or the stimulus may be the choreographer's desire to express a particular concept or emotion or a fas-

cination with a particular choreographic idea. Such stimuli may, of course, influence the work even if the choreographer is producing it for a specific purpose, though, as with any artist, it is rare that a choreographer's motives and intentions can be clearly analyzed—particularly during the actual working process.

The methods by which different choreographers create their work also vary. Some work closely with the dancers from the beginning, trying out ideas and taking suggestions from the dancers themselves before pulling all of the material together. Others start with clear ideas about the shape of the piece and its content even before going into the studio. The 19th-century choreographer Marius Petipa used small models to work out the groupings of his dances. The amount that any choreographer can do without dancers is limited, because the notation of dance is relatively undeveloped. Whereas a composer can write a complete symphony without meeting the orchestra that is going to play it, dance notation is mostly used in recording rather than creating dances (see below *Dance notation*).

**The three-phase choreographic process.** The choreographic process may be divided for analytical purposes (the divisions are never distinct in practice) into three phases: gathering together the movement material, developing movements into dance phrases, and creating the final structure of the work.

*Gathering the movement material.* The way in which the choreographer accumulates movement material depends on the tradition in which he works. In certain dance forms it may be simply a question of creating variations within a traditional pattern of movements. For example, dancing masters in the Italian courts of the 14th and 15th centuries simply invented variations on existing dances and published them in dance manuals bearing their own names. Even today many ballet choreographers use as raw material for their pieces the traditional steps and enchaînements that dancers learn in class. The same is true for many of today's performers of Indian or Middle Eastern dance forms; they may not strictly follow the traditional structure and sequence of movements passed down to them, but they remain faithful to their characteristic styles, retaining the traditional quality of movement and not introducing steps or movements widely different from the original.

Building on tradition

In modern Western forms choreographers have worked less within established traditions, creating instead a vocabulary and style of movement to suit their own personal visions. But even in the work of pioneering choreographers, it is possible to trace major influences. Martha Graham's early work, in the 1920s, for example, was strongly influenced by the American Indian and Southeast Asian dance forms used by her mentor, Ruth Saint Denis. Merce Cunningham's technique owed a great deal to classical ballet. Even Vaslav Nijinsky's ballet *Le Sacre du printemps* (*The Rite of Spring*), which audiences at its first performance in 1913 regarded as a complete break with known dance forms, may have been influenced by the rhythmic-movement exercises of the music teacher Émile Jaques-Dalcroze and by the interest in archaic dance forms already generated by Isadora Duncan and Michel Fokine.

Although each choreographer draws material from diverse sources and often employs contrasting styles, most dance works of a single choreographer show a characteristic style of movement. Dances, however, are rarely if ever a loose collection of isolated movements. One of the most important features of any choreographer's style is the way in which movement material is connected into dance phrases.

*Developing movements into phrases.* A phrase, loosely speaking, is a series of movements bound together by a physical impulse or line of energy and having a discernible beginning and end. (A rough analogy can be made with the way a singer phrases a multiplicity of notes within a single breath.) Many factors work to make the spectator perceive a series of movements as a phrase. The first is the recognition of some kind of logical connection between the movements that prevents them from appearing arbitrary and isolated. It may be that one movement flows easily and naturally into another within the phrase and that there are no awkward transitions or that there is some clearly visible pattern to the movement (such as the basic three-step phrase in the waltz). Rhythm is a significant factor, and movements are often clearly linked by a recognizable pattern of accents. A movement's accent is measured by its force and duration; thus, a hard, sharp movement has a strong accent, while a soft, gradual movement has a weak one. Even a single movement, such as a head roll, may begin with a strong accent and end with a weak one. In phrases that have perfectly regular rhythm, the strong and weak accents recur in the same sequence and always over the same duration of time.

Dance phrases vary both in length and shape. A phrase may begin with a very forceful movement, or maximum output of energy, that gradually comes to a pause, or it may have its climax somewhere in the middle or at the end. Other dance phrases, in contrast, have an even distribution of energy. These factors determine the way in which the phrase is perceived by, and the effect that it produces on, the spectator. Long, repetitive, evenly paced phrases produce a hypnotic effect, while a series of short phrases with strong climaxes appears nervous and dramatic. One of the distinguishing features of Graham's early style was her elimination of linking steps and fluid transitions between movements, so that many of her dance phrases were short, stark, and forceful.

Effects of various phrases on the spectator

Once a phrase has been constructed, it can be built onto in many different ways. Perhaps the simplest ways are repetition, in which the same phrase is simply repeated, and accumulation, in which the original phrase is repeated with a new phrase added on each time. Separate dance phrases may also be repeated according to a pattern, one of the most basic being the alternation of two phrases, and another being the passing of one or more phrases from one dancer to another in canonic form. Material within a dance phrase can also be developed in a number of ways to create new material. The simplest of these is a straightforward reversal of the sequence of movements in the phrase, but more complex principles of motif and development and of theme and variation are also common. The principle of theme and variation works on the same initial dance phrase being repeated in a number of different ways; for example, with different numbers of people, at different speeds, with different styles of movement (jerky or smooth), or with different dramatic qualities (happy or sad). In motif and development, material from within the phrase is developed in new ways, for example, by embellishing it with other movements (the same jump but with different arm movements), by imitating it on a different scale (the same jump, only bigger or smaller), or by fragmenting it and repeating only small details.

Motif and development, theme and variation

*Creating the final structure.* The third phase of the choreographic process, creating the overall structure of the dance, may be influenced by a variety of considerations, including the purpose of the dance. If the work is to be a narrative piece, the plot will obviously determine the way in which the dance material is to be structured. It may have to follow a strict succession of events, create characters in a particular order, and bring the drama to climax at the proper moments. Similarly, if the dance forms part of a ritual, the material may have to strictly follow sanctioned form and procedure.

The music determines the structure of a dance work, too—by its length, its arrangement of fast and slow movements, and its treatment of theme. Many of Balanchine's works follow the structure of the accompanying score very closely; this is reflected in pieces with such titles as *Symphony in Three Movements* (1972), set to music by Stravinsky, or *Concerto barocco* (1940), set to music by Bach. Many dance forms actually have the same names as musical forms—such as the rondo, which, by repeating an initial movement in alternation with various contrasting movements, follows the same scheme as its musical counterpart.

A dance's purpose and its musical score are outside influences on its structure. But structure may also be organic; in other words, an entire dance piece may arise from a continuous development of movement ideas, each movement working off of the movement that came before.

Organic development of structure

Richard Alston's *Doublework* (1978), for example, derived its structure from the exploration of the duet form and the repetition of dance material in different contexts. Other movement ideas that may develop in this way are the use of contrasting sections of movement (a section of fast, energetic dancing followed by a slow, meditative passage), the deployment of different numbers and configurations of dancers (a solo followed by an ensemble followed by a trio, and so on), and the manipulation of different floor patterns or different areas of space (a section of leaping movements contrasted with movement executed very close to the ground).

Movement usually develops organically even when the overall structure of the piece is imposed by a plot or piece of music. In the case of narrative ballets, choreographic ideas may develop into formal motifs while still retaining the ability to represent certain actions or situations in the plot. For example, in Ashton's *Fille mal gardée* (1960) ribbons represent the lovers' emotions; tied into a love knot, they signify their passion, and transformed into a skipping rope and cat's cradle, they show their innocence. But at the same time, the ribbons are used in a purely formal way, embellishing certain movements or creating elaborate patterns that can be enjoyed solely for their beauty. In even the most dramatic ballets the representation of emotions and events is heavily stylized, and the ordering of the plot is determined as much by aesthetic as by dramatic logic. Many narrative ballets, like those of Petipa, contain sections of nondramatic dance that develop according to the kind of formal choreographic principles described above.

Finally, the structure of a dance reflects the tradition in which it is created and performed. Ballets in the 19th-century classical tradition tend to last an entire evening and are divided into several acts, with the tragic death or happy marriage of the protagonists occurring at the end. Modern dances are often much shorter, and a single program may include up to a half-dozen pieces. In a performance of the Indian dance form bharata natya, sections of dramatic and abstract dance follow one another in strict succession for a period lasting up to four and a half hours, while in the kathakali dance form of southwestern India, a single performance of alternating dance and music may go on for 16 hours.

## DANCE NOTATION

Since dance is a performing art, the survival of any dance work depends either on its being preserved through tradition or on its being written down in some form. Where tradition is continuous and uninterrupted, changes in style and interpretation (inevitable when different dancers perform the same material) may be corrected and the dance preserved in its original form. But when a tradition is broken (if, for instance, the cultural traditions of one ethnic group encroach on those of another), then dances may not only change radically but may even disappear. For this reason methods of recording dance are important in the preservation of its history.

The *Nāṭya-śāstra* and Indian classical dance

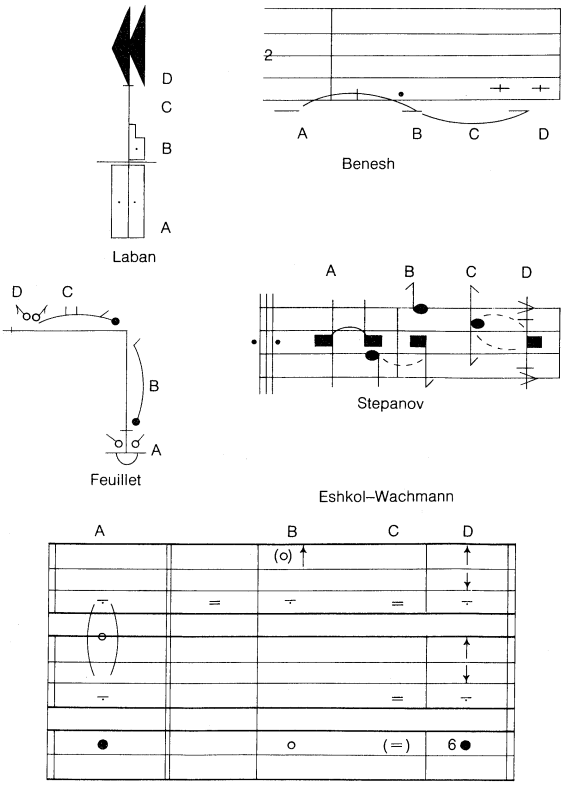
Evidence of dance records dates to the ancient Egyptians, who used hieroglyphs to represent dance movements. In India the earliest book discussing dance, the *Nāṭya-śāstra* ("Treatise on the Dramatic Arts"; variously dated from the 2nd century BC to the 3rd century AD), still survives. This work, which is sacred in Indian culture, codifies dance into a series of rules determining the gestures used to depict different themes and emotions. The bharata natya, a classical dance form based on this treatise, is a good example of a dance tradition that has survived unbroken for many centuries. It only began to founder during the 19th century, partly because Westerners and Indians alike began to deplore its associations with prostitution, but was saved from disappearing altogether when it was developed into a concert form at the beginning of the 20th century. One reason for the long survival of the bharata natya was its importance in religious ceremonies of Hinduism; in addition, when Indian dances were rarely being performed and were in danger of being lost or of degenerating beyond recognition, the *Nāṭya-śāstra* provided a record of traditional principles and styles for their later revival. Even today, not all dance instructors are familiar with these principles, and purists still fear that certain dances are in danger of disappearing or being completely distorted.

**Prominent notation methods.** The absence in the West of any reliable form of notation until the 20th century resulted in a relative paucity of dance traditions when compared to other art forms. While the music, art, and literature of many centuries past is available today, either in the original or in a reproduced form, there is no complete record of any of the ballets choreographed before the 19th century. Even those works that form the backbone of ballet's classical tradition (*Swan Lake, Giselle,* and *The Sleeping Beauty,* for example) have not survived in forms that fully resemble the original choreography.

During the Renaissance dances were recorded through a simple form of verbal abbreviation, with one letter standing for each individual step—as in B for branle or R for reverence. This method was adequate because the dances of that time were simple and the individual steps were well known. By the 17th century the increasingly complex floor patterns of certain dances, particularly those of the court ballets, led to the emergence of track-drawing systems, the most sophisticated of which was published in 1700 by Raoul-Auger Feuillet in his *Chorégraphie, ou l'art de décrire la danse* ("Choreography, or the Art of Describing the Dance"). Feuillet's work recorded foot positions and combinations of steps as well as floor patterns, but it was unable to register movements in the upper part of the body.

The track-drawing system

Subsequent ballet masters turned to a form of notation using stick figures, the first of which was *La Sténochorégraphie,* published in 1852 by the Frenchman Arthur Saint-Léon. The disadvantage of this system was that it could not record the timing or musical coordination of movements, so that later attempts to produce a system were based on musical notes that would give not only anatomical detail but also the duration of the movement. In the 19th century the most advanced system of this kind was published in *Alphabet des mouvements du corps humain* (1892; *Alphabet of Movements of the Human Body*), by Vladimir Stepanov, a dancer at the Mariinsky Theatre in St. Petersburg (now the Kirov State Academic Theatre




*Comparison of five systems of dance notation.*
(A) Starting position: stand with feet together. (B) Step forward on the right foot (count 1). (C) Spring into the air. (D) Land to the left, feet together, knees bent (count 2).

of Opera and Ballet in Leningrad). Stepanov's system was used to record many ballets in the Mariinsky's repertoire; the recordings were the basis of subsequent reconstructions of those ballets by the Sadler's Wells Ballet in London.

Stepanov's system still had many disadvantages, one of the most significant being that it was strongly geared toward ballet and could not accommodate the wider range of movements being developed through modern dance techniques. In 1928 Rudolf Laban, a Hungarian dancer, teacher, and choreographer, developed a complex series of principles for analyzing the full range of human movement. His system for recording movements in dance—widely known as Labanotation—had the advantage of being able to record not only the positions of the body and the pattern of the steps but also the way in which movements should be executed (*i.e.,* whether they should be relaxed or forceful and where the accent of the movement should lie).

Labano-ation

Choreology, developed by Joan and Rudolf Benesh in 1955, is based on a more clearly visual rather than symbolic form of notation. It is written on a five-line stave, recording the dancer's position as viewed from behind. The top line shows the position of the top of the head; the second, the shoulders; the third, the waist; the fourth, the knees; and the fifth, the feet. Special symbols such as lines, dots, and crosses indicate what each part of the body is doing—for example, whether a limb is straight or flexed and in which direction (to the side or front or in a circle) each part is moving. Other symbols show the quality or dynamics of the movement, its rhythm and accent, and the group formations of the dancers. Choreology is now the most frequently used system, particularly in Britain. In 1958 Noa Eshkol and Abraham Wachmann proposed a mathematical system in which movement (of the joints, for example) was analyzed anatomically, in degrees of circular movement in either positive or negative directions, with positions of the body being fixed in relation to two coordinates.

Video recording is more readily accessible than written notation, though it fails to represent the three-dimensional nature of dance and is unable to record movements when one dancer is concealed behind another. It may be useful when used in tandem with some form of written notation, particularly as it can provide a record of how individual dancers interpret particular roles.

**Difficulties of notation.** The problem with all systems of dance notation is that few choreographers—and even fewer dancers—are literate in them. As currently practiced, dance notation is mostly used only for the recording, rather than the creating and learning, of dances. Given the present method of creating in the studio, it is impossible for a choreographer to take an overall view of the work; it is difficult to make changes or to experiment in the way a composer can, because the choreographer is limited by the relatively short period of time allowed for rehearsals and by practical considerations such as the dancers' availability and fatigue.

Even the best system of notation cannot succeed completely, because it cannot alter the fundamental nature of dance. Like any other performing art, dance is essentially ephemeral, existing only at the time of its performance. It can never be properly recorded or preserved, since the way in which dancers interpret a work—their styles, technical abilities, and physical appearance—always change the work each time it is performed.

Ephemeral nature of dance

## THEATRICAL ELEMENTS

Music, design, and drama have all played important roles in the evolution of dance, and in many cultures dance has actually been inseparable from these arts. The Greek word *mousikē,* for example, referring to music, poetry, and dance as one form, reflected the integral relation between these three arts in classical Greek drama. In the early European ballets, dance, music, drama, and spectacle were equally inseparable.

Even where dance is perceived as an independent art form, most choreography is still accompanied by one or more of these elements. Choreographers generally regard them as integral parts of the works. Sound and visual effects, for example, can clarify the dramatic effect of a dance movement and can also help the spectator to perceive more fully its aesthetic qualities. In a general way, music, design, and drama also work together to heighten the experience of dance as something removed from everyday experience, inspiring a special attention in the spectator.

**Rhythm and music.** The most important element of dance is music, and it is rare for dance of any kind—social, theatrical, or religious—to develop without musical accompaniment. The close relation between dance and music is based on the fact that both are organized around rhythmic pattern; thus, the rhythm of the accompanying music may be used to determine the rhythm of the dance, to give it emphasis, or to help the dancers maintain the same beat.

*Rhythm.* Nearly all physical activity is done rhythmically, as in the beating of the heart, the flow of the breath, and the actions of walking and running. Work activities such as digging, sawing, scrubbing, or planting also tend to fall into a regular rhythm, because that is the most efficient and economical way of working the muscles and pacing the effort. When the rhythm is perfectly even, a regular pattern of time and force is established—each inhalation and exhalation of the breath and each stride or stroke of the saw taking the same amount of time and using the same amount of energy. In dance, too, the setting up of regular, efficient rhythms may also be important in allowing the dancer to continue dancing for a long time, whether the dancer be a Ṣūfī dervish or a disco dancer.

Individual dance movements also have a natural rhythm that determines the way in which they can be executed. A high leap, for example, can take only a certain amount of time (the force of gravity preventing a very prolonged duration and the height of the leap precluding a very quick one). Thus, the rhythm, or pattern of accents, imposed on the leap can be neither very sharp nor very sustained.

Even though choreographers are limited to those rhythms permitted by the various dance movements, they do not always use those that are most natural and efficient. It may be easier for a dancer to perform a section of runs and jumps at a moderate, evenly paced rhythm, but this may not produce the effect that the choreographer wants.

Choreographers vary dance rhythms for many reasons, the most basic being the wish to create different qualities of movement—a slow, even rhythm, for example, to create softness and fluidity, or a fast, asymmetrical rhythm to make the movement attenuated or uneven. Varying the qualities of movement may also have a dramatic function, rhythm often determining whether a movement appears joyous, calm, or anguished. Also, choreographers following a musical score may manipulate the rhythms of the dance movements either to match or counterpoint those of the music.

Effects of varying rhythm

Rhythm is a vital element of all dances in all cultures, even in those African and Asian dances whose complex rhythms are often imperceptible to the Western observer. In these forms, the drummer may play a different rhythm with each hand, one setting the basic pulse and the other producing a pattern of sound to reflect the mood or meaning of the dance. The dancer, too, may set up one rhythm in the stamping of the feet while marking out another in the torso, arms, or head, thus producing a highly varied and irregular pattern of sounds and movements. It is rare for dance not to follow any kind of rhythmic organization, just as poets who do not follow a strict metre still emphasize and manipulate the rhythms of language.

*Music.* Many of the terms used in reference to dance rhythm, such as tempo, dynamics, and beat, are derived from music, as most dance is either set to music or accompanied by it. Particularly in cases where the choreographer sets the dance to a previously composed score, the music may determine both the length and structure of the work and even the exact phrasing of the movements. At its simplest, there may be an exact correspondence between the notes and the dance steps, as in a basic waltz melody. On a more complex scale, as in the music visualization popular with such choreographers as Ruth Saint Denis, dancers or groups of dancers are assigned to

specific instruments and are choreographed in such a manner that they duplicate on stage the relationships among the instruments in the orchestra. Balanchine was said to have translated music into spatial terms, manipulating the floor patterns and the grouping of the dancers so that they corresponded to the appearance and development of particular chord sequences, rhythmic patterns, melodies, or sections of counterpoint. Nijinsky, on the other hand, in *L'Après-midi d'un faune* (1912; *Afternoon of a Faun*), used Debussy's music purely for atmosphere, permitting it to set the mood rather than influence the organization of movements.

Music setting the style and drama

Music can determine the style or dramatic quality of a dance. In fact, composers are often instructed to emphasize or clarify the drama already inherent in the choreography. In Western ballet it is common for important characters to have their own musical themes expressing and identifying their personalities or for whole sections of music to be written in the style of the character dancing to them—as in the sweet, tinkling music that Tchaikovsky composed for the Sugar Plum Fairy in *The Nutcracker*. In plotless dances music and movement also reflect and reinforce each other, as in Ashton's *Monotones* (1965–66), where the clear, uncluttered lines of the choreography reflect the limpidity of Satie's music.

Certain choreographers in the second half of the 20th century have worked either without music or in such a way that music and dance remain wholly independent of each other. Merce Cunningham choreographed in silence, so that while the music helped to determine the overall mood of the dance, it rarely affected the dance's phrasing and structure and often did not even last for the same length of time. Cunningham believed that too close a correspondence between dance and music would not really help the audience to perceive the two forms more clearly but, rather, would have the opposite effect of each canceling the other out. Other choreographers, such as Jerome Robbins in *Moves* (1959), used complete silence even in performance, so that the natural sounds of the dance movements formed the only accompaniment, leaving the spectator to concentrate solely on the patterns and rhythms of the movement. Others have used natural or electronic sounds and even spoken words in an effort to separate dance from a close relationship with music while still providing it with some relationship to sound.

*Union of dance and music.* It is likely that music accompanied dance from earliest times, either through sounds such as stamping, clapping, and singing that the dancers made themselves, through percussion, or through various wind instruments such as pipes or flutes. In modern Afro-Caribbean dances it is possible to discern the effects that drumming and percussive-sounding movements can have on dancing—in maintaining the dancer's beat, providing accompaniment, and intensifying the dance's emotional power. A slow, heavy beat can create a mood of tension or expectancy, while a fast beat may build the dance to a joyous or dramatic climax. The rhythms of the drums, reinforced by clapping and stamping, can amplify the rhythms of the movements (the sway of the pelvis, the rippling of the spine) as well as set up a complex counterpoint with them to produce variations in tempo and phrasing.

Clapping and stamping can also play an important role in producing the hypnotic effect necessary to certain ritual dances, uniting both spectators and dancers in a single world of sound and clearing their minds of everyday preoccupations. In the war and hunting dances of many tribes, sound is often used in an imitative way, with the dancers uttering war cries or animal sounds in order to further their transformation into warriors or the hunters' prey.

In many Indian and Asian classical dances, stamping also plays an important role in maintaining the beat. Music, too, is very important, and many dances are accompanied by specific songs or musical compositions. In the Middle Eastern *raqṣ sharqī,* the song or music establishes the mood or narrative situation of the dance, which the performer then interprets through movement. In the Indian bharata natya the dancer is accompanied by a singer, who marks the movements with a tiny pair of cymbals while singing out instructions to the dancer. Bells tied around the dancer's ankles also accompany the movements with their sound. Just as in Western theatre dance, the music accompanying these different dance forms is important both for its dramatic function—emphasizing moments of climax or different emotional states—and for its ability to increase the spectator's pleasure in and awareness of the movement.

Music enhancing awareness of movement

Social dance is nearly always accompanied by music, which not only helps to keep the dancers in time with each other but also increases the power and excitement of the dance, encouraging the dancers to abandon themselves to their movements. Sometimes individual dances have developed in response to a new musical form, as in jazz and rock and roll; but dance has also had an important influence on music, as in the Renaissance, when musicians were required to produce music to accompany the new dances that were developing.

Choreographers and composers alike often feel limited and frustrated when they have to create their own works within the limits of someone else's artistic conception. The most fruitful relationship is often one in which an element of collaboration exists between composer and choreographer from the start. Fokine's collaboration with Stravinsky on *The Firebird* (1910) is an example of both score and choreography emerging from long and detailed discussion, during which each artist remained sensitive to the other's wishes and to the overall idea of the work. There are no rules, however, and while some choreographers dislike being subjected to the limitations and demands of a musical score, others regard them as important creative stimuli.

Most dances have a traditional relationship with particular musical works or with particular kinds of music. Although ballet has always had a close relation to classical (as opposed to popular) music, many people have found unacceptable its use of established masterpieces that were not specially composed for ballet. It was not until the 20th century that this practice came into being, with Isadora Duncan performing to Wagner, Brahms, and Chopin and Léonide Massine choreographing his symphonic ballets to the works of Berlioz, Brahms, and Tchaikovsky.

During the 20th century a close relationship has also existed between modern dance and contemporary music, often music of a highly experimental nature. Thus, choreographers have used, or even commissioned, works from composers such as Schoenberg, Webern, Berio, Copland, and Cage. But it is common for both contemporary ballet and modern dance to use a variety of musical forms: modern dance may use early classical or non-Western music, while ballet may be performed to popular music. Also, as mentioned above, the concept of musical accompaniment has been stretched to include any kind of natural sound, electronic score, spoken text, or even silence.

**Set and design.** Just as music can enhance the mood of a dance and influence the way in which the spectator interprets its dramatic content, so visual elements such as costume, makeup, masks, props, lighting, and stage sets may also amplify certain qualities of dance movement. Because set and design are vital elements of theatre, they are most important in those types of theatre dance, whether dramatic or abstract, in which dancers perform before nonparticipating spectators. Therefore, most discussion of the use of visual elements in dance centres on theatre dance.

Such visual elements as costume and makeup do play a role in participatory social and ritual dances, however. In most war and hunting dances the participants not only imitate the movements of warriors or prey but also use weapons, masks, makeup, and animal skins to heighten the realism of the dance. The wearing of animal skins is a common means in many such dances to magically acquire the animals' strength or agility—hence the eagle feathers worn in the headdresses of many North American Indians or the deerskin shoes traditionally worn by the Scots.

In other ritual dances the dancers' clothes may well possess magical or religious significance. The Ṣūfī dancer begins his ritual by divesting himself of a black cloak that is symbolic of the tomb. Body painting in symbolic colours is characteristic of many tribal dances as a means of keeping away evil spirits, while the embroidery on a

number of European national costumes is often a relic from the days when it functioned as a magic charm. Most important of all, the wearing of special clothes in ritual dances, as in rituals not involving dance, is a way of signaling and preserving the sacred quality of the occasion and removing it from ordinary life.

In festive dances, too, clothes and ornamentation play an important role in embellishing the movement and heightening the atmosphere of gaiety, pomp, or excitement. Social dances frequently have special clothes associated with them—such as the evening suits and voluminous sequined dresses of ballroom dancing or the tight, black clothes of rock and roll. Such clothes are not only the fashion of the era but also the uniform that identifies the dancer more strongly with the dance and the other dancers. Like music, clothes can help dancers surrender their everyday selves to the dance.

Makeup and costume in Indian classical dance

In theatre dances everywhere, the use of visual effects is crucial to the power of the dance. In the Indian kathakali, facial makeup is central to the portrayal of character. Differently coloured beards are used to represent good or bad characters, while the colour of the makeup is even more revealing: a green and red painted face represents an evil and ferocious character, a green and white face is for heroes and noblemen, a pinkish-yellow face is for women characters and sages, and black and red makeup is used for female demons.

The bharata natya dancer relies more purely on the mudras for character portrayal, but makeup and costume are still highly important. The graceful, sinuous lines of the dancer's movements are emphasized by the bare torso and flowing skirt or trousers, while the intricate detail of the mudras is reflected in the rich jewels, flowers, and decoration of the costume.

*Costume and stage sets in Western theatre dance.* Masks have also been used as a means of characterization in many dance forms, from ancient Egypt to the early European court ballets. One reason early ballet dancers were limited in their dance technique was that the masks they wore to represent different characters were so elaborate and their wigs and clothes so heavy that it was scarcely possible to jump or to move across the floor with any speed or lightness.

The early ballets not only had elaborate costumes but also were performed in spectacular settings. *The Mountain Ballet,* performed in the early 17th century, had five enormous mountains as its stage scenery, in the middle of which was a "Field of Glory." The dance historian Gaston Vuillier later described the scene:

> Fame opened the ballet and explained its subject. Disguised as an old woman she rode an ass and carried a wooden trumpet. Then the mountains opened their sides, and quadrilles of dancers came out, in flesh coloured attire, having bellows in their hands, led by the nymph Echo, wearing bells for headdresses, and on their bodies lesser bells, and carrying drums. Falsehood hobbled forward on a wooden leg, with masks hung over his coat, and a dark lantern in his hand.

It was even known for ballets to be staged outdoors, with mock sea battles staged on artificial lakes.

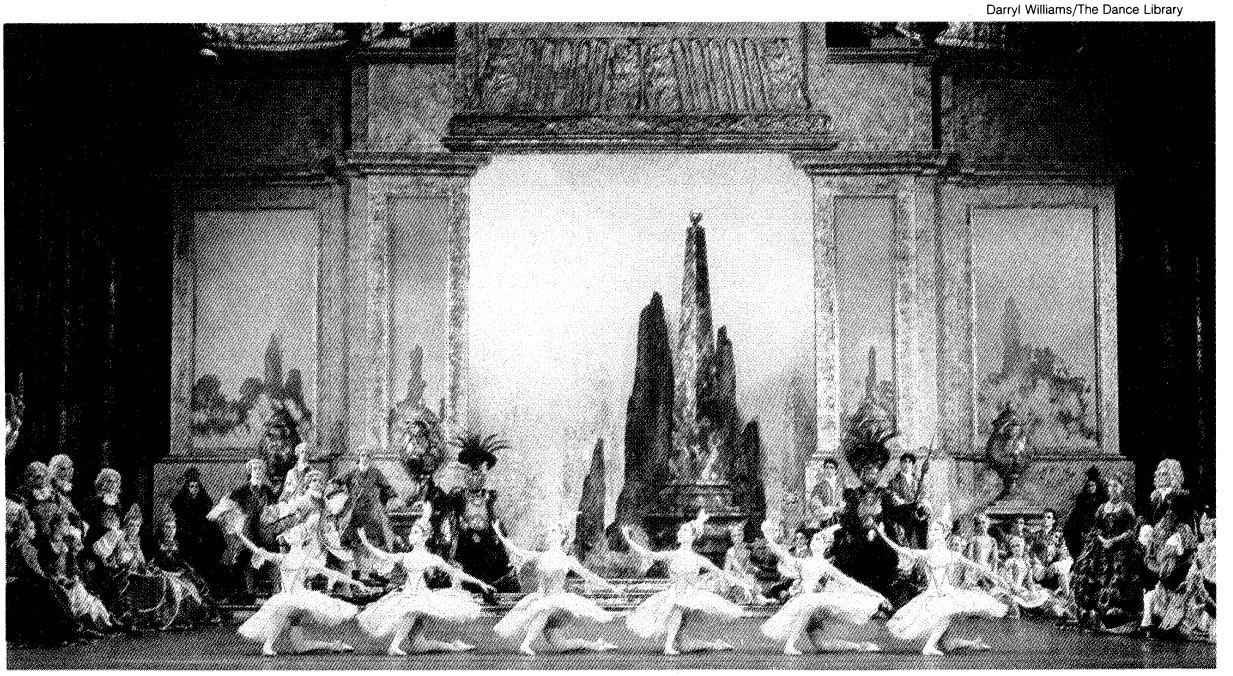
Gradually, as dancers shed their encumbering costumes and stage designs were simplified, dance movement and mime became more important in the depiction of plot and character. Set design and costume were tailored to the ballet's theme and atmosphere, rather than swamping the choreography with their elaborate opulence. The development of gas lighting meant that magical effects could be created with simple painted scenery, and though wire contraptions were sometimes used to fly the ballerina (as a sylph or bird) across the stage, the development of pointe work (dancing on the toes) meant that the dancer could appear weightless and ethereal without any artificial aids. In place of highly decorative mythological or classical scenes, there were poetic evocations of landscape, and the ballerinas were either dressed in simple white dresses or in colourful national dress. The poet, critic, and librettist Théophile Gautier described the typical "white" or ethereal Romantic ballet as follows:

> The twelve marble and gold houses of the Olympians were relegated to the dust of the storehouse and only the romantic forests and valleys lit by the charming German moonlight of Heinrich Heine's ballads exist.... This new style brought a great abuse of white gauze, of tulle and tarlatans and shadows melted into mist through transparent dresses. White was almost the only colour used.

This unity of dance and design was not to last, however. By the end of the 19th century most of the productions mounted at the Mariinsky Theatre in St. Petersburg were lavish spectacles in which set and costume had little relevance to the ballet's theme, being designed simply to please the audiences' taste for opulence. At the beginning of the 20th century one of the first revolutionary steps that Michel Fokine took in trying to change this state of affairs was to dress his dancers in costumes as nearly authentic as possible—for example, by replacing the prevailing tutu

Fokine and the unity of dance and design

Darryl Williams/The Dance Library



Palace set design for *The Sleeping Beauty,* with painted scenery and period costumes for the members of the court; Sadler's Wells Royal Ballet.



Four renowned ballerinas in classical ballet costumes, 1845; (left to right) Carlotta Grisi, Marie Taglioni, Lucile Grahn, and Fanny Cerrito.

By courtesy of the Theatrical Museum of La Scala, Milan

with clinging draperies (as in the Egyptian costumes for *Eunice* [1908]) and by dispensing with the dancers' shoes. (Actually, the theatre management did not allow dancers to go barefoot, but they had red toenails painted onto their tights to achieve the same impression.)

This move was part of Fokine's general commitment to the idea that movement, music, and design should be integrated into an aesthetic and dramatic whole. His collaboration with designers such as Léon Bakst and Alexandre Benois were as important as his musical collaboration with Stravinsky. Sets and costumes not only reflected the period in which the ballet was set but also helped to create the dramatic mood or atmosphere—as in *Le Spectre de la rose* (1911; "The Spirit of the Rose"), where the exquisite rose-petaled costume of the spectre, or spirit, seemed almost to emit a magical perfume, and where the simple naturalism of the sleeping girl's bedroom emphasized her dreaming innocence.

In the newly emerging modern dance, experiments with set, lighting, and costume design were also significant. One of the pioneers in this field was Loie Fuller, a solo dancer whose performances in the 1890s and early 1900s consisted of very simple movements with complex visual effects. Swathing herself in yards of diaphanous material, she created elaborate shapes and transformed herself into a variety of magical phenomena. These illusions were enhanced by coloured lights and slide projections playing across the floating material.

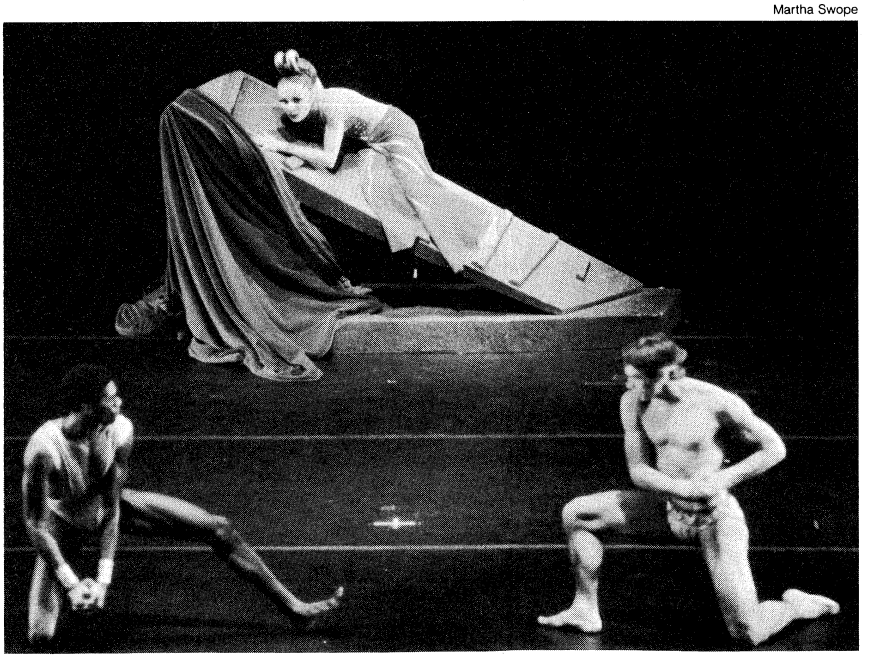
Elaborate lighting and costumes were also used by Ruth Saint Denis, whose dances frequently evoked ancient and exotic cultures. At the opposite extreme Martha Graham, who began her career as a dancer with Saint Denis' company, strove to eliminate all unnecessary ornamentation in her designs. Costumes were made out of simple jersey and cut along stark lines that clearly revealed the dancers' movements. Simple but dramatic lighting suggested the mood of the piece. Graham also pioneered the use of sculpture in dance works, replacing painted scenery and elaborate props with simple, free-standing structures. These had a number of functions: suggesting, often symbolically, the place or theme of the work; creating new levels and areas of stage space; and also illuminating the overall design of the piece.

Modern trend toward minimal visual effects

While it has remained common for choreographers to use elaborately realistic sets and costumes, as in Kenneth MacMillan's *Romeo and Juliet* in 1965, most choreographers have tended to adopt a minimal approach, with costumes and scenery simply suggesting the ballet's characters and location rather than representing them in detail. One reason for this development has been the move away from narrative to plotless, or formal, works in both ballet and modern dance, where there is no longer any need for visual effects to provide narrative background. Balanchine set many of his works on a bare stage with the dancers dressed only in practice costumes, feeling that this would allow the spectators to see the lines and patterns of the dancers' movements more clearly.

*Effects of design in theatre-dance performance.* Set, costume, and lighting design are important in narrative as well as formal dance in helping the audience maintain the special attention that theatre demands. They can also influence strongly the way in which the choreography is perceived, either by creating a mood (sombre or festive, depending on the colour and ornamentation used) or by strengthening a choreographic image or concept. In Richard Alston's *Wildlife* (1984) the geometrically shaped kites suspended from the flies actually inspired some of

Martha Swope



Set for *Phaedra's Dream*, by Martha Graham, for which there is no scenery except the single free-standing structure that is incorporated into the dance.

the dancers' sharply angled movements as well as making them visually more striking in performance.

Costume, set design, and the perception of movement

Costume, too, can alter the appearance of movement: a skirt can give fuller volume to turns or to high leg extensions, while a close-fitting leotard reveals every detail of the body's movements. Some choreographers, trying to emphasize the nontheatrical or nonspectacular aspects of dance, have dressed their dancers in ordinary street clothes in order to give a neutral, everyday look to their movements, and they have often dispensed entirely with set and lighting.

Set design and lighting (or their absence) can help to frame the choreography and to define the space in which it appears. The space in which a dance occurs has, in fact, a crucial influence on the way movement is perceived. Thus, a small space can make the movement look bigger (and possibly more cramped and urgent), while a large space can lessen its scale and possibly make it appear more remote. Similarly, a cluttered stage, or one with only a few lighted areas, may make the dance appear compressed, even fragmented, while a clearly lighted, open space may make the movement appear unconfined. Two choreographers who have been most innovative in their use of set and lighting are Alwin Nikolais and Merce Cunningham. The former has used props, lighting, and costumes to create a world of strange, often inhuman shapes—as in his *Sanctum* (1964). The latter has often worked with sets that almost dominate the dancing, either by filling the stage with a clutter of objects (some of which are simply things taken from the outside world, such as cushions, television sets, chairs, or bits of clothing) or—as in *Walkaround Time* (1968)—by using elaborate constructions around which the dance takes place, often partly concealed. As with his use of music, Cunningham's sets have often been conceived independently of the choreography and are used to create a complex visual field rather than to reflect the dancing.

Separating dance from everyday life

Perhaps the most important influence on the way spectators perceive dance is the place in which it is performed. Religious dances usually take place within sacred buildings or on sacred ground, thus preserving their spiritual character. Most theatre dance also occurs in a special building or venue, heightening the audience's sense that it has entered a different world. Most venues create some kind of separation between the dancers and the audience in order to intensify this illusion. A theatre with a proscenium stage, in which an arch separates the stage from the auditorium, creates a marked distance. Performance in the round, in which the dancers are surrounded by spectators on all sides, probably lessens both the distance and the illusion. In dance forms that do not traditionally take place in a theatre, such as Afro-Caribbean dance, the intimacy between audience and dancer is very close, and the former may often be called upon to participate.

The theatre space not only influences the relationship between the audience and the dancer but is also closely related to the style of the choreography. Thus, in the early court ballets, spectators sat on three sides of the dancers, often looking down at the stage, because the intricate floor patterns woven by the dancers, rather than their individual steps, were important. Once ballet was introduced into the theatre, however, dance had to develop in such a way that it could be appreciated from a single, frontal perspective. This is one reason turned-out positions were emphasized and extended, for they allowed the dancer to appear completely open to the spectators and, in particular, to move sideways gracefully without having to turn away from them in profile.

Many modern choreographers, wishing to present dance as part of ordinary life and to challenge the way in which people look at it, have used a variety of nontheatrical venues to dispel the illusion or glamour of the performance. Choreographers such as Meredith Monk, Trisha Brown, and Twyla Tharp, working in the 1960s and '70s, performed dances in parks, streets, museums, and galleries, often without publicity or without a viewing charge. In this way dance was meant to "happen" among the people instead of in a special context. Even the most surprising or nonglamorous venue, however, cannot entirely dispel the sense of distance between dancer and audience and between dance and ordinary life.

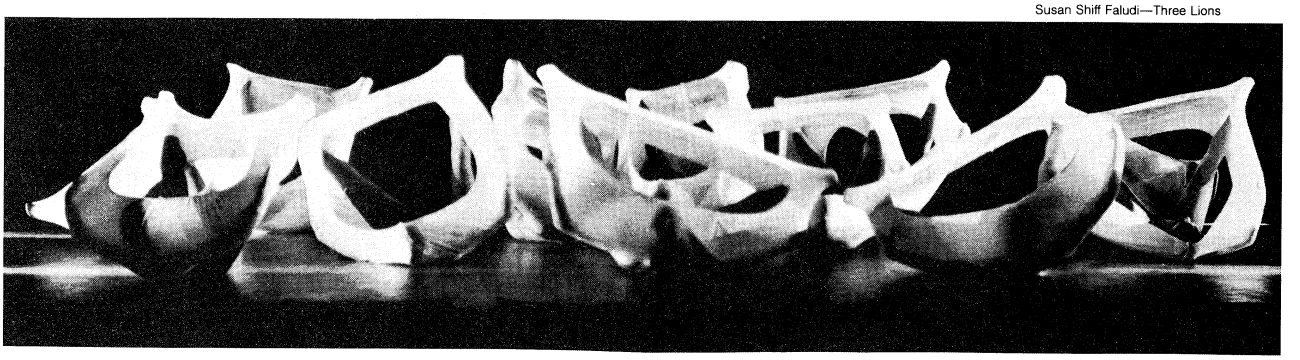
**Drama.** Throughout history there has been a rough division between dramatic dance, which expresses or imitates emotion, character, and narrative action, and purely formal dance, which stresses the lines and patterns of movement itself (see above *Dance as dramatic expression or abstract form*). The type and function of dramatic dance vary considerably, including full-length theatrical works (in which dance is used to tell a story and present specific characters), hunting dances (in which the dancers' movements imitate those of a particular animal), and courtship dances (which may contain only a few pantomimic gestures, such as a lift, a curtsy, or a mock kiss, to convey meaning).

The expressive quality of dance movement

Because dance movements are often closely related to everyday forms of physical expression, there is an expressive quality inherent in nearly all dancing. This quality is used extensively in dramatic dance to communicate action or emotion—for example, the aggression in stamping movements, the exhilaration communicated by jumping, and the dragging motions of despair. Mime, or narrative gesture, is also used. Mime can either imitate movement realistically—in a death scene, for example, where the killer assumes a ferocious expression and imitates strangling a victim—or it can function as a symbol—as in the circling movement of the arms in ballet to represent dancing or in pointing to the fourth finger to represent marriage. Dance movements are often accompanied by other elements, such as masks, costume, music, acting, singing, recitation, and even film, to help communicate the dramatic content.

*Cultural distinction between dramatic and formal dance.* Curt Sachs argued that the division between dramatic and formal dance in tribal cultures followed the division between hunting and planter cultures. While the accuracy of his claim may be hard to establish, it can help to illuminate the different types and function of dance that lie at the root of such a division. In hunting dances (and war dances as well) the dancers' movements are dramatically charged, expressing a state of excitement or aggression and frequently imitating the movements of animals or fighting men, even to the point of manipulating weapons.

Susan Shiff Faludi—Three Lions



*Sanctum.* Choreography, sound score, costumes, decor, and lighting by Alwin Nikolais; Alwin Nikolais Dance Company.

Imitative sounds increase the power of the illusion, as does the wearing of masks, makeup, or animal skins. The effect on both dancer and spectator is to be drawn into a fictional world, in which the dancers become the people or animals that they represent and the story or situation enacted by the dance takes on an immediate reality. Any successful dramatic dance should, in fact, produce this effect, even if the dancers do not actually feel the emotions they are representing or the spectators respond as if the imitation were real.

In the dances of planter cultures, Sachs argued, the movements tend to be smaller and not directly imitative. The groupings of the dancers and the floor patterns traced by their steps, on the other hand, tend to be much more complex and ordered. In addition, the sequence of movements tends to be more repetitive and the dancers' movements are more uniform. Such formal dances are often performed as part of a ritual propitiation of the gods in order to assure good weather and successful harvests. Although their movements may not be imitative, the repetitive patterns often represent such natural occurrences as the cycle of the seasons, the waxing and waning of the Moon, and the growing of vegetation, and they even evoke more abstract entities such as space and time.

Effects of formal dance movement

The effect may thus be one of fusing the dancers and spectators with some aspect of the natural world. At the same time the dance may produce an effect similar to the repetitive chanting of prayer or meditation, emptying the mind of its usual preoccupations and focusing it on the object of worship. In fact, the power of dance in achieving this type of spiritual discipline is peculiarly strong, since the repetitive movements work kinesthetically as well as aurally and visually. As a consequence, mind and body are equally absorbed into the ritual.

Even where formal dances are not part of a ritual (as in modern plotless dance works), the movement of the dancers may produce an effect not dissimilar to that described above. Space, time, and the force of gravity may be made apparent to the spectator through the trajectories that the dancers make in space, through the configurations that they form on the dance floor, through the duration of the dance phrases, and through the alternating sensations of weight and weightlessness created by falls and jumps. In a similar way, too, the audience may experience a special focusing of attention, a draining of the usual habits of perception through the kinesthetic, visual, and aural power of the movement and music.

Many extant tribal dances can be categorized as either imitative or formal, as can the European folk dances that developed out of earlier tribal dance forms. Courtship dances, the descendants of ancient courtship and fertility dances, still retain overt imitations of flirtatiousness. Other dances have similarly retained their early formal character, even, in some cases, retaining the symbolic significance of their patterns. In Ukrainian dances descended from pagan Moon-worshiping ritual, the circling of the dancers represents the way the Moon influences the work in the fields, and the final pivot represents the flourishing of the corn. In Armenian carpet-weaving dances, the complex floor patterns mimic the action of the work process.

*Drama in Western theatre dance.* When dance developed into a form of spectacle, particularly of a secular kind, it was frequently allied to the telling of a story and the depiction of characters. Mimed gesture was often prominent in such dance dramas—for example, in ancient Greece, where the gestures of the chorus illustrated the drama's major themes. There the mime was often naturalistic: a hand on the head to represent grief or the stretching upward of the arms to express worship. During the later, cosmopolitan period of the Roman Empire, dance and mime were popular entertainment for audiences drawn from a variety of linguistic backgrounds. The highly sophisticated pantomime used by these dancers formed the basis of the improvised mime drama of the 16th-century Italian commedia dell'arte and, later, the techniques of 20th-century mime artists such as Marcel Marceau.

The early European court ballets were also oriented toward dramatic spectacle, though the dance movement itself was not highly expressive and mimed gesture was limited. Other dramatic elements, usually visual effects or speech, communicated the essential points of the story. One of the first choreographers to extend dance movement so that it could be dramatically expressive was the Englishman John Weaver, who in his ballet *The Loves of Mars and Venus* (1717) experimented with giving the characters gestures to express their individual personalities. Later in the 18th century Jean-Georges Noverre reacted against the purely decorative form into which ballet had developed. He believed that mime should be as close to natural gesture as possible and that dance movement should not be meaninglessly decorative but should reflect the ballet's action.

Mime in ballet

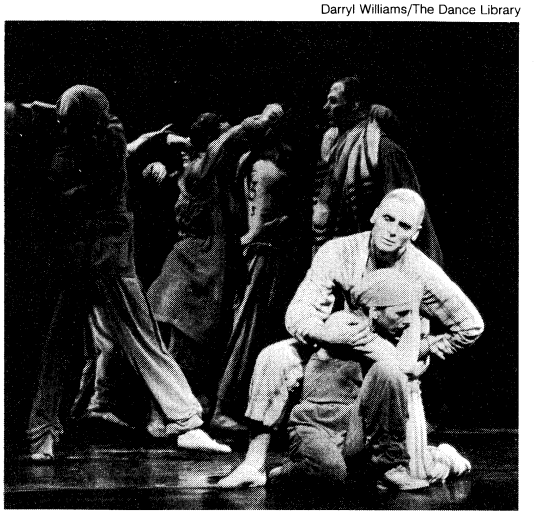
Noverre's ideas were partly realized in the Romantic ballet of the early 19th century, which strove to give movement a greater poetic expressiveness. Developments in dance technique, notably that of dancing *en pointe*, gave dancers a wider range of movement to express character and action, although conventional or symbolic mime was also used to tell parts of the story. By the end of the century, however, choreography was once again seldom concerned with plot and character, and long sections of mime (often incomprehensible even to the dancers) were used to tell whatever story there was in the dance. The reforms proposed by Fokine at the beginning of the 20th century, like those of Noverre two centuries before, demanded more naturally expressive mime and dance movement that illuminated theme and character and were an essential component of the dance.

Fokine's own work reflected these ideas faithfully. He experimented with angular movement reminiscent of archaic Greece in *Daphnis et Chloé* (1912; *Daphnis and Chloe*), developed individual styles for different characters (such as the jerky wooden movements of the puppet Petrushka), and brought mime much closer to natural gesture than the symbolic code previously used. This naturalism still characterizes ballet; the expressive qualities of dance movement and simple, dramatic gestures almost entirely displace conventional mime, and even in revivals of the 19th-century classics, traditional mime is usually kept to a minimum so that audiences have no trouble following it.

The founders of modern dance, Isadora Duncan, Mary Wigman, Martha Graham, and Doris Humphrey, also reacted against the lack of expression in ballet. Like Fokine, they believed that most ballet dancing was mere decorative acrobatics, but while Fokine was happy to continue using exotic or archaic themes for his new, naturalistic ballets, these later choreographers believed that dance should address subjects of greater relevance and profundity. The kinds of movement with which the modern dance choreographers expressed these themes had little of conventional ballet technique about them. Eschewing mime, particularly that associated with ballet, as well as

Dramatic concerns of modern dance

Darryl Williams/The Dance Library



A dramatic moment in *Valley of Shadows*, a ballet by Kenneth MacMillan about life in a Nazi concentration camp; Sadler's Wells Royal Ballet.

the traditional ballet vocabulary, they sought to make the whole body dramatically expressive. (See below *Theatre dance: Modern dance.*)

During the 20th century, ballet, like modern dance, moved toward a concern with more serious issues. In works such as Antony Tudor's *Jardin aux lilas* (1936; *The Lilac Garden*), Peter Darrell's *Prisoners* (1957), Gerald Arpino's *Clowns* (1968), and Kenneth MacMillan's *My Brother, My Sisters* (1978), choreographers engaged subject matter ranging from emotional and psychological conflict to war and social issues.

The avant-garde

In the avant-garde dance of the 1970s and '80s, experiments were made in expanding narrative potential by incorporating nondance elements (almost turning full circle back to the early court ballets). At times dance was accompanied by mime, acting, and singing as well as a multitude of visual effects. In some cases choreographers collaborated with artists working in other forms, such as music, drama, and the visual arts, and they thought of dance less as a single discipline than as a broadly based theatre art. Most of these experimental works had some kind of dramatic or conceptual content, although they avoided conventional forms of narration and expression. Events were rarely presented in chronological order, and the distinction between reality, symbolism, and fantasy was often blurred.

## Types of dance

The division of dance into types can be made on many different grounds. Function (*e.g.,* theatrical, religious, recreational) is an obvious ground, but distinctions can also be made between tribal, ethnic, and folk dance, between amateur and professional, and above all between different genres and styles.

Defining genre and style

Genre and style are relatively ambiguous terms. They depend on analyses of movement style, structure, and performance context (*i.e.,* where the dance is performed, who is watching, and who is dancing) as well as on a cluster of general cultural attitudes concerning the role and value of dance in society. Genre usually refers to a self-contained formal tradition such as ballet, within which there may be further subgenres, such as classical and modern ballet. (Some critics consider modern dance as an independent genre with a subgenre of postmodern dance, but others subsume both categories under ballet, along with other theatre dance forms such as jazz.) Within subgenres, different styles can be distinguished, such as those of Ashton, MacMillan, and Balanchine in modern ballet and Graham and Cunningham in modern dance. Style as used here embraces many elements, including a preference for certain kinds of movement (fast, slow, simple, or intricate) or for particular kinds of energy and attack (sharp, edgy, and hard, as opposed to soft and fluid). It also embraces different ways of phrasing movement or of arranging dancers into groups, as well as an interest in certain kinds of music or design.

Perhaps the most obvious division between types is that between theatre and non-theatre dance. The separation of dancer and spectator in theatre dance has tremendous influence on the style of the dance itself and on its reception as an art form. In theatre dance the professionalism of dancer and choreographer, the presentation of dramatic and formal movement, the use of visual effects, and even the philosophical question of the role of the spectator reach their most sophisticated level. In non-theatre dance the unity of dancer and spectator, of observation and participation, means that the dance styles and even the function within the social group are quite different from those of theatre dance.

Of course, the division between the two types is not as clear in practice as in theory. For example, although ethnic and folk dances are not, in theory, theatre dances, many of them are preserved, choreographed, and presented to audiences in theatrical settings. Some scholars have even argued that ballet is an ethnic dance form that grew out of the European dance tradition. On the other hand, Indian and Southeast Asian dance forms are usually called ethnic dances, but within these traditions there are numerous classical dances whose theatrical settings and elaborate choreographies qualify them as theatre dances.

Among the theatre dances, this section discusses the two major Western genres, ballet and modern dance, as well as Indian classical dance. Among the non-theatre dance forms, this section discusses tribal and ethnic dance, folk dance, and social dance.

### THEATRE DANCE

**Ballet.** *Basic characteristics.* Ballet has been the dominant genre in Western theatre dance since its development as an independent form in the 17th century, and its characteristic style of movement is still based on the positions and steps developed in the court dances of the 16th and 17th centuries (see above *Basic steps and formations*). Perhaps the most basic feature of the ballet style is the turned-out position of the legs and feet, in which the legs are rotated in the hip socket to an angle of 90 degrees and the feet point outward. This position gives the body an open, symmetrical appearance, and it also facilitates the high leg extensions and fast turns typical of ballet. Openness is most characteristic of the ballet dancer's stance, for the head is nearly always lifted and the arms held out in extended curves. Even when the dancer executes fast or energetic steps, the arms rarely move in a way that is not fluid, calm, and gracefully extended, and they are frequently held in positions that either frame the face or form a harmonious relation to the position of the legs. The body is nearly always held erect, apart from controlled arches in the back or a slight turning of the shoulders toward or away from the working leg. This positioning of the shoulders, called épaulement, gives a sculpted, three-dimensional quality to the dancer's positions.

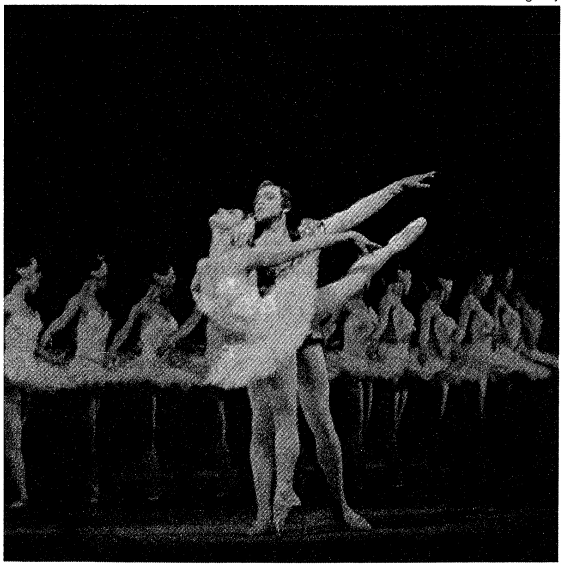
The open stance

Whenever the ballet dancer's foot is not flat on the floor, it is pointed, and, of course, women dancers (and occasionally men) frequently dance on the tips of their toes with the aid of blocked shoes. Dancing *en pointe* lends lightness and airiness to the dancer's movements, and the pointed toe extends the line of the leg—particularly when it is raised in the air, as in an arabesque.

The vertical line

The long vertical line is the other basic feature of ballet: verticality in the upright stance of the body, in the high leg extensions, and above all in the aerial quality of the movement. Ballet dancers rarely move close to the ground, and they frequently seem to defy gravity through the height of their jumps and their vigorous batterie (beating together of the legs in midair), through the speed and multiplicity of their turns, and through the fast linking steps that seem to move them effortlessly, almost without touching the floor, from one virtuoso movement to another. In ballet the stress and effort of the dancer's movements are always concealed beneath the fluid, graceful surface of the dance

Novosti Press Agency



The Bolshoi Ballet performing the classical ballet *Swan Lake;* choreography by Marius Petipa and Lev Ivanov.

and the perfect repose of the face and torso. This gives the dance its characteristic qualities of dignity and control, which it inherited from the early court ballets, where the movements were designed to show off the aristocratic polish of the dancers.

Ballet has, of course, undergone many stylistic alterations. The Romantic style of the early to mid-19th century was much softer—less studded with virtuosic jumps and turns—than the classical style of the late 19th and early 20th centuries. Russian ballet, frequently regarded as the paradigm of the classical school, is itself a blend of the soft and decorative French school, the more brittle and virtuosic style of the Italians, and the vigorous athleticism of Russian folk dances.

The design of classical ballet is traditionally symmetrical in the shapes made by the dancers' bodies, in the grouping of the dancers on stage, and even in the structure of the whole dance. For example, if two principal dancers perform a pas de deux (a dance for two), other dancers on stage remain still, are arranged in symmetrical framing patterns, or move in harmony with the main dancing, not distracting from it. Even when large groups of dancers move, they are usually arranged in formal lines or circles. Jean-Georges Noverre in the 18th century and Michel Fokine in the first decades of the 20th both argued that this kind of formal symmetry was detrimental to the dramatic naturalism of ballet. Fokine's own choreography encouraged the use of less rigid and artificial groupings in later ballet, as in Kenneth MacMillan's dramatic works, where the crowd scenes are often composed of asymmetrical groups that rarely seem artificial.

*Innovations in the 20th century.* Fokine's reforms were a major influence on the development of 20th-century ballet. Particularly in the works that he created for Sergey Diaghilev's company, the Ballets Russes, he showed the range of different dance styles that classical ballet was capable of absorbing, helping to pave the way for more radical innovation. For example, in *Chopiniana* (1908; later called *Les Sylphides*), a virtually plotless ballet that recalled the earlier Romantic tradition, Fokine created a soft and uncluttered style that contained no bravura feats of jumping, turning, or batterie. Arm movements were simple and unaffected, the grouping of the dancers had a fluid, plastic quality, and above all there was a flowing, lyrical line in the phrasing and movement.

Fokine's non-ornamental style

In *Prince Igor* (1909) and *L'Oiseau de feu* (1910; *The Firebird*) Fokine incorporated the vigorous style and athletic steps of Russian folk dances. These works revealed his talent for organizing large crowds of dancers on stage and transforming their previously ornamental function into a powerful dramatic force. Neither ballet is longer than a single act, because Fokine believed that the full-length ballet was generally both an excuse for, and the cause of, useless choreographic padding, and that a work should last only as long as its theme required.

For all its stylistic variations, Fokine's choreography was couched largely in the classical idiom. Two other choreographers working with the Ballets Russes, Vaslav Nijinsky and his sister Bronislava Nijinska, produced works of a more radical nature. In *Jeux* (1913; "Games"), Nijinsky was one of the first choreographers to introduce a modern theme and modern design into ballet. Based on his own (rather erroneous) idea of a tennis match, the choreography incorporated sporting movements and dancers in modern dress. In *The Rite of Spring,* perhaps Nijinsky's most innovative work, the dancers were arranged in massed groupings and executed harsh, primitive movements, the legs turned in, the arms hanging heavily, and the heads lolling to one side. Unlike Fokine, Nijinsky was prepared to risk ugliness in his search for a truly authentic style, and the audiences were almost as deeply shocked by the choreography as by the discordant sounds and jagged rhythms of Stravinsky's score.

Nijinsky's *Rite of Spring*

In her ballet *Les Noces* (1923; "The Wedding"), which took its theme from the marriage ceremonies of Russian peasants, Nijinska created a stark and heavily weighted style of movement. There were few elevations, and the dancers were frequently crouched or bent over, with their heads hanging low to the floor. They were also arranged in large groups, so that the overall impression was less that of individual bodies moving together than of large shapes and blocks of movement.

Although there are similarities between the works of these choreographers and the modern-dance forms that emerged in the 1920s and '30s, there is little evidence to suggest any direct influence. The major significance of Fokine, Nijinsky, and Nijinska was in their bringing ballet out of its remote, courtly past by using modern themes and subjects and by introducing modern intellectual and artistic influences into the classical art form.

The style of later 20th-century ballet was influenced not only by the Ballets Russes but by modern dance as well. It became common for choreographers to extend the traditional ballet vocabulary with modern-dance techniques, such as curving and tilting the body away from the vertical line, working on or close to the floor, and using turned-in leg positions and flexed feet. Balanchine, influenced by jazz, used syncopated rhythms in his phrasing and incorporated steps from such popular dances as ragtime and rock and roll. His movements were usually wide, almost exaggerated in shape and volume, and frequently characterized by speed and by hard, clear accents.

Influence of other genres

Despite these changes ballet retains significant traces of its courtly and classical past. Although there are exceptions, such as those noted above, ballet dancers still tend to dance in the calm, erect, and dignified manner of their aristocratic forebears. Illusion and spectacle remain important; nearly all works are performed on a proscenium stage in a large theatre, where the performers are distanced from the audience, and productions are frequently elaborate and expensive. Ballet companies still, therefore, tend to be large organizations, receiving some kind of patronage or state subsidy.

**Modern dance.** *Expressionism.* Modern dance, the other major genre of Western theatre dance, developed in the early 20th century as a series of reactions against what detractors saw as the limited, artificial style of movement of ballet and its frivolous subject matter. Perhaps the greatest pioneer in modern dance was Isadora Duncan. She believed that ballet technique distorted the natural movement of the body, that it "separated the gymnastic movements of the body completely from the mind," and that it made dancers move like "articulated puppets" from the base of the spine. Duncan worked with simple movements and natural rhythms, finding her inspiration in the movements of nature—particularly the wind and waves—as well as in the dance forms that she had studied in antique sculpture. Elements that were most characteristic of her dancing included lifted, far-flung arm positions, an ecstatically lifted head, unconstrained leaps, strides, and skips, and, above all, strong, flowing rhythms in which one movement melted into the next. Her costumes, too, were unconstrained; she danced barefoot and uncorseted in simple, flowing tunics, with only the simplest props and lighting effects to frame her movements.

The style of Isadora Duncan

Duncan believed that dance should be the "divine expression" of the human spirit, and this concern with the inner motivation of dance characterized all early modern choreographers. They presented characters and situations that broke the romantic, fairy-tale surface of contemporary ballet and explored the primitive instincts, the conflicts and passions of man's inner self. To this end they sought to develop a style of movement that was more natural and more expressive than ballet. Martha Graham, for example, saw the back, and particularly the pelvis, as the centre of all movement, and many of her most characteristic movements originated from a powerful spiral, arch, or curve in the back. Doris Humphrey saw all human movement as a transition between fall (when the body is off-balance) and recovery (when it returns to a balanced state), and in many of her movements the weight of the body was always just off-centre, falling and being caught.

Instead of defying gravity, as in ballet, modern dancers emphasized their own weight. Even their jumps and high extensions looked as if they were only momentarily escaping from the downward pull of the Earth, and many of their movements were executed close to, or on, the floor. Graham developed a wide repertoire of falls, for example,

Emphasizing gravity

and Mary Wigman's style was characterized by kneeling or crouching, the head often dropped and the arms rarely lifted high into the air.

As ballet sought to conceal or defy the force of gravity, so it also strove to conceal the strain of dancing. Modern dance, on the other hand—particularly the work of Graham—emphasized those qualities. In the jagged phrases, angular limbs, clenched fists, and flexed feet, in the forceful movements of the back and the clear lines of tension running through the movement, Graham's choreography expressed not only the struggle of the dancer against physical limitations but also the power of passion and frustration. Movements were always expressive gestures, never decorative shapes. Often the body and limbs appeared racked and contorted by emotion, for these choreographers, like Nijinsky, were not afraid to appear ugly (as indeed they did to many of their contemporaries).

The structure of early modern dance works responded in part to the fragmented narrative and symbolism characteristic of modernist art and literature. Graham often employed flashback techniques and shifting time scales, as in *Clytemnestra* (1958), or used different dancers to portray different facets of a single character, as in *Seraphic Dialogue* (1955). Groups of dancers formed sculptural wholes, often to represent social or psychological forces, and there was little of the hierarchical division between principals and corps de ballet that operated in ballet.

*Merce Cunningham.* The Expressionist school dominated modern dance for several decades. From the 1940s onward, however, there was a growing reaction against Expressionism spearheaded by Merce Cunningham. Cunningham wanted to create dance that was about itself—about the kinds of movement of which the human body is capable and about rhythm, phrasing, and structure. Above all, he wanted to create dance that was not subservient to the demands of either narrative or emotional expression. This did not mean that Cunningham wanted to make dance subservient to music or design; on the contrary, though many of his works were collaborations, in the sense that music and design formed a strong part of the total effect, these elements were often conceived—and worked—independently of the actual dance. Cunningham believed that movement should define its own space and set its own rhythms, rather than be influenced by the set and the music. He also felt that it was more interesting and challenging for the spectator to be confronted with these independently functioning elements and then to choose for himself how to relate them to one another.

Cunningham's balletic style

Believing that all movement was potential dance material, Cunningham developed a style that embraced an extraordinarily wide spectrum, from natural, everyday actions such as sitting down and walking to virtuosic dance movements. Elements of his style even had a close affinity to ballet: jumps tended to be light and airy, the footwork fleet and intricate, and the leg extensions high and controlled. He placed greater emphasis on the vertical and less emphasis on the body's weight and the force of gravity. Like those of Graham, many of Cunningham's movements centred on the back and torso, but they tended less toward dramatic contractions and spirals than toward smaller and more sharply defined tilts, curves, and twists. The arms were frequently held in graceful curves and the feet pointed.

Cunningham's phrases were often composed of elaborate, coordinating movements of the head, feet, body, and limbs in a string of rapidly changing positions. The arrangement of performers on stage was equally complex: at any one moment there might have been several dancers, in what seemed like random groupings, all performing different phrases at the same time. With no main action dominating the stage, the spectator was free to focus on any part of the dance.

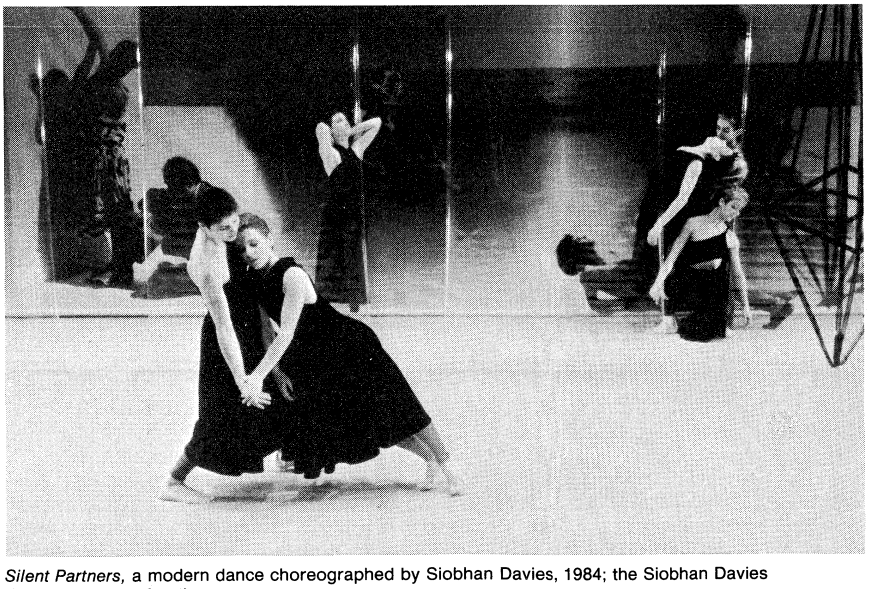
While Graham's works were usually structured around the events of a narrative, Cunningham's works usually emerged from the working through of one or more choreographic ideas, whose development (*i.e.*, the ordering of movements into phrases or the number of dancers working at any one time) might then have been determined by chance. Deriving its movements from such formal origins did not mean that Cunningham's works lacked expressive power. One of his pieces, *Winterbranch* (1964), started out as a study based on moving into a space and falling, but it produced a powerful effect on audiences, who variously interpreted it as a piece about war, concentration camps, or even sea storms. Cunningham believed that the expressive qualities in dance should not be determined by a story line but should simply rise out of movement itself. "The emotion will appear when the movement is danced," he claimed, "because that is where the life is."

*Postmodernism.* During the 1960s and '70s a new generation of American choreographers, generally referred to as postmodernist choreographers, took some of Cunningham's ideas even farther. They also believed that ordinary movement could be used in dance, but they rejected the strong element of virtuosity in Cunningham's technique and the complexities of his phrasing and structure, insisting that such a style interfered with the process of seeing and feeling the movement clearly. Consequently, the postmodernists replaced conventional dance steps with simple movements such as rolling, walking, skipping, and running. Their works concentrated on the basic principles of dance: space, time, and the weight and energy of the dancer's body.

The focus on movement

Postmodernists discarded spectacle as another distraction

Darryl Williams/The Dance Library



*Silent Partners*, a modern dance choreographed by Siobhan Davies, 1984; the Siobhan Davies dance troupe performing.

from the real business of movement. Costumes were often ordinary practice or street clothes, there was little or no set and lighting, and many performances took place in lofts, galleries, or out-of-doors. Narrative and expression were discarded, and the dance structures were usually very simple, involving either the repetition and accumulation of simple phrases or the working through of simple movement games or tasks. In Tom Johnson's *Running Out of Breath* (1976) the dancer simply ran around the stage reciting a text until he ran out of breath.

Most avant-garde modern-dance companies have been quite small and have occupied a position on the fringe of the dance world, attracting only small and specialist audiences. Although "mainstream" modern dance now attracts large audiences in both Europe and North America, it too was for many decades a minority art form, often playing to only a handful of spectators. Modern-dance companies were then, and still are, relatively small. Partly because they lack funding, they tend to use less elaborate costume and staging, and they perform in small theatres where contact with the audience is close.

*The musical.* Perhaps the most genuinely popular of all the subgenres within ballet and modern dance are the dance forms associated with the musical, such as tap, jazz, ballroom, and disco. In musicals of both stage and screen, dance is an essential ingredient along with song, acting, and spectacle. Although the dancing is often mechanical and unoriginal, in the work of such dancers and choreographers as Fred Astaire and Gene Kelly it can rise to the status of a genuine art form. Sometimes, as in Jerome Robbins' choreography for the dances of the rival gangs in *West Side Story* (1957), it creates a powerful dramatic effect. Other innovative choreographers include Agnes deMille, whose dances in *Oklahoma!* (1943) were the first ever used to advance the plot, and Bob Fosse, particularly known for his work on the film *All That Jazz* (1979).

**Indian classical dance.** The six recognized schools of Indian classical dance developed as a part of religious ritual in which dancers worshiped the gods by telling stories about their lives and exploits. Three main components form the basis of these dances. They are natya, the dramatic element of the dance (*i.e.,* the imitation of character); nritta, pure dance, in which the rhythms and phrases of the music are reflected in the decorative movements of the hands and body and in the stamping of the feet; and nritya, the portrayal of mood through facial expression, hand gesture, and position of the legs and feet.

The style of movement in Indian classical dance is very different from that of Western ballet. In ballet the emphasis is frequently on the action of the legs—in jumps, turns, and fast traveling steps, which create ballet's characteristic qualities of height, speed, and lightness—while the body itself remains relatively still and the arms simply frame the face or balance the body. In Indian dance, however,



Sitakumari giving the Nataraja, God of Dance, pose in the style of bharata natya, one of the principal Indian classical dance forms.
Nick Sidle/The Dance Library

Differences from Western ballet

the legs are usually bent, with the feet flat rather than lifted and pointed. Jumps are usually low (though light), and the dancer rarely covers much ground or performs intricate steps, the complexity of the footwork lying more in elaborate stamping rhythms. (These stamping rhythms enhance the musicality of the dance; many dancers wear bells around their ankles, supplying their own accompaniment as well as counterpoint to the rhythms beaten out by the musicians.) The torso, face, arms, and hands are extremely active. The head is quite mobile, with subtle changes of direction and a characteristic side-to-side movement emphasizing the dancer's changing facial expressions. The movement of the torso is graceful and fluid, shifting from side to side or turning on the axis of the spine, while the movement of the hands and arms is subtle and elaborate, every gesture having a narrative function. Indian dancers have a vast repertoire of gestures through which they express complex events, ideas, and emotions. There are, for example, 13 gestures of the head, 36 different glances, and 67 mudras, or hand gestures, that can, in various combinations, yield several thousand different meanings.

While these qualities characterize Indian classical dance in general, there are significant variations in each school. Bharata natya is perhaps the most delicate and elegant of

Copyright © 1961 Mirisch Pictures Inc. All Rights Reserved



Members of the street gang "Sharks" dance the choreography of Jerome Robbins to the music of Leonard Bernstein in *West Side Story,* film version, 1961.

all the forms. It is traditionally, though not exclusively, performed by women. In the dancer's typical stance the legs are bent, often turned out at the hips, and the body is held upright. Even in movements where the torso bends or spirals, it remains lifted, never dropped and heavy. The feet perform small stamping movements against the ground; for example, the heel may be lifted and the leg extended to the front or back and then brought back. This is a small movement at some points and at others a larger lunge, but none of the steps travels far off the spot. Stamping movements are also made by raising the foot, bringing it down on the ball, and then bringing down the heel. These quick, shifting steps maintain a complex rhythmic relationship with the musical accompaniment. Sometimes such steps include a light spring from one foot to the other.

While the feet are executing the basic step sequence, the arms, hands, and head are also performing intricate movements. The arms are always supported at the elbow, never loosely hanging, and they may be stretched to the side or above the head or bent at the elbows in many different positions. In executing the mudras, the hands convey different meanings according to the position of the fingers and the way the palms are cupped or splayed. The neck moves from side to side, the head nods or turns, the eyes dart and glance in different directions, and the body tilts or leans. Each of these different movements contributes to the rhythmic and visual complexity of the dance.

Kathakali is a dance-drama performed by men, as it is considered too vigorous and difficult for women (although women may study it and perform certain extracts). The dancers wear elaborate headdresses and costumes as well as extensive makeup. The makeup can take up to four hours to apply and allows the dancer to absorb himself in the role he is about to perform. The basic kathakali stance is a deep bend, with the legs turned in and the feet resting on the outside of the soles, giving the dancer a bandy-legged look. This position allows him to survive the long performances without getting sore feet.

Much kathakali dancing is vigorous. The stamping steps are larger and more energetic than in bharata natya, the legs are lifted higher, the lunges are deeper, and the jumps are bigger. Generally the dancers travel farther and with greater agility. The arm, body, and head movements are also more dramatically expressive: the body crouches and twists furiously, the arms make larger, more imitative gestures (as in fight scenes), and the facial expressions are highly exaggerated. (Kathakali dancers have such control over their facial muscles that they can laugh with one side of their faces and cry with the other.) Mudras also help dramatize the story, although they do not always signify the same things as in bharata natya.

Orissi and manipuri use more sinuous movement, in which the spine and torso are elegantly curved. The most characteristic element in kathak is the *chukra,* a brilliant, whipping turn executed on the spot. In this style the feet work closely together, often with one crossed in front of the other and stamping out unusually complex rhythms. The dancer also uses a gliding walk similar to the pas de bourrée in ballet. There is more swaying of the body in kathak than in bharata natya, and the wearing of a full skirt emphasizes the speed and excitement of the dancer's turns.

The influence of Indian dance can be seen throughout Asia. In Japan, for example, the dancer makes use of a fan to create an additional repertoire of gestures. It may be opened to suggest the reading of a book, whirled and dropped to the ground to show the falling of leaves, or, appearing above the dancer's sleeve, used to signify that the Moon has risen. In Java the dancers' faces remain impassive, but their hand gestures are elaborate, and they also manipulate long, floating scarves to give their movements a weightless, ethereal quality.

## TRIBAL AND ETHNIC DANCE

Ballet, modern dance, and Indian classical dance are forms of theatre dance, the dancers usually being highly trained professionals performing for audiences in particular venues and on special occasions. Tribal and ethnic dance, on the other hand, may be characterized by a number of almost opposite features. They are not necessarily the province of trained specialists (although they may be); such dances may be participatory (*i.e.,* with no real distinction between dancer and spectator); and while they may take place in special venues or on special occasions, these are often intimately related to the everyday life of the community.

**Tribal dance.** A tribal society is essentially a self-contained system. While it may possess sophisticated cultural and social structures, its technological and economic structures are generally primitive. Consequently, by the late 20th century such societies had become increasingly rare, and many tribal dances had either died or become transformed.

Some tribal dances have been preserved, however, even in cases where tribes have been absorbed into other social structures, as a means of preserving cultural identity and a sense of historical continuity. This is quite common in many African states. A frequently cited case is that of King Sobhuza II, the Ngwenyama ("Lion") of Swaziland, who in 1966 joined his people in a six-day *Incwala,* or ritual ceremony. Dressed in animal skins and elaborate plumage, Sobhuza performed dances that would ensure the renewal of the land, the king, and the people.

U. Bagel—ZEFA



South African performing a tribal dance in a traditional animal skin costume with elaborate plumage during a ceremonial gathering of regional bands.

In extant tribal societies, such as the Hopi Indians of northeastern Arizona, dance retains most of its traditional form and significance. The Hopi still dance as a form of worship, with specific dances for different ceremonies. Such dances, however, as in any other tradition, have undergone inevitable change and development throughout history, and they cannot be used as accurate evidence of what the tribal dances of early man were like. Generalizing about tribal dance is made difficult not only by the lack of evidence concerning its origins and the rapid dying of extant forms but also by the fact that the term tribal covers so many different kinds of dance. Tribal dances not only vary from one tribe to another but also fall into many different categories, such as weapon dances, fertility dances, Sun and Moon worshiping dances, initiation dances, war dances, and hunting dances.

Difficulty of defining tribal dance

The following are two examples of tribal dance that have survived into the 20th century. The musicologist Curt Sachs quoted a description of the fertility dance of the Cobéua Indians of Brazil:

> The dancers have large [artificial] phalli . . . which they hold close to their bodies with both hands. Stamping with the right foot and singing, they dance . . . with the upper parts of their bodies bent forwards. Suddenly they jump wildly along with violent coitus motions and loud groans. . . . Thus they carry

> the fertility into every corner of the houses . . . ; they jump among the women, young and old, who disperse shrieking and laughing; they knock the phalli one against another.

Joan Lawson described the tree-worship dance performed both in Australia and up the Amazon River:

> A solemn circling of the tree is followed by an ecstatic raising of the head and hands to the branches, leaves, and fruit. Hands are then gradually run down the trunk and finally the men kneel or lie grovelling at the roots. They hope that by so doing the strength of the tree will enter into them.

An interesting parallel with tribal dances may be found in the break-dancing and "body-popping" craze that swept the United States and Britain in the 1980s. While the dancers clearly were not members of a tribe in any strict sense, they were often members of a distinct group or crew that had its own style and identity. These crews were part of a larger group of young people, again with its own style and customs, that could be differentiated from other groups such as punks or skinheads. The two dance forms were characterized by an energetic spinning action, whereby the dancer propelled himself around on his neck, head, or shoulders and by small, jerky movements of the joints that traveled in a wave through his body. Rival crews often competed with one another in the street, showing off the skill and ingenuity of their moves.

**Ethnic dance.** In describing many dances, reference is often made to their ethnic, rather than their tribal, origins. An ethnic dance is simply a dance that is characteristic of a particular cultural group. Under this definition even the polka, which is almost always considered a social dance, may be called ethnic, as it began in a culturally distinct region of Europe. Flamenco, which began as an improvised dance among Andalusian gypsies, combines toe and heel clicking with body movements similar to Indian dance. Indian dances may be regarded as a general ethnic type, but there are numerous forms and traditions within the type: some are classical (see above *Indian classical dance*), while others are popular, being danced by nonspecialists for communal festivities and for recreation.

Use of the term ethnic

In this discussion of the art of dance, it is most useful to reserve the designation ethnic for those genres that, while perhaps in a state of transition, are still practiced by a unique cultural group, still retain some of their original communal or ritual functions, and have not yet reached the professionalized state of classical or folk dance.

The many Afro-Caribbean dance forms are usually considered to constitute a distinct ethnic form because they share certain characteristic movements. As in Indian dance, the legs are frequently bent, with the feet stamping out rhythms against the ground. The torso and back are also very mobile, executing sinuous rippling actions or more jerky, rhythmic movements. The body is frequently bent slightly forward, and there is greater use of the hips, which sway and circle in syncopated rhythms. Gestures and facial expressions are used in some narrative dances, but they tend to be much less sophisticated or strictly codified than in Indian dance.

In performance today, most Afro-Caribbean dance companies are made up of both dancers and drummers, the percussion marking out the rhythm and helping to intensify the emotion. Frequently the dancers take turns performing, and there is usually a great deal of informal communication among members of the company on stage. Participation by the audience is often encouraged at the end of the performance, reflecting the communal, rather than theatrical, origins of the form.

## FOLK DANCE

When tribal societies in Europe gave way to more structured societies, the old dance forms gradually developed into what are now called folk or peasant dances. For a long time these retained much of their original significance and therefore could have received the modern classification of "ethnic." The Maypole dance, still sometimes performed in England, is a descendant of older tree-worshiping dances, the ribbons that the dancers hold as they dance around the pole symbolizing the tree's branches. The morris dance, also called the moresque because the blackened faces of the dancers resembled the Moors, is a survival of early weapon dances, which were not war dances but an ancient form of religious worship. The types and styles of these different dances were numerous, and, as with tribal dances, many were lost so that information about them often remains sketchy. In the 20th century, efforts to collect national music and dances were made by, among others, Cecil Sharp in England and Béla Bartók in Hungary. These efforts resulted in the revival of certain dances, but they are now danced mainly for recreation, and their original significance has been lost. It is in this conscious revival or preservation of ethnic and national dances for purposes of entertainment that modern folk dance has its origin.

Although different areas and countries have different styles of dance, most of them share common formations and styles of movement. The earliest and simplest formation, the closed circle, is found in all folk dances and derives from the ritual of circling around an object of worship. The dancers grasp one another by the hands, wrists, shoulders, elbows, or waists and face the centre of the circle. In more complex forms dancers move into and out of the circle to perform individual movements or to join into couples; or as the dancers circle, they may weave around one another. In some dances there are two concentric circles, sometimes the inner one of men and the outer one of women.

Another common formation, the chain, involves a long line of dancers, often holding hands or linked by handkerchiefs. The leader may trace a complex, serpentine pattern for the others to follow. Processional dances may travel a long way—even through an entire village. The dancers are mostly in couples, with the procession halting at times for them to dance together.

Many folk dances today are performed in sets, groups of about eight dancers who may perform in all of the above formations but within a restricted space. In other dances, individuals may leave the group and dance on their own.

Folk dance steps are usually quite simple variations on walking, hopping, skipping, and turning. (See above *Basic steps and formations: Folk dance.*) Depending on the particular dance form, these steps may be long, slow, and gliding or short, fast, and springing. The hips are usually held still, though in more vigorous dances the men in particular may crouch, kneel, or even lie on the floor. Some dances involve large jumps and lifts, usually with the man seizing the woman by the waist, lifting her into the air, and possibly turning with her.

There are numerous kinds of holds. For example, two dancers may face each other and hold hands with the arms crossed, link arms, or use a hold similar to that of ballroom dancers. Individual folk dances may also contain distinctive motifs: the dancers may clap their hands, wave handkerchiefs, or clash sticks with one another. Some

M. Howard—ZEFA



Portuguese folk dancers from Algarve performing one of the traditional regional folk dances.

dances contain elements of mime—not only the bows and curtsies of courtship dance but also gestures such as those performed in certain Slavic harvest dances, where the arms are brought up to the chest and opened outward as if presenting something.

Many European folk dances are characterized by a strong emphasis on pattern and formation. The dancers frequently move in an ordered relation to one another, and the steps follow clearly delineated floor patterns on the ground. The circle is the simplest pattern, but the chain, the procession, and the longways dance are also common. (Some of the more complicated patterns are probably due to the influence of the court dances, which systematized and polished the more robust peasant forms.) Although there are numerous exceptions to the rule, the emphasis in many of those dances is on the footwork, rather than on large or vigorous movements of the body.

### SOCIAL DANCE

When the early European folk dances—particularly the courtship forms—were incorporated into court dances, they lost many of their boisterous and pantomimic elements. The man no longer thrust forward to embrace the woman or lifted her vigorously into the air, but simply knelt and took her hand. The woman's earlier violent resistance dwindled into a coquettish turn of the head, and energetic strides and runs gave way to simple gliding steps, often forming intricate patterns that were punctuated with small poses, bows, and curtsies.

The social, as opposed to the theatrical, forms that these early court dances inspired gradually became more elaborate and more lively, with small lifts, jumps, and turns being included, as in the galliard and lavolta. Gradually, too, the emphasis began to switch from the tight group formations of many earlier dances to the individual couple. By the end of the 18th century, in dances such as the waltz and, subsequently, the polka, people simply danced in pairs, with group formations reserved for public display. At the same time these dances came to be danced by all classes of people. Steps were simplified, and dancers no longer needed special instruction to perform them.

In the 20th century, ballroom dances became very popular, with new dances, such as the tango and fox-trot, and new variations gradually added to the repertoire. Like the waltz and polka, ballroom dances placed importance on nimble leg- and footwork, with almost no hip movement and the torso only slightly swaying to the rhythm of the dance. The advent of jazz, however, led to other forms of social dance as Western music fell under the influence of the descendants of African slaves in America. During the jazz era of the 1920s, dances like the Charleston and the Black Bottom not only showed the syncopated rhythms, bent knees, crouched torsos, and hip and pelvic movements of African dance but also broke through the dominance of the couple form. People might still dance opposite each other in pairs, but they no longer held each other or danced in unison, and it was perfectly permissible for the dancer to dance singly. As a consequence, dancers no longer followed a set pattern of steps but invented their own within the general style.

Afro-American influence

A dancer without a partner was free to choose the distance and direction in which to travel. Much more vigorous movements of the torso, legs, and arms were possible, as the dancer did not have to worry about getting in his partner's way. The dancer might jump, kick his legs, stretch his arms out to the side or above the head or swing them through the air and might crouch, extend his body, or twist with complete freedom. The lindy and rock and roll brought back contact between the dancers, but it was of a very acrobatic and individualistic kind. The influence of African dance could still be seen in disco and other popular forms, particularly in the characteristic swaying of the hips and the jerky, percussive movements of the torso marking the rhythms of the music.

Henri Cartier-Bresson—Magnum



Ballroom dancing.

**BIBLIOGRAPHY**

*General works:* ANATOLE CHUJOY and P.W. MANCHESTER (comps. and eds.), *The Dance Encyclopedia,* rev. and enl. ed. (1967), a standard reference source with articles about all forms of dance, containing almost 300 photographs; G.B.L. WILSON, *A Dictionary of Ballet,* 3rd ed. (1974), a comprehensive reference source; and CURT SACHS, *World History of the Dance* (1937, reprinted 1965; originally published in German, 1933), a classic study of the dance in all forms, with special focus on origins, although some of Sachs's arguments have been challenged by more recent anthropological studies. LOUIS HORST, *Pre-Classic Dance Forms* (1937, reprinted 1968), a study of early dances; RICHARD KRAUS and SARAH ALBERTI CHAPMAN, *History of the Dance in Art and Education,* 2nd ed. (1981); and WALTER SORELL, *Dance in Its Time* (1981), analyze the subject within a wide cultural and social context. ROGER COPELAND and MARSHALL COHEN (eds.), *What Is Dance?: Readings in Theory and Criticism* (1983), is a collection of essays on the nature of dance and its different styles and forms. Physiological aspects of dance and the mechanics of human movements are discussed in KENNETH LAWS, *The Physics of Dance* (1984).

*Choreography:* Contemporary works on choreography include FREDERICK RAND ROGERS (ed.), *Dance, a Basic Educational Technique: A Functional Approach to the Use of Rhythmics & Dance as Prime Methods of Body Development & Control, and Transformation of Moral & Social Behavior* (1941, reprinted 1980); PEGGY VAN PRAAGH and PETER BRINSON, *The Choreographic Art: An Outline of Its Principles and Craft* (1963); LA MERI (RUSSELL MERIWETHER HUGHES), *Dance Composition: The Basic Elements* (1965); and DORIS HUMPHREY, *The Art of Making Dances* (1959, reprinted 1981). Reflections on the creative process involved in some of the choreographer's major dance works can be found in MARY WIGMAN, *The Language of Dance* (1966; originally published in German, 1963).

*Dance notation:* One of the first inventors of dance notation, RAOUL-AUGER FEUILLET, showed the floor pattern of dances in his work *Orchesography* (1706, reprinted 1971; originally published in French, 1700). Other early works on the subject include ARTHUR SAINT-LÉON, *La Sténochorégraphie: Ou, art d'écrire promptement la danse* (1852); and V.I. STEPANOV, *Alphabet of Movements of the Human Body: A Study in Recording the Movements of the Human Body by Means of Musical Signs* (1958, reissued 1969; originally published in French, 1892). Modern works include RUDOLF BENESH and JOAN BENESH, *An Introduction to Benesh Movement-Notation: Dance,* rev. and extended ed. (1969); NOA ESHKOL and ABRAHAM WACHMANN, *Movement Notation* (1958), supplemented with *Movement Notation Survey 1973: Eshkol-Wachmann Movement Notation* (1973); and NOA ESHKOL, MICHAL SHOSHANI, and MOOKY DAGAN, *Movement Notations: A Comparative Study of Labanotation (Kinetography Laban) and Eshkol-Wachmann Movement Notation* (1979). Current studies of dance notation are found in the periodicals *Ballet News* (monthly); and *Dance Notation Journal* (semiannual).

*Theatrical aspects of dancing:* MARY CLARKE and CLEMENT CRISP, *Making a Ballet* (1974), offers observations on the different relationships between choreographers and dancers, designers, and composers, and their *Design for Ballet* (1978) is a lavishly illustrated survey of costume and set design. MERLE ARMITAGE (ed.), *Martha Graham, the Early Years* (1937, reprinted 1978); and LE ROY LEATHERMAN, *Martha Graham: Portrait of the Lady as an Artist* (1966), explore the nature of this prominent choreographer's work. See also DAVID VAUGHAN, *Frederick Ashton and His Ballets* (1977); and MERCE CUNNINGHAM, *Changes: Notes on Choreography,* edited by FRANCES STARR (1969). For the analysis of technical components of dance as theatre, see ELIZABETH R. HAYES, *Dance Composition and Production for High Schools and Colleges* (1955, reissued 1981); CYRIL W. BEAUMONT and STANISLAS IDZIKOWSKI, *A Manual of the Theory and Practice of Classical Theatrical Dancing (Classical Ballet) (Cecchetti Method),* rev. ed. (1977); and articles on the technical aspects of the art in *Dance Chronicle: Studies in Dance and the Related Arts* (quarterly).

*Ballet:* Development of the ballet as a theatre art is reflected in the early writings of some dancing masters such as CESARE NEGRI MILANESE, *Nuove inventioni di balli: opera vaghissima* (1604; reissued in 1969 as *Le gratie d'amore*), a richly illustrated treatise. See also CLAUDE FRANÇOIS MÉNESTRIER, *Des Ballets anciens et modernes selon les régles du théâtre* (1682, reprinted 1972), the first printed history of the ballet; DERYCK LYNHAM, *The Chevalier Noverre: Father of Modern Ballet* (1950, reprinted 1972), a biographical history; and THÉOPHILE GAUTIER, *The Romantic Ballet as Seen by Théophile Gautier,* trans. from French by CYRIL W. BEAUMONT (1932, reprinted 1980). LINCOLN KIRSTEIN, *Dance: A Short History of Classic Theatrical Dancing* (1935, reprinted 1970), and *Movement & Metaphor* (1970, reissued as *Four Centuries of Ballet,* 1984), are brilliant analyses of the component parts of ballet and its developments, based on a wide survey of works. An authoritative historical study is JOAN LAWSON, *A History of Ballet and Its Makers* (1964, reprinted 1976). SELMA JEANNE COHEN, *Next Week, Swan Lake: Reflections on Dance and Dances* (1982), is a witty and illuminating discussion of some basic issues in dance criticism. For the analysis of ballet techniques, see JEAN-GEORGES NOVERRE, *Letters on Dancing and Ballets* (1930, reissued 1966; originally published in French, rev. ed. 1803–04), a reformer's statement of the principles of ballet techniques, which are still valid; and CARLO BLASIS, *An Elementary Treatise upon the Theory and Practice of the Art of Dancing* (1944; originally published in French, 1820), a book by an Italian dancer and choreographer who codified the techniques of classic ballet. For the librettos of most famous ballets, see CYRIL W. BEAUMONT, *Complete Book of Ballets: A Guide to the Principal Ballets of the Nineteenth and Twentieth Centuries,* rev. ed. (1949, reprinted 1956), supplemented with his *Ballets of Today* (1954), and *Ballets, Past & Present* (1955); and WALTER TERRY, *Ballet Guide: Background, Listings, Credits, and Descriptions of More Than Five Hundred of the World's Major Ballets* (1982). Contributions of the Russian ballet are discussed by RICHARD BUCKLE, *Diaghilev* (1979, reprinted 1984), and *Nijinsky,* 2nd ed. (1975); and by NATALIA ROSLAVLEVA, *Era of the Russian Ballet* (1966, reprinted 1979).

*Modern dance:* Various aspects of modern dance and its forms are studied in WALTER TERRY, *The Dance in America,* rev. ed. (1971, reprinted 1981); WALTER SORELL, *The Dance Has Many Faces,* 2nd ed. (1966); and JOHN MARTIN, *Introduction to the Dance* (1939, reissued 1965), especially good on the theory of the early modern dance. Also see SELMA JEANNE COHEN (ed.), *The Modern Dance: Seven Statements of Belief* (1966, reprinted 1973), a collection of essays by important choreographers; SALLY BANES, *Terpsichore in Sneakers: Post-Modern Dance* (1980), a survey of the subject; and JOSEPH H. MAZO, *Prime Movers: The Makers of Modern Dance in America* (1977), which contains useful analyses of many choreographers' works. Current research in choreography is presented in the periodicals *Dancing Times* (monthly); *Dance Research Journal* (semiannual); and *Dance Magazine* (monthly).

*Indian classical dance:* Analysis of regional dancing is found in KAY AMBROSE, *Classical Dances and Costumes of India,* 2nd ed. (1983), a well-documented description of basic forms and styles; BERYL DE ZOETE, *The Other Mind: A Study of Dance in South India* (1953, reissued 1960); and RINA SINGHA and REGINALD MASSEY, *Indian Dances: Their History and Growth* (1967), a well-illustrated guide.

*Folk and social dance:* Folk dance is the subject of CECIL J. SHARP and A.P. OPPÉ, *The Dance: An Historical Survey of Dancing in Europe* (1924, reprinted 1972); and JOAN LAWSON, *European Folk Dance* (1953, reprinted 1980). Periodicals include *Arabesque: A Magazine of International Dance* (bimonthly); *American Square Dance* (monthly); and *Square Dancing* (monthly). Social and ballroom dances are analyzed in ARTHUR H. FRANKS, *Social Dance: A Short History* (1963); and PHILIP J.S. RICHARDSON, *The Social Dances of the Nineteenth Century in England* (1960).

(J.R.Ma.)

# The History of Western Dance

The peoples of the West—of Europe and of the countries founded through permanent European settlement elsewhere—have a history of dance characterized by great diversity and rapid change. Whereas most dancers of the East repeated highly refined forms of movement that had remained virtually unchanged for centuries or millennia, Western dancers showed a constant readiness, even eagerness, to accept new vehicles for their dancing. From the earliest records, it appears that Western dance has always embraced an enormous variety of communal or ritual dances, of social dances enjoyed by many different levels of society, and of skilled theatrical dances that followed distinct but often overlapping lines of development.

The articles FOLK ARTS and POPULAR ARTS cover in greater detail the unique nature, techniques, forms, and functions, and the historical developments of each of these kinds of Western dance. In addition, the article DANCE, THE ART OF, covers the aesthetics and the varieties of dance, both Western and non-Western. Aspects of Eastern dance are detailed in the article EAST ASIAN ARTS.

The West cannot always be clearly distinguished from the non-West, especially in such countries as the Soviet Union where some dances are Asian and others European in origin and character. This article focuses on the dance of Western peoples, noting where appropriate the influence of other cultures.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 625.

The article is divided into the following sections:

From antiquity through the Renaissance 958
Dance in the ancient world 958
Ancient Egyptian dance
Dance in classical Greece
Ancient Roman dance
Jewish dance
Christianity and the Middle Ages 960
Ecclesiastical attitudes and practices
Dance ecstasies
Dance and social class
The Renaissance world and the art dance 961
Court dances and spectacles
The birth of ballet
During the 17th, 18th, and 19th centuries 962
The maturing of ballet 963
Technical codifications and dance scholarship
Varieties of the ballet
Early virtuosos of the dance
The reign of the minuet 964
English social dance 964
Dance in colonial America 965
The rise of the waltz 965
The Romantic movement in dance
Spread of the waltz
Offspring and rivals
Foundations of modern ballet 966
Noverre and his contemporaries
The Romantic ballet
Theatre and ballroom dance 967
The 20th century 967
Diaghilev and his achievements 967
The Ballets Russes
The continuing tradition
The Soviet ballet
Modern dance 968
New rhythms, new forms 969
Latin-American and jazz dances
Dance contests and codes
Effect on folk dancing
Experiments
The dance since 1945 969
Social dance
Dance in the theatre
Dance and the film
Bibliography 970

## From antiquity through the Renaissance

Before written records were left, a vast span of time elapsed about which scholars can only speculate. Pictorial records in cave paintings in Spain and France showing dancelike formations have led to the conjecture that religious rites and attempts to influence events through sympathetic magic were central motivations of prehistoric dance. Such speculations have been reinforced by observation of dances of primitive peoples in the contemporary world, though the connection between ancient and modern "primitives" is by no means accepted by many scholars. If the dances recorded in early written records represented a continuity from prehistoric dances, there may have been prehistoric work dances, war dances, and erotic couple and group dances as well. One couple dance surviving in the 20th century, the Bavarian-Austrian *Schuhplattler,* is considered by historians to be of Neolithic origin, from before 3000 BC.

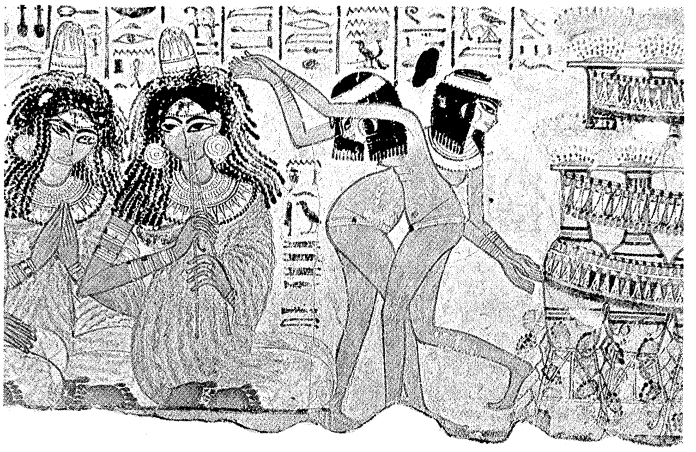
### DANCE IN THE ANCIENT WORLD

In the civilizations of Egypt, Greece and its neighbouring islands, and Rome, written records supplement the many pictorial remains. Written records alone provide information about ancient Jewish dancing. There are still conjectures about the style, pattern, and purpose of ancient dances, but there is far more concrete evidence.

**Ancient Egyptian dance.** Formalized ritual and ceremonial dances in which the dancing priest–king represented the person of a god or the servant and regenerator of his people were practiced in Egypt. These dances, culminating in ceremonies representing the death and rebirth of the god Osiris, became more and more complex, and ultimately they could be executed only by specially trained dancers. From Egypt also come the earliest written documentations of the dance. These records speak of a class of professional dancers, originally imported from the interior of Africa, to satisfy the wealthy and powerful during hours of leisure and to perform at religious and funerary celebrations. These dancers were considered highly valuable possessions, especially the pygmy dancers who became famous for their artistry. One of the pharaohs prayed to become a "dance dwarf of god" after his death, and King Neferkare (3rd millennium BC) admonished one of his marshals to rush such a "dance dwarf from the Land of Spirits" to his court.

Professional dancers

There is considerable agreement that the belly dance, now performed by dancers from the Middle East, is of African origin. A report of the 4th century BC from Memphis in Egypt described in detail the performance of an apparently rumba-like couple dance with an unquestionably erotic character. The Egyptians also knew acrobatic exhibition dances akin to the present-day adagio dances. They definitely were aware of the sensual allure of the sparsely clad body in graceful movement. A tomb painting from Shaykh 'Abd al-Qurnah, now in the British Museum, shows dancers dressed with only rings and belts, apparently designed to heighten the appeal of their nudity. These figures probably were intended to entertain the dead as they had been entertained in life.

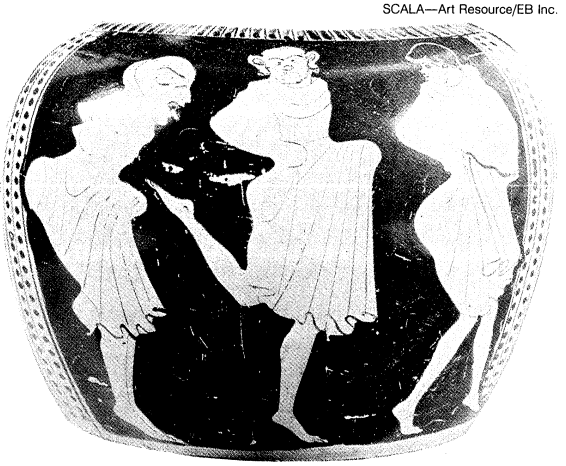


Egyptian dancing, detail from a tomb painting from Shaykh 'Abd al-Qurnah, Egypt, *c.* 1400 BC. In the British Museum.
By courtesy of the trustees of the British Museum

Egypt, then, presented a dancing scene that was already varied and sophisticated. In addition to their own danced temple rituals and the pygmy dancers imported from the headwaters of the Nile, there were Hindu dancing girls from conquered countries to the east. This new dance had none of the long masculine strides or the stiff, angular postures seen in so many Egyptian stone reliefs. Lines of movement undulated softly, nowhere bending sharply or breaking. These Asiatic girls brought a true feminine style to Egyptian dance.

**Dance in classical Greece.** Many Egyptian influences can be found in the Greek dance. Some came by way of Crete, others through the Greek philosophers who went to Egypt to study. The philosopher Plato (*c.* 428–348/47 BC) was among them, and he became an influential dance theoretician. He distinguished dances that enhance the beauty of the body from awkward movements that imitate the convulsions of ugliness. The Apis cult dances of Egypt had their equivalent in the Cretan bull dance of about 1400 BC. It inspired the labyrinthine dances that, according to legends, Theseus brought to Athens on his return with the liberated youths and maidens.

SCALA—Art Resource/EB Inc.



Kordax dance, Greek vase painting, 5th century BC. In the Museo Nazionale Tarquinise, Italy.

Another dance form that originated in Crete and flourished in Greece was the *pyrrhichē,* a weapon dance. Practiced in Sparta as part of military training, it was a basis for the claim of the philosopher Socrates that the best dancer is also the best warrior. Other choral dances that came to Athens from Crete include two dedicated to Apollo and one in which naked boys simulated wrestling matches. Female characteristics were stressed in a stately and devout round dance in honour of the gods, performed by choruses of virgins.

Numerous vase paintings and sculptural reliefs offer proof of an ecstatic dance connected with the cult of Dionysus. It was celebrated with a "sacred madness" at the time of the autumnal grape harvest. In his drama *Bacchae,* Euripides (*c.* 480–406 BC) described the frenzy of Greek women, called bacchantes or maenads. In their dance for generation and regeneration, they frantically stamped the ground and whirled about in rhythmic convulsions. Such dances were manifestations of demoniacal possession characteristic of many primitive dances.

Dionysian dances

The Dionysian cult brought about Greek drama. After the women danced, the men followed in the disguise of lecherous satyrs. Gradually the priest, singing of the life, death, and return of Dionysus while his acolytes represented his words in dance and mime, became an actor. The scope of the dance slowly widened to incorporate subjects and heroes taken from the Homeric legends. A second actor and a chorus were added. In the lyric interludes between plays, dancers re-created the dramatic themes in movements adopted from the earlier ritual and bacchic dances. In the comedies, they danced the very popular kordax, a mask dance of uninhibited lasciviousness. In the tragedies, the chorus performed the *emmeleia,* a dignified dance with flute accompaniment.

These dances and plays were executed by skilled amateurs. At the end of the 5th century BC, however, there came into being a special class of show dancers, acrobats, and jugglers, the female members of which were evidently hetaerae, members of a class of courtesans. No doubt influenced by Egyptian examples, they entertained guests at lavish banquets. The historian Xenophon (*c.* 430–*c.* 355 BC) in his *Symposium* tells of the praise Socrates lavished on a female dancer and a dancing boy at one such occasion, finally himself emulating their beautiful movements. Elsewhere, Xenophon describes a dance representing the union of the legendary heroine Ariadne with Dionysus, an early example of narrative dance.

**Ancient Roman dance.** There was a striking difference between the Etruscan and the Roman peoples in their approach to the dance. Little is known about the Etruscans, who populated the area north of Rome up to Florence and flourished between the 7th and 5th century BC. But it is apparent from their lavish tomb painting that dance played an important part in their enjoyment of life. Women were enthusiastic participants in Etruscan dancing; funerary chain dances were performed by groups of women, and lively, energetic couple dances are portrayed in Etruscan frescoes. They were performed without masks in public places and showed a distinct courting character.

Roman antagonism to dance seems to reflect a sober rationalism and realism. Nonetheless, Rome did not entirely evade the temptations of dance. Before about 200 BC, dances were evidently in the form of choral processions only. There were agricultural processions headed by priests, and weapon dances of the Salii, a congregation of the priests of Mars who walked around in a circle while rhythmically beating their shields. Dancing was an important part of Roman festivals—the celebrations of Lupercalia and Saturnalia featured wild group dances that were precursors of the later European carnival.

Antagonism of the Romans

Later, Greek and Etruscan influences began to spread, though people who danced were considered suspicious, effeminate, and even dangerous by the Roman nobility. One public official did not believe his eyes when he watched dozens of the daughters and sons of well-respected Roman patricians and citizens enjoying themselves in a dancing school. About 150 BC all dancing schools were ordered closed, but the trend could not be stopped. And though dance may have been alien to the Roman's inner nature, dancers and dancing teachers were increasingly brought from abroad in the following years. The statesman and scholar Cicero (106–43 BC) summed up the general opinion of the Romans when he stated that no man danced unless he was insane.

A form of dance that enjoyed great popularity with the Romans under the emperor Augustus (63 BC–AD 14) was the wordless, spectacular pantomime that rendered dramatic stories by means of stylized gestures. The performers, known as *pantomimi,* were at first considered more or less as interpreters of a foreign language, since they



Funeral dance, Etruscan fresco from a tomb cover, 5th century BC. In the Museo e Gallerie Nazionale di Capodimente.
SCALA—Art Resource/EB Inc.

came from Greece. They refined their art until the two dancer-mimes Bathyllus and Pylades became the star performers of Augustan Rome. The stylized performance of the dancer, who wore a mask appropriate to the theme of his dance, was accompanied by musicians playing flutes, horns, and percussion instruments and a chorus that sang about the action between dance episodes.

**Jewish dance.** When dance is mentioned in the Old Testament it is distinguished by its joyousness. Words such as leaping and whirling describe the energy and vitality of ancient Hebrew dances. As in other early societies, dancing is most often connected with ritualistic activity. Ring dances may have been performed in the worship of the golden calf; the prohibition against making graven images that resulted from this worship explains the lack of evidence of Jewish dances in the visual arts.

Hebrew dances were performed by both men and women, though usually the sexes were separated. Victory dances were performed by groups of women; men participated in ecstatic whirling dances designed to evoke prophecy. Festival dances were performed by both groups—one of the most important was the water-drawing festival on the first night of Sukkoth, which was celebrated by a torchlit procession dance that lasted through the night.

Weddings provided another important occasion for ritual dancing. Dancing with the bride was considered an act of devotion, and the officiating rabbi always complied with pleasure. During the Diaspora of the early Christian Era many of the ritual dances disappeared, but the bridal dance continued as a tradition. In the Middle Ages a wedding dance called the *mitzvah* was performed; because of the segregation of the sexes men danced with the bridegroom and women with the bride. Later, men could dance with the bride either by wrapping their hands in a cloth or by holding a cloth between them to signify their separation.

## CHRISTIANITY AND THE MIDDLE AGES

Dancing was traditional also among the tribes of barbarians to the north, as attested by the writings of the Christian missionaries. Wherever they went, they found the same fertility-rite dances—if in different guise, the same charm dances to induce good and ward off evil, the same warrior and weapon dances to bolster fighting morale, and the same uncontrolled expressions of the joy of life, which the missionaries attributed to the devil.

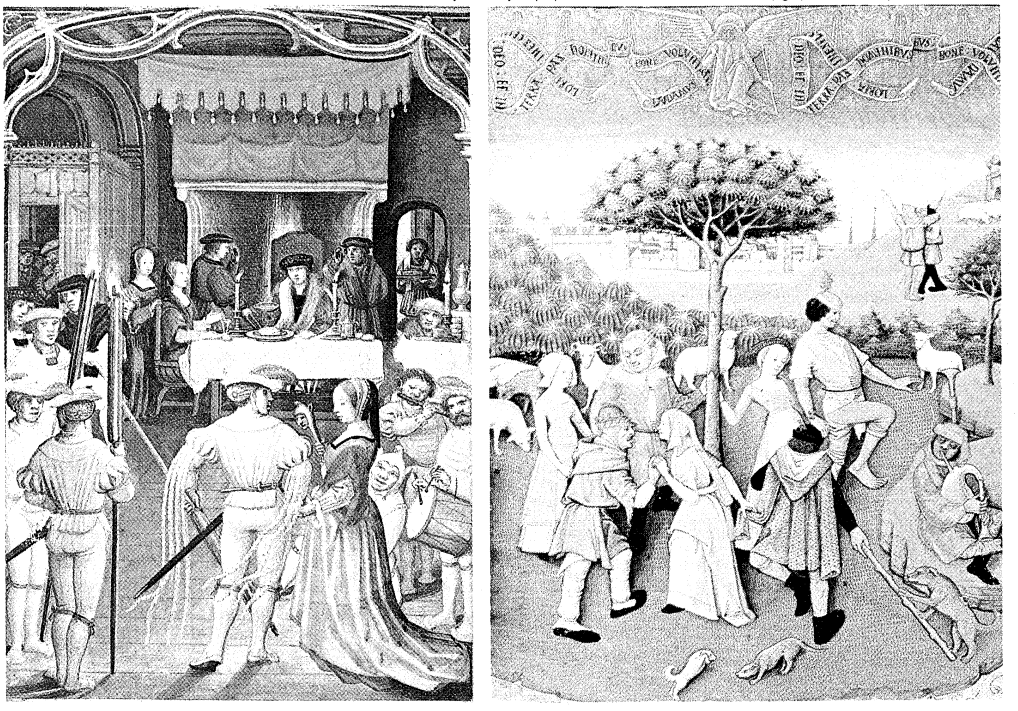
Erotic dancing was not the exclusive property of heathen societies. In Byzantium, the Christian emperor Justinian I (483–565) married the notorious Theodora, a dancer who had appeared in the nude in theatrical performances. About 500, St. Caesarius of Arles reported a sacrificial banquet ending in some demoniacal dancing rites performed to the accompaniment of lewd songs. The Anglo-Saxons had little girls performing dances at Easter in which a phallus was carried in front of them.

**Ecclesiastical attitudes and practices.** The attitude of the Christian Church toward dance was not unanimous. On the one side there was the ascetic rejection of all manifestations of lust and ecstasy, and dance was seen as one of the strongest persuasions to sexual permissiveness. On the other side, some early Church Fathers tried to find functions for pagan dances in Christian worship. St. Basil of Caesarea in 350 called dancing the most noble activity of the angels, a theory later endorsed by the Italian poet Dante. St. Augustine (354–430) was strictly against dancing, but, despite his great influence in the medieval church, dancing in churches continued for centuries.

Ambivalence of the church

Charlemagne, the Holy Roman emperor at the beginning of the 9th century, officially prohibited all kinds of

By courtesy of (left) the trustees of the British Museum, (right) the Bibliotheque Nationale, Paris



*Late medieval court couple dancing and rustic dancing.*
(Left) Torch dance from the *Golf Book,* attributed to Simon Bening, school of Bruges, *c.* 1500. In the British Museum (Add. MS. 24,098). (Right) Peasant round dance from the *Hours of Charles d'Angoulême,* French, late 15th century. In the Bibliothèque Nationale, Paris (MS. lat. 1173, fol. 20 v).

dancing, but the ban was not observed. The Teutonic peoples were accustomed to dancing as part of their religious rites. On Christian feast days, which coincided with their ancient rites of expelling the winter, of celebrating the arrival of spring, and of rejoicing that the days grew longer again, they revived their old ritual dances, though these were camouflaged with new names and executed to different purpose. In this manner previously sacred dances became more and more secularized. After such secularization, two lines of development were open: the social dance or the assimilation of dance into theatrical spectacle by the joculators, travelling comedians who combined the arts of dancer, juggler, acrobat, singer, actor, mime, and musician in one person.

**Dance ecstasies.** There were two kinds of dance peculiar to the Middle Ages, the dance of death, or danse macabre, and the dancing mania known as St. Vitus' dance. Both originally were ecstatic mass dances, dating from the 11th and 12th centuries. People congregated at churchyards to sing and dance while the representatives of the church tried in vain to stop them. In the 14th century another form of the dance of death emerged in Germany, the *Totentanz,* a danced drama with the character of Death seizing people one after the other without distinctions of class or privilege. The German painter Hans Holbein the Younger (1497/98–1543) made a famous series of engravings of this dance.

The dance of death

The St. Vitus' dance became a real public menace, seizing hundreds of people, spreading from city to city, mainly in the Low Countries, in Germany, and in Italy during the 14th and 15th centuries. It was a kind of mass hysteria, a wild leaping dance in which the people screamed and foamed with fury, with the appearance of persons possessed. In these convulsive, frantic, and jerky dances, religious, medical, and social influences probably interacted in response to such things as the epilepsy-like seizures of persons suffering from the Black Death. Italy was afflicted with tarantism, an epidemic presumably caused by the bite of venomous spiders. Its effects had to be counteracted by distributing the poison over the whole body and "sweating it out," which was accomplished by dancing to a special kind of music, the tarantella.

**Dance and social class.** In western Europe by the 12th century, society had developed into three classes, the nobility, the peasantry, and the clergy. This separation contributed to the development of the social dance. The knights created their own worldly and spiritual ideals, exemplified in tournaments and courtly entertainments that were praised in song and poetry by the troubadours and minnesingers. The couple dances of the knights expressed the polished and aristocratic notions of courtly love. The round dances of the peasants were executed by circles or lines of people, often singing and holding each other by their hands. The rustic choral round had strong pantomimic leanings and unpolished expressions of joy and passion. And while the choral rounds almost always were executed to the singing of the participants, the court dances of the knights generally were accompanied by instrumental playing, especially of fiddles, and when there was singing, it emerged from the spectators rather than the performers.

Courtly couples and rustic rounds

From the late Middle Ages, graphic artists frequently recorded what dancing looked like in all its different manifestations. How dancing adapted to the idealism of knightly love is shown in manuscript illuminations and tapestries. Paintings of the Flemish painter Pieter Bruegel the Elder (*c.* 1525/30–69) leave no doubt that the peasants enjoyed celebrating with dances of uninhibited stamping and cavorting.

## THE RENAISSANCE WORLD AND THE ART DANCE

France had set the fashion in court dance during the late Middle Ages; with the Renaissance, however, Italy became the centre of the new developments in dance. The Renaissance brought greater mixing of social classes, new fortunes and personal wealth, and greater indulgence in worldly pleasures and in the appreciation of the human body. The period emerged as one of the most dance-conscious ages in history.

**Court dances and spectacles.** Celebrations and festivities proliferated. The itinerant jugglers of the Middle Ages became highly respected and much sought after as dancing masters. They quickly assumed the function of instructing the nobility not only in the steps but also on posture, bearing, and etiquette. They became responsible for the planning and realization of the spectacular festivities. The social prestige of this newly developing profession grew constantly.

Some of these dancing masters were highly learned men, and their treatises leave no doubt about their scholarly ambitions. Many of them were Jewish, descended from the Klesmorim, a group of medieval Jewish entertainers. The first dancing master known by name was Domenico da Piacenza, who in 1416 published the first European dance manual, *De arte saltandi et choreas ducendi* ("On the Art of Dancing and Directing Choruses"). His disciple, Antonio Cornazano, a nobleman by birth, became an immensely respected minister, educator of princes, court

The rise of the dancing master

By courtesy of (right) the Viscount De L'Isle; photograph, (left) SCALA—Art Resource/EB Inc.



*Renaissance dances.*
(Left) Court dance in early balletic form as seen in "Catherine de Médicis Receiving the Polish Ambassador," tapestry designed by François Quesnel, *c.* 1575. In the Uffizi, Florence. (Right) "Queen Elizabeth Doing a Leaping Turn of the Volta, Assisted by the Earl of Leicester," oil painting by an unknown artist, 16th–17th century. In the collection of the Viscount De L'Isle, Penshurst Place, Kent.

poet, and dancing master to the Sforza family of Milan, where about 1460 he published his *Libro dell'arte del danzare* ("Book of the Art of the Dance"). Such books record little about the actual steps and the melodies to which they were performed, but they are eloquent in the description of the *balli*—works that were invented by the dancing masters themselves. Adapting steps from the various social dances, they used them in a kind of dance pantomime.

In France, numerous forms developed from the branle, a round dance of peasant origin that became fashionable in the courts. One of the most frequently mentioned of all the dances of the 15th century was the *morisca,* or *moresque,* a romanticized version of dances from Moorish Spain. These were first mentioned in 1446 by a Bohemian traveller who visited Burgos, Spain. Later, in Portugal, he encountered similar forms. Sometimes religious motifs of the legendary fight between Charlemagne and the Turkic invader Timur entered the *morisca,* but usually it was performed as a double-file choral dance. It had nothing to do, as was long believed, with the English masked Morris dance, now considered to be a survival from a primitive religious cult.

From such choral dances the ballet emerged. At the court entertainments throughout Savoy and northern Italy, sumptuous spectacles with mythological, symbolical, or allegorical content became increasingly popular. At these early stages, however, pantomime and dance are not easily distinguished. Famous examples of these spectacles are the presentation of the story of Jason and the Golden Fleece at the marriage of Philip the Good of Burgundy in 1430, and the dinner ballet on the same, though widely enlarged, subject staged for the wedding of the Duke of Milan in 1489.

Tudor England of the early 16th century had similar pageants, with the participants disguising "after the manner of Italie." Like the Italian *balli,* the English masque offered an almost unlimited choice of performing variations, from a simple dance in masks to the most elaborate spectacle interspersed with songs, speeches, and pantomimes. As for the actual dances, Robert Copland's *Maner of Dauncynge of Bace Daunces after the Use of Fraunce,* published in 1521 as an appendix to a French grammar, leaves no doubt that the English upper class of that period was thoroughly familiar with continental dance. But whereas the nobility preferred dances of slow, measured, and dignified stature, stylishly performed and modelled upon the standards of the French court, the peasants continued their boisterous dancing, in England as elsewhere, very much as they had for centuries.

In England in the late 16th century, Queen Elizabeth I gave dancing a further boost. She was a skilled practitioner of the galliard and the volta, with its tight embraces by high-leaping couples. She enjoyed watching the English country dances—the chain, ring, and round dances of ancient origin and constantly new invention. These dances apparently provided a continuous infusion of new vitality into court dances. The nobles vied with one another in the execution of the jig, a sprightly and swift dance of "the folk" accompanied by songs. Dancing schools flourished everywhere in London, giving public displays and contributing considerably to the reputation of "the dancing English." Another extremely important contribution to dance was provided by Spain, which in the late 16th and early 17th centuries enjoyed a cultural renaissance. It was the "golden age" of Cervantes in literature, of Lope de Vega and Pedro Calderón de la Barca in the theatre, of El Greco and Diego Velásquez in painting. With the growth of Spain's empire in the Americas, dances of Afro-American origin found their way back to Europe. The sarabande and the chaconne were brought from Central America before 1600. Both were considered outspokenly obscene in their suggestions of sexual encounters. They became extremely popular in the harbours of Andalusia, where they were polished and their pantomimic literalness somewhat moderated. From there they crossed the Pyrenees and were integrated into the canon of the French court dance.

Spread of Spanish dance

Other dances from abroad played major roles in the shaping of Spain's national dances. The *canarie* of African origin certainly sired the Aragonese jota, while the sarabande brought forth the seguidilla. The Afro-Cuban *chica* lived on in the fandango, and the flamenco dances of the Andalusian Gypsies retained their Moorish heritage into the 20th century. It can be presumed that this exchange of dances was not a one-way traffic, that the European conquerors and colonists similarly influenced the dancing habits of the people in other lands.

**The birth of ballet.** Meanwhile, dance became the subject of serious studies in France. A group of writers calling themselves La Pléiade aimed for a revival of the theatre of the ancient Greeks with its music, song, and dance. In Catherine de Médicis (1519–89), the Florentine wife of Henry II, the Italian dancing masters found an influential sponsor in Paris. She called to Paris the Italian musician and dancing master Baltazarini di Belgioioso, who changed his name to Balthazar de Beaujoyeulx (early 16th century to 1587). There had been previous fetes in both France and Italy that offered masquerades, pantomimes, and dances with allegorical and symbolical subjects, but none of them compared to the splendours of the *Ballet comique de la reine* that Beaujoyeulx staged in 1581 for Catherine.

*Ballet comique de la reine*

This "ballet" told the story of the legendary sorceress Circe and her evil deeds. Spoken texts alternated with dances amid magnificently decorative settings. The performers, recruited from the nobility, moved on the floor more like animated costumes than individual dancers. They came together in strikingly designed groups, and they set up geometrical floor patterns that had highly symbolic meanings. (To audiences of the period, for example, three concentric circles represented Perfect Truth, and two equilateral triangles within a circle stood for Supreme Power.) The ballet, which ended in an act of homage to the royal majesties present, had a distinct political moral. Circe had to render her might to the absolutist power of the king of France as the supreme symbol of a peaceful and harmonious world.

The *Ballet comique* launched the species known as *ballet de cour,* in which the monarchs themselves participated. The idealized dances represented the supreme order that France itself, suffering from internal wars, lacked so badly. The steps were those of the social dances of the times, but scholars became aware of how these native materials might be used to propagate the Greek revival. They thoroughly analyzed and systematized the dances, and in 1588 the priest Jehan Tabourot, writing under the pen name Thoinot Arbeau, published his *Orchésographie,* which he subtitled "a treatise in dialogue form by which everyone can easily learn and practice the honest exercise of the dance." This was the first book containing reliable descriptions of how, and to what kind of music, the *basse danse,* pavane, galliard, volta, courante, allemande, gavotte, *canarie, bouffon, moresque,* and 23 different variations of the branle were performed.

## During the 17th, 18th, and 19th centuries

Under kings Louis XIV and Louis XV, France led western Europe into the age of the Rococo in the arts. The Rococo began as a movement toward simplicity and naturalness, a reaction against the stilted mannerisms and preciousness to which the earlier Baroque art was considered to have degenerated. It was a great age of and for dancing, with the minuet the symbol of its emphasis on civilized movement. This formal dance, the perfect execution of which was almost a science in itself, reflected the Rococo idea of naturalness. The statement that "the dance has now come to the highest point of its perfection" by the composer Jean-Philippe Rameau (1683–1764) suggested how conscious the French were of the great strides dance had made. That this was particularly the case in France was confirmed by the English poet and essayist Soame Jenyns (1704–87) in his lines "None will sure presume to rival France, / Whether she forms or executes the dance." None, however, excelled the estimation of his profession by the dancing master in Molière's *Le Bourgeois Gentilhomme* (1670):

> There is nothing so necessary to human beings as the dance . . . Without the dance, a man would not be able to do anything. . . . All the misfortunes of man, all the baleful reverses with which histories are filled, the blunders of politicians and the failures of great leaders, all of this is the result of not knowing how to dance.

### THE MATURING OF BALLET

Dance was finally deemed ready for an academy of its own. In 1661, 13 dancing masters who had been members of a professional guild of medieval origin, together with some musicians, composers, and the makers of instruments, were granted a charter by Louis XIV for the Académie Royale de Danse.

**Technical codifications and dance scholarship.** The academicians were charged with setting up objective standards for perfecting of their arts, with unifying the rules of dance training, and with issuing licenses to dancing instructors. Though the nobility continued for some time to participate in the *ballets de cour,* and Louis himself danced in them until 1669, the dance became more and more the province of highly trained specialists.

After 1700 ballet and social dance took separate paths. But while the ballet continued to absorb new ideas from the folk and social dance, its practitioners and theoreticians looked down on those more common forms. A profusion of books on dance began to appear—treatises, instructions, and analyses as well as the first attempts to record dances by means of written notation. The first history of dance was Claude-François Menestrier's *Des ballets anciens et modernes* ("On Dances Ancient and Modern"; 1682). The second major work of European dance literature, after Arbeau's *Orchésographie,* was Raoul Feuillet's *Chorégraphie, ou l'art de décrire la danse* ("Choreography, or the Art of Describing the Dance"; 1700). It became the standard grammar for the dances practiced at the turn of the century, describing them in minute detail and notating them by a system devised by Feuillet. This indicated the position of the feet and directions, combinations, and floor patterns of the steps and leaps. The notations system was unable, however, to register the movements of the upper parts of the body. Feuillet provided as well a complete definition of the principles of the dance first described by the Académie in the 1660s. These included the *en dehors* (*i.e.,* the turnout of the body and its limbs), the five classical positions of the feet, the port de bras (*i.e.,* the positions and movements of the arms), and the leaps to the *grande élévation,* the aerial movements of the dance.

Early works about ballet

In 1706 Feuillet's influential book was translated into English by John Weaver (1673–1760), a dancer, choreographer, and teacher who worked mainly at the Drury Lane Theatre, London. In 1717 he produced one of the first serious ballets without words, *The Loves of Mars and Venus.* Weaver was the first dance teacher to insist that dance instructors should have a thorough knowledge of human anatomy. In 1721 he published his *Anatomical and Mechanical Lectures upon Dancing,* which became a standard work of international importance. Germany also was represented in the field of dance scholarship, most notably by Leipzig Gottfried Tauber in *Der rechtschaffene Tanzlehrer* ("The Correctly Working Dance Teacher"; 1717). These books strongly emphasized the contributions of dance to general education and manners. In this period dance was considered the basis of all education, and well-to-do parents went to great pains to have their children properly instructed.

**Varieties of the ballet.** As the technical demands of performance became greater and the amateurs gave way to the professionals, performance of the ballet moved from the dance floor onto the stage. There it gradually shed its declamations and its songs and concentrated on telling a story through the gestures of dance and mime alone. But this purifying process took time. For decades different forms of mixed-media spectacles were seen, from the *comédies-ballets* of Molière (1622–73) and the composer Jean-Baptiste Lully (1632–87) to the *opéras-ballets* of André Campra (1660–1744) and Rameau, which were successions of songs and dances on a common theme. The first ballet to be performed without the diversions of speech or song was *Le Triomphe de l'amour* (*The Triumph of Love;* 1681), choreographed by Charles-Louis Beauchamp (1636–*c.* 1719) to Lully's music. Originally a *ballet de cour,* it was revived for the stage with a professional cast. Its star, Mlle Lafontaine, became ballet's first *première danseuse* exactly 100 years after the *Ballet comique* had been produced.

An even more dramatic form known as *ballet d'action* came into being in 1708, when two professional dancers presented an entire scene from the tragedy *Horace* by Pierre Corneille (1606–84) in dance and mime. Weaver's silent ballets, whose expressive dance much impressed English audiences, also encouraged Marie Sallé, a highly ambitious dramatic dancer. Despairing of the *opéras-ballets* of Paris, she went to London, where she performed in pantomimes and produced a miniature dance-drama of her own, *Pygmalion* (1734). In it she appeared in a flimsy muslin dress and loose, flowing hair rather than the heavy costumes and elaborate wigs usually worn by ballerinas. Thus lightened, the dancer was able to move with much greater freedom.

**Early virtuosos of the dance.** The era of the great dancer was at hand. Marie Sallé (1707–56) was the greatest dancer-mime and an important innovator of her day. Her popularity was rivalled by the Brussels-born Marie Camargo (1710–70), who excelled Sallé in lightness and sparkle. She used the entrechat, a series of rapid crossings of the legs that previously had been used only by male dancers. To show off properly her entrechats and other lithe footwork, she shortened her skirt by several inches, thereby contributing to costume reform. Both ballerinas were depicted by Nicolas Lancret (1690–1743), a painter known for his festive scenes, and both were praised by the writer and philosopher Voltaire (1694–1778), who carefully compared their respective virtues. Both, however, were surpassed by the Italian dancer Barberina Campanini (1721–99), whose fame is less adequately recorded in dance history. By 1739, she had taken Paris by storm, demonstrating jumps and turns executed with a speed and brilliance hitherto unknown. She offered ample proof that the Italian school of dance teaching had by no means died out with the earlier exodus of so many of its best practitioners to the French courts. Despite the great public acclaim that these ballerinas attracted, they were overshadowed by Louis Dupré (1697–1744), known as "The Great Dupré" and "the god of the dance." In grace, majesty, and allure, he was unsurpassed, giving the male dancer a prominence he held for a century. Dupré was also the first of a direct line of great dance teachers that was unbroken in the late 20th century.

Sallé, Camargo, and Campanini

Giraudon—Art Resource/EB Inc.



"La Camargo Dancing," oil painting of Marie Camargo by Nicolas Lancret, 1730. In the Musée des Beaux-Arts, Nantes, France.

## THE REIGN OF THE MINUET

In the realm of the social dance, the years between 1650 and 1750 were called "the age of the minuet" by the dance and music historian Curt Sachs.

*The French dance suite.* At the great balls of the French court at Versailles, the minuet was the high point of the festivities, which culminated in a suite of dances. The opening branle, led by the king and his escort, was a measured circling around, one couple after another. Next came the courante, which had been toned down from its earlier rather capricious figurations. Over the years it assumed a continuously greater dignity until it was danced with such gravity and sobriety that it was termed the "doctor dance." It went quickly out of fashion, however, after 1700. Following the courante in the succession was the gavotte, which opened in the form of a round dance. A couple separated to each perform a short solo, then returned to the original circle. Sometimes the suite was extended through an allemande (French: "German"), an old dance form that was introduced into France from the heavily German-speaking province of Alsace in the 1680s. This dance, with its turning couples, the lady on the arm of the gentleman, was a relative of the German ländler and a precursor of the waltz.

The dance suite

*Form of the minuet.* But the unrivalled king of the social dances was the minuet, named from the *pas menu* ("small step"), a term used at least as early as the 15th century. The earliest surviving specimen was composed by Lully in 1663. Mozart composed a series of 12 minuets as late as 1789. It originated as a folk dance in Poitou, but as a court dance it took its form from the courante. Though today it looks mannered, even artificial, in its time it was looked upon as the most beautiful and harmonious of dances, and to execute it perfectly required prolonged and careful study:

> The *minuet* was performed in open couples; spectators and partners were saluted with ceremonial bows. With dainty little steps and glides, to the right and to the left, forward and backward, in quarter turns, approaching and retreating hand in hand, searching and evading, now side by side, now facing, now gliding past one another, the ancient dance play of courtship appears here in a last and almost unrecognizable stylization and refinement. (Curt Sachs, *World History of the Dance,* trans. Bessie Schönberg, W.W. Norton & Co., Inc., 1937.)

In spite of the great popularity of the minuet before the French Revolution, it was the object of much barbed commentary in the late 18th century. Voltaire compared the metaphysical philosophers of his time with the dancers of the minuet, who, in their elegant attire, bow and mince daintily across the room showing off their charms, move without progressing a single step, and end up at the very spot from which they began.

## ENGLISH SOCIAL DANCE

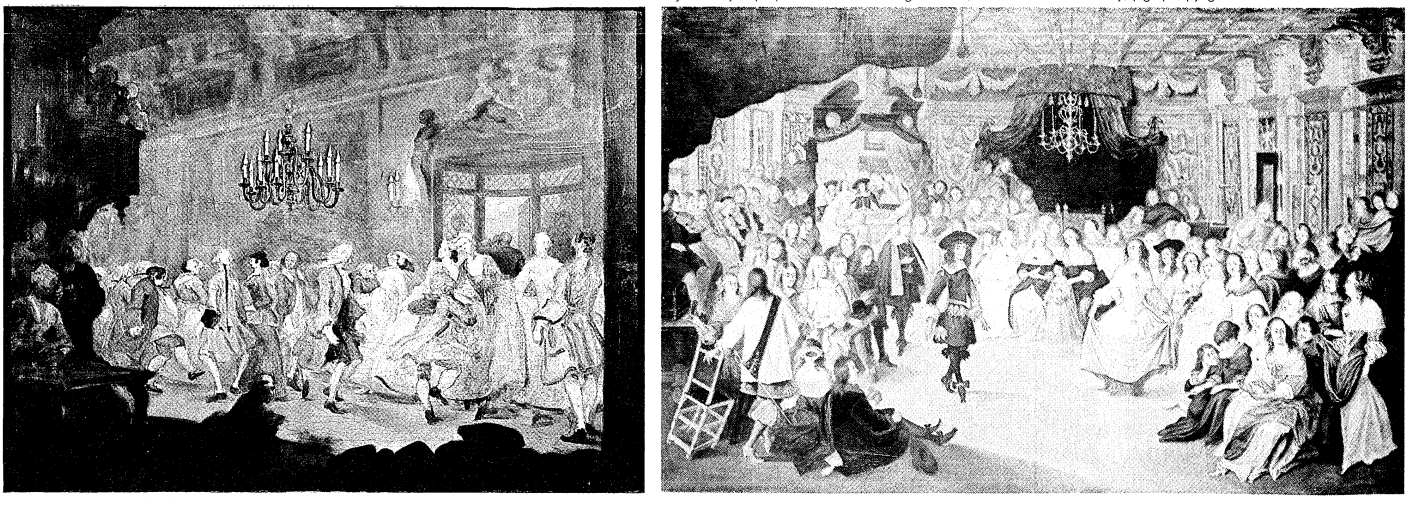
England thoroughly democratized the dance. Though the English Puritanism of the 17th century stigmatized dance as one of man's most sinful occupations, even Oliver Cromwell, lord protector of England under the Puritan rule in the 1650s, could not prevent the appearance of *The English Dancing Master* (issued 1650; dated 1651), by the bookseller and publisher John Playford (1623–*c.* 1686). This was a collection of English traditional dances and tunes. It had 18 editions in 80 years, each one adding to the repertoire. Its 900 choral dances of rustic origin, which formerly had been danced in the open air but were now usually performed indoors, included an enormous variety of forms and patterns. It was written in straightforward, matter-of-fact language, with no discrimination of dances by social class. Its instructions could be understood and its dances performed by anyone. People could enjoy dancing as a playful, sportive activity rather than as an exercise of courtly etiquette.

These "country dances" could as well be city dances, as is suggested by such names as "Mayden Lane" and "Hide Park" from London locales. Others were named for persons—"Parson's Farewell" and "My Lady Foster's Delight"—and that there were foreign influences can be surmised from "The Spanish Jeepsie" and "A la Mode de France." At the same time, native jigs and hornpipes continued to flourish. The English were particularly fond of the Morris dance. This dance may have received its name from the blackened faces of some of its participants, suggestive of the African Moors, but its origins were in the ancient ritual dances. It was a vigorous male dance, in the form of a dance procession through town streets. Its participants, in the disguises of such popular characters as the fool or the Queen of May, wore jingling bells around their ankles and sometimes galloped about on hobby horses. Other dancers wore antlers, tails, and similar animal masking.

The Morris dance

About 1700 the English country dances began to appear on the Continent, where they were somewhat formalized and sometimes substantially altered. In France they were named contredanses. The longways, dances with double lines of dancers facing one another, became *contredanses anglaises;* the rounds became the *contredanses françaises,* which were also known as cotillions and quadrilles. These figure dances, which quickly spread to Spain, Germany, Poland, and other countries, were the dances of the rising middle class. By no means revolutionary in their content, they were nonetheless a distinct declaration of rationality and common sense in dance, a counterbalance to the artificialities and mannerisms of the aristocratic court dances. The orthodox dance teachers might bemoan the decline from the standards that were epitomized in the minuet,

By courtesy of (left) the Southwark Borough Council, South London Art Gallery; (right) Copyright reserved



*17th- and 18th-century dance.*
(Left) Contredanse from "Masked Ball at Wanstead Assembly," oil painting by William Hogarth, *c.* 1745. In the South London Art Gallery. (Right) Courante from "Charles II at a Ball at The Hague," oil painting by Hieronymus Janssens (*c.* 1624–93). In Windsor Castle, Berkshire.

but the townspeople and peasants, unconcerned with such niceties, continued in their uncomplicated knowledge that dancing could be fun.

### DANCE IN COLONIAL AMERICA

*Attitudes.* The English colonists in America had mixed opinions about dance. There was the complete disapproval of those who saw only its inherent licentiousness, but from others came at least a tacit toleration of the obviously irrepressible urge to dance. The South, more heavily populated by colonists with aristocratic backgrounds, was generally more inclined to dance than the North, where religious fervour had motivated much of the migration from England. But what was allowed and even encouraged in Connecticut was strictly forbidden in Massachusetts. The general consensus was apparently that dancing in itself was not bad, but that no punishment could be severe enough for what was regarded as lascivious dancing. The Quakers, who had settled mainly in Pennsylvania, were very much against dancing, and in 1706 they complained bitterly about a dancing and fencing school being tolerated in Philadelphia. They feared that the school's teachings would tend to corrupt their children.

*External and internal influences.* Nonetheless, Playford's *The English Dancing Master* was by no means unknown in America. There were also dancing masters and dancing mistresses to instruct in and lead the dances that had been brought from the Old World. There were society balls in the cities along the coast, and on the inland frontiers the settlers of the widely scattered farmsteads often came together for exuberant feasting and social dancing. Here dancing was considered a socializing virtue expressed in this anonymous observation:

> I really know among us of no custom which is so useful and tends so much to establish the union and the little society which subsists among us. Poor as we are, if we have not the gorgeous balls, the harmonious concerts, the shrill horn of Europe, yet we delight our hearts as well with the simple negro fiddle.

Slave dances

What the colonists saw of American Indian dancing they found very strange and primitive, and there was virtually no exchange of dancing customs between the groups. The situation differed, however, with regard to the black slaves, who in the 17th century had brought their own songs and dances from their native lands in Africa.

During religious holidays in New Amsterdam, blacks danced in the streets to the musical accompaniment of three-stringed fiddles and drums constructed from eel pots and covered with sheep-skins. Dutch families joined in the festivities. When New Amsterdam became New York, however, the English discouraged dancing between whites and blacks; blacks went on to develop the characteristic dance style that would so deeply affect social dancing in the 19th and 20th centuries.

By courtesy of the Museum of Fine Arts, Boston



"Rustic Dance after a Sleigh Ride," oil painting by William Sidney Mount, 1830. In the M. and M. Karolik Collection of the Museum of Fine Arts, Boston.

Early in the 18th century, rather rough theatrical entertainments, acts of acrobatic skill or pantomimes in which dances played an increasing role, began to spread through the American colonies. These often amateurish showings got a mighty boost when the first professional companies came from Europe, about the middle of the century, to perform plays and harlequinades with incidental dances.

### THE RISE OF THE WALTZ

The age of the minuet was followed by that of the waltz. As the French Revolution approached, the minuet, a form that exuded the essence of earlier decades, died a natural death. The English country dances, expressing the self-satisfaction of the bourgeoisie, fared little better.

**The Romantic movement in dance.** The young people, whose preferences led the way in creating new forms, had lived through the revolutionary events of the 1780s and '90s. They now looked to dance as a way to unleash deeper emotion, to satisfy the needs of body and soul, and to mobilize more vital and dynamic expression than that permitted by the sober and decorous rules of the dancing masters. The overflow of feeling and the striving for horizons broader than those understood by the traditional canons of French Rationalism were among the factors that generated the Romantic movement in the arts of Europe. This new direction was clearly expressed in the waltz, a dance filled with the Dionysian spirit.

Dionysius revived

Like much of the spirit of the Romantic movement, the waltz was of German origin. It paralleled the Sturm und Drang movement in German literature, which featured the new forms of prose and poetry by Johann von Goethe and Friedrich Schiller. One of the most glowing advocates of the waltz was Goethe, who time and again praised it, nowhere more than in his novel *Die Leiden des Jungen Werthers* (1774; *The Sorrows of Werter,* 1779): "Never have I moved so lightly. I was no longer a human being. To hold the most adorable creature in one's arms and fly around with her like the wind, so that everthing around us fades away." Even the aristocrats who formed the Congress of Vienna in 1815, which sought to restore law and order to Europe following the upheavals brought on by Napoleon, delighted in performing this earliest of all nonaristocratic ballroom dances.

**Spread of the waltz.** The waltz started as a turning dance of couples. It was especially popular in south Germany and Austria, where it was known under such different names as *Dreher,* ländler, and *Deutscher.* More than any other dance it appeared to represent some of the abstract values of the new era, the ideals of freedom, character, passion, and expressiveness. This may explain somewhat its eruption into the limelight of international popularity. This popularity was scaled in 1787 when it was brought to operatic stage. Vienna became the city of the waltz, for there it surpassed everything in wild fury. It swept over national frontiers, and in 1804 the French were reported to be passionately in love with this light, gliding dance. "A waltz, another waltz" was the common cry from the ballroom floor, for the French could not get enough of the dance.

Some guardians of the public morality disapproved of the "mad whirling" of the waltz and it did not arrive in England until 1812. At the Prussian court in Berlin it was forbidden until 1818, though Queen Luise had danced it while still a princess in 1794. The guardians could do no more than delay its total victory, and it conquered the world without sanction of courts, of dancing masters, or of other powers. After many centuries of leadership, France no longer set the fashions. In 1819 Carl Maria von Weber's *Invitation to the Dance* represented the declaration of love of classical music to the waltz. Shortly thereafter began the age of the Viennese waltz kings, most notably expressed by the Strauss family.

**Offspring and rivals.** The waltz sired a great variety of offspring throughout Europe. Germany developed such variations of the waltz as the schottisch, with turns like those of the waltz. France had its airy *balance valse,* and the Americans later on had their Boston waltz, a slower,



Waltzers in "Le Bal à Bougival," oil painting by Pierre Renoir, 1883. In the Museum of Fine Arts, Boston.
By courtesy of the Museum of Fine Arts, Boston

gliding variant. About 1840 a serious rival to the waltz emerged in the polka, a Bohemian dance that took its name from the Czech word *půlka,* "half step." It was full of fiery vigour. Another Bohemian folk dance finding favour in the dance halls was the *rejdovák* or redowa, while Poland's mazurka and krakowiak enjoyed great popularity. No ball could be concluded without a galop, in which couples galloped through the hall with accelerated polka steps, an exhausting exercise that required considerable reserves of stamina.

## FOUNDATIONS OF MODERN BALLET

The ideals of naturalness, character, soul, passion, and expressiveness came to govern the ballet.

**Noverre and his contemporaries.** The French dancer-choreographer-teacher Jean-Georges Noverre (1727–1810) was the first major reformer of ballet. He defined his artistic positions in *Lettres sur la danse et sur les ballets* (*Letters on Dancing and Ballets*), published in 1760 and continuously reprinted ever since. He worked in Paris, London, Stuttgart, and Vienna, and his influence spread as far as St. Petersburg. He preached the dignity of the ballet and tried to purge it of its excessive artificialities and conventions. He choreographed subjects of mythology and history in highly dramatic narrative forms. He collaborated with some of the major composers of the period, including Mozart, on his ballets.

The influence of Noverre

Noverre was not alone, and the others around him were full of the same zest to give a new meaning to ballet. In Vienna he had a feud with the Italian choreographer Gasparo Angiolini (1731–1803) over Noverre's reforms of the *ballet d'action.* Angiolini claimed these for his teacher, the Austrian choreographer Franz Hilverding (1710–68). In Bordeaux, Noverre's pupil Jean Dauberval premiered in 1789 *La Fille mal gardée* (*The Ill-Guarded Maiden*), usually called *Vain Precautions* in English, which became the first durable ballet comedy. It introduced the *demi-caractère* dance, which featured what were considered to be "true-to-life" characters. In London, still another pupil, Charles Didelot, created in 1796 *Flore et Zéphyre.* This was the first attempt to bestow on the individual dances within the ballet a certain period and local coloration, and to break the uniformity of step and movement of the corps de ballet by assigning individual tasks to its various members. Later, Didelot thoroughly reformed the ballet school in St. Petersburg, which had existed since 1738. There he also created the first ballets of the Russian national repertory. Among these were the first ballets to be based on the works of the Russian writer Alexandr Pushkin (1799–1837), whose stories continued to be used as ballet libretti for many decades.

In Milan, Salvatore Viganò, who had worked under Dauberval and Didelot and who had choreographed in 1801 the first performance of Beethoven's *Creatures of Prometheus,* was praised by the French writer Stendhal for his thrilling ballets based, among other subjects, on Shakespeare's *Othello* and *Coriolanus.* He was followed by Carlo Blasis, who was more noted as a teacher and theoretician. His *Traité élémentaire, théorique, et pratique de l'art de la danse* (1820; *Elementary Treatise upon the Theory and Practice of the Art of Dancing*) became the standard work of ballet teaching for the 19th century. In 1837 he founded the Imperial Ballet Academy, through which Milan became, with Paris and St. Petersburg, a third ballet centre of world renown.

**The Romantic ballet.** During the 1830s and '40s the Romantic movement flooded ballet stages with nature spirits, fairies, and sylphids. The cult of the ballerina replaced that of the male dancer, whose last and greatest representative had been the Italian dancer Gaetano Vestris (1729–1808). The techniques of female dancing were greatly improved. Skirts were shortened further, and blocked shoes permitted toe dancing. Choreographers strove for a more expressive vocabulary and to highlight the individual qualities of their dancers.

*La Sylphide* (1836) stated a main subject of the Romantic ballet, the fight between the real world and the spiritual world. This theme was enhanced and expanded in *Giselle* (1841) and *Ondine* (1843). Paris and London were the taste setters, and it was London that in 1845 witnessed the *Pas de quatre,* for which the French choreographer Jules Perrot brought together, for four performances, four of the greatest ballerinas of the day, the Italians Marie Taglioni (1804–84), Carlotta Grisi (1819–99), and Fanny Cerrito (1817–1909), and Lucile Grahn (1819–1907). After this the decline of Romantic ballet was rapid, at least in these cities. It continued to flourish into the early 1860s, however, in Copenhagen under the choreographer Auguste Bournonville, whose repertoire was kept alive by the Royal Danish Ballet into the second half of the 20th century. Russia, under the French-born Marius Petipa (1819–1910) and his Russian aide Lev Ivanov (1834–1901), built a world-famous ballet culture of its own. Linked at first with Paris, it gradually developed its own balletic idiom from native as well as imported sources. The high point of the classical ballet under the tsars was reached with the St. Petersburg productions of *The Sleeping Beauty* (1890), *The Nutcracker* (1892), and *Swan Lake* (1895), all with music composed by Peter Tchaikovsky, and *Raymonda* (1898), composed by Aleksandr Glazunov (1865–1936). While the ballet prospered in St. Petersburg and Moscow, it waned in Paris. Its ballerinas even appeared in male roles, as in *Coppélia* (1870). In Milan the extravaganzas of Luigi Manzotti (1838–1905) offered anything but dancing while glorifying the progress of mankind through material discoveries and inventions. The 19th century also saw an unprecedented increase in travel and in cross-cultural influences. Many seemingly exotic dance styles arrived on the Western scene. Troupes from as far as India and Japan appeared at expositions in Paris and London, starting a lively interest in folk and ethnic dancing. Ballerinas of the Romantic ballet travelled from one European city to another, from Milan to London to Moscow. The Austrian dancer Fanny Elssler toured the Americas in the early 1840s for two years, visiting Havana twice. The great choreographers, too, went from city to city. The language of dance became a medium of international communication without regard for difference in geography or spoken language.

Russian rise, Western decline

THEATRE AND BALLROOM DANCE

Other dance entertainments of a lighter kind gained immense popularity during the 19th century. In Paris the all-female cancan became the rage. Its electrifying high kicks were shockingly exhibitionistic and titillating. After 1844 it became a feature of the music halls, of revues, and of operetta. It was raised to musical prominence by operetta composer Jacques Offenbach (1819–80) and vividly depicted by the painter Henri de Toulouse-Lautrec (1864–1901). London enjoyed the Alhambra and Empire ballets, which were mostly classical ballets with spectacular productions. But it was America that provided the widest variety. There were patriotic spectacles, popular after the Revolutionary War, such as *The Patriot, or Liberty Asserted,* in which dance figured prominently.

Giraudon—Art Resource/EB Inc.



Cancan with "Jane Avril Dancing," oil on cardboard by Henri de Toulouse-Lautrec, 1892. In the Louvre, Paris.

More important and of longer range results were the minstrel shows, extravaganzas, burlesques, and vaudevilles. These represented a confluence of a wide assortment of dance and theatrical influences, especially from black culture. White men affected black faces and black dances, and black men affected the faces and dances of whites. Dances were tap and soft-shoe, the buck-and-wing, and similar routines. Theatrical productions offered all kinds of dance, from European-imported ballets through entirely native exhibitions of female beauty verging on the striptease. American dancers began to establish reputations both in America and Europe. The ballerina Augusta Maywood (1825–76?) was the first American dancer to perform at the Paris Opéra.

During the 19th century there was also an enormous increase in the number of public ballrooms and other dancing establishments in the fast-growing cities of the West. Here were first encountered American imports such as the barn dance, then called the military schottische; the Washington Post, a very rapid two-step in march formation; and the cakewalk, which contorted the body to degrees previously unknown. For the first time Europe found in the New World a new infusion of blood for its dancing veins. The tempo of the dances quickened, reflecting perhaps the quickening pace of life and the great social changes of the century.

## The 20th century

Two trends were evident during the first years of the 20th century, before World War I. As if aware of some impending catastrophe, the wealthy society of Europe and the Americas indulged itself to the full in quicker waltzes and faster galops. At the same time, it tried to revive the minuet, gavotte, and pavane, producing only pale and lifeless evocations. There had hardly ever been such a frantic search for new forms, such radical questioning of values previously taken for granted, such a craze among the youth of all nations for individual expression and a more dynamic way of life. All the arts were deeply influenced by the rapid accumulation of discoveries in the physical and social sciences and an increasing awareness of social problems.

New directions in the dance

Overall, it was an incredibly lively time for the dance, which never before had generated so many new ideas or attracted so many people. The ballet was completely rejuvenated under the leadership of Russian impresario Sergey Diaghilev (1872–1929). It inspired some of the foremost composers and painters of the day, becoming the primary theatre platform for the most up-to-date work in the arts. Proponents of another reform movement, "modern dance," took their cue from the American dancer Isadora Duncan to strike in another way at the artificialities that Romantic ballet had generated. It took vigorous roots in Germany, where its expressionistic forms earned it the name *Ausdruckstanz* ("expressionistic dance"). The ballroom dances were thoroughly revolutionized through infusions of new vitality from South American, Creole, and black sources. With the overwhelming popularity of Afro-American jazz, the entire spirit and style of social dancing altered radically, becoming vastly more free, relaxed, and intimate through the following decades.

There was also a renewal of interest in the folk dances that had been the expressions of the common people in past centuries. This was fostered partly through special folk-dance societies, partly through various youth movements that saw that these dances might assist in shaping new community feelings. Theatrical dance of all kinds, from the highly stylized, centuries-old dances of the Orient to exhibitions of naked female flesh, reached new heights of popularity.

DIAGHILEV AND HIS ACHIEVEMENTS

The artistic consequences of Diaghilev's Ballets Russes were enormous. Diaghilev's interest in dance began while he was a member of a small circle of intellectuals in St. Petersburg who fought to bring Russia's arts onto the wider European scene. The painters Alexandre Benois and Léon Bakst were his earliest collaborators.

**The Ballets Russes.** The Russian ballet troupe that Diaghilev took to Paris in 1909 boasted some of the best dancers from the imperial theatres in St. Petersburg and Moscow. They set all Paris ablaze. No living person could remember ballets of such quality. For the next 20 years the Ballets Russes, which never appeared in Russia, became the foremost ballet company in the West. Diaghilev, who never choreographed a ballet himself, possessed a singular flair for bringing the right people together. He became the focus of the ballet world, striving for the integration of dance, music, visual design, and libretto into a "total work of art" in which no one element dominated the others.

Between 1909 and 1929, the contributions of many of the finest dancers and choreographers and of some of the most avant-garde, style-setting painters and composers made the Diaghilev company the centre of creative artistic activity. The group became a haven for Russian dancers who emigrated after the 1917 Revolution. It was the first large, permanently travelling company that operated on a private basis and catered to a cosmopolitan Western clientele.

Michel Fokine (1880–1942) was the first choreographer to put Diaghilev's ideas into practice. He worked with contemporary composers, notably the Russian Igor Stravinsky

The work of Fokine

(1882–1971) and the Frenchman Maurice Ravel (1875–1937). Stravinsky composed the score for two of Fokine's best known ballets, *L'Oiseau de feu* (*The Firebird;* 1910) and *Petrushka* (1911); both are based on old Russian folktales. He drew also upon many eminent composers of the past, such as the Russians Aleksandr Borodin (1833–87) and Nicolay Rimsky-Korsakov (1844–1908), and the Pole Frédéric Chopin (1810–49). His major scenic artists were Benois and Bakst, whose contributions to theatrical design had influences beyond the sphere of ballet. Among his dancers were the Russians Anna Pavlova (1881–1931), who left after the 1909 season to dance with her own company throughout the West as well as the Orient, and Vaslav Nijinsky (1890–1950), who succeeded Fokine as the company's choreographer. A classic dancer, Nijinsky was an anticlassic choreographer, specializing in turned-in body movements and in unusual footwork. In 1912 Nijinsky choreographed *L'Après-midi d'un faune* (*Afternoon of a Faun*) to music written by the French Impressionist composer Claude Debussy (1862–1918)—it is the only Nijinsky ballet still performed. The following year he created *Le Sacre du printemps* (*The Rite of Spring*) to Stravinsky's music. The unconventional ballet was considered scandalous and nearly caused a riot at its Paris premiere.

After Nijinsky's career was cut short by his insanity, the dancer Léonide Massine (1896–1979) assumed the role of choreographer. He quickly became noted for his wit and the precisely characterizing gestures of his dancers. His musical collaborators included Stravinsky; Manuel de Falla (1876–1946), whose work was full of the flavour of his native Spain; Ottorino Respighi (1879–1936), noted for his musical evocations of Italian landscapes; and Erik Satie (1866–1925), a Frenchman known for his originality and eccentricity. Massine's designers included leading painters of the School of Paris such as André Derain (1880–1954) and Pablo Picasso (1881–1973). Following Diaghilev's death, Massine created a furor in the 1930s with his ballets based on symphonies by Tchaikovsky and Johannes Brahms. It was considered inappropriate to use symphonic music for dance, and the incorporation of the style and movements of modern dance into the plotless ballets added to the controversy.

Another of Diaghilev's choreographers was Nijinsky's sister, Bronisława Nijinska (1891–1972), who became famous for her massive ensemble groupings, used to great effect in *Les Noces* (*The Wedding;* 1923), and her talent for depicting the follies of contemporary society. Diaghilev's last choreographic discovery was the Russian-trained George Balanchine (1904–83). Balanchine's 1928 ballet, *Apollon Musagète,* was the first of many collaborations with Stravinsky and led the way to the final enthronement of neoclassicism as the dominant choreographic style of the following decades.

**The continuing tradition.** When Diaghilev died his was no longer the only ballet company touring the world. Anna Pavlova's company visited places in Europe, the Americas, Australia, and the Orient that had never heard of, let alone seen, ballet. A troupe assembled by Ida Rubinstein (1885–1960) had Nijinska as a choreographer and Stravinsky and Ravel as composers. The Ballets Suédois featured, from 1920 to 1925, another group of avant-garde, largely French and Italian composers, painters, and writers. New dancers came from the schools in Paris, London, and Berlin that were directed by self-exiled Russian teachers. Important developments took place in London, where Dame Marie Rambert (1888–1982), a Diaghilev dancer, founded the Ballet Rambert, and Ninette de Valois founded the company that became in 1956 the Royal Ballet. In New York, Balanchine set up the School of American Ballet in 1934. From it he drew the dancers for the several companies that led ultimately to the founding of the New York City Ballet in 1948.

**The Soviet ballet.** Although Diaghilev's achievements were ignored there, the Soviet Union in the 1920s abounded with the daring choreographic experiments of Fyodor Lopukhov (1886–1973) and others. Despite the official imposition of "socialist realism" as the criterion of artistic acceptability in 1932, ballet gained enormous popularity with the Soviet people. They loved their dancers, who were superbly trained by generations of teachers under the leadership of Agrippina Vaganova (1879–1951).

## MODERN DANCE

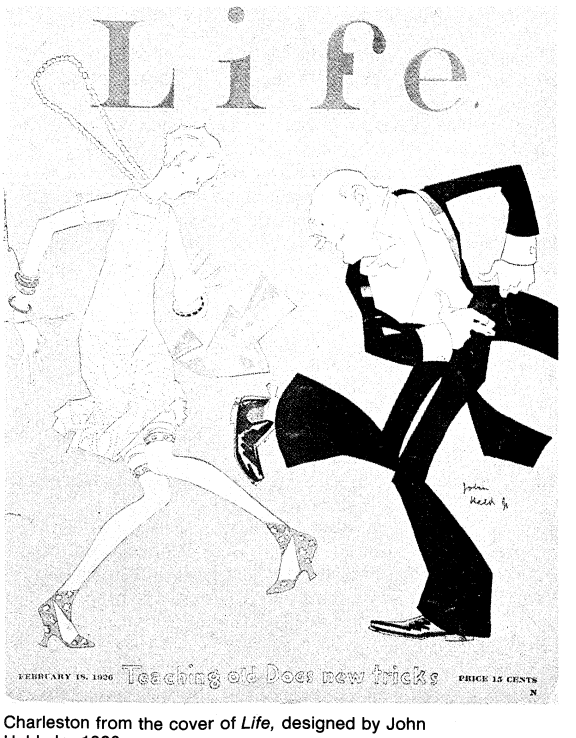
Despite the recovery of ballet from its sterility in the late 19th century, other dancers questioned the validity of an art form so inescapably bound to tradition by its relatively limited vocabulary. They wished to change radically the culture concept of expressive stage dancing. In a period of women's emancipation, women stepped to the front as propagandists for the new dance and toppled the conventions of the academic dance. They advocated a dance that arose from the dancer's innermost impulses to express himself or herself in movement. They took their cues from music or such other sources as ancient Greek vase paintings and the dances of Oriental and American Indian cultures.

The pioneers of this new dance were Isadora Duncan (1877–1927), who stormed across European stages in her loosely flying tunic, inspiring a host of disciples and imitators, and Ruth St. Denis (1877–1968), who surprised American and European audiences with her Oriental-style dances. With her partner Ted Shawn (1891–1972) she founded (1915) Denishawn, which, as a school and performing company, became the cradle of America's early protagonists of modern dance; notable among them were Martha Graham, Doris Humphrey, and Charles Weidman (1901–75).

American and German schools

In the German *Ausdruckstanz* the central figure was Rudolf Laban (1879–1958), who was more a theoretician and teacher than a choreographer. His researches into the physiological impulses to movement and rhythm crystallized in a formidable system of physical expression. His system of dance notation, known most widely as Labanotation, provided the first means for writing down and copyrighting choreographies. His most prolific disciples were Kurt Jooss (1901–79) and Mary Wigman (1886–1973). Jooss became known for his dances containing strong elements of social commentary. Wigman had also studied with Émile Jaques-Dalcroze (1865–1950), who developed eurythmics, a system of movement originally designed to train professional musicians in rhythm. Wigman blended features of both men's techniques into her own new style of dance. When she toured the United States in the 1930s, Americans became aware that they were not alone in their search for new forms of expressive

Culver Pictures




Charleston from the cover of *Life,* designed by John Held, Jr., 1926.

dance. She left behind one of her closest collaborators, Hanya Holm, who became another major figure on the American scene.

Across the United States schools opened, producing small groups of dancers who performed on college campuses and on small stages in the cities. Each choreographer and company brought different materials, artistic points of view, and performing styles to the dance. Perhaps the single element common to all of the many facets of modern dance was the search for new and valid forms of artistic expression.

### NEW RHYTHMS, NEW FORMS

The changes in the social climate that were evident in the new century had a notable influence on the ballrooms.

**Latin-American and jazz dances.** The younger generation in Europe eagerly took up the more vivacious, dynamic, and passionate social dances from the New World. The turning dances of the 19th century gave way to such walking dances as the two-step, the one-step, or turkey trot, the fox-trot, and the quickstep, performed to the new jagged rhythms. These rhythms were African in origin, whether from the Latin-American tangos and rumbas or from the Afro-American jazz. It is impossible to say how far this music was reduced in intensity from its original forms, but its influence was enormous in shaping the ragtime popular before 1918, the syncopated rhythms and mellower swing that followed it, the acrobatic jitterbug of the 1930s and 1940s, and the rock and roll of the next decades.

**Dance contests and codes.** After 1912, when ballroom tango became the rage of the dancing world, even elegant hotels invited their clientele to their "tango teas." In fashionable restaurants professional dance couples demonstrated the new styles. In 1892 New York City saw one of the first cakewalk competitions, and in 1907 Nice advertised the first tango contest. After the first world dance competition in 1909, in Paris, this became an annual event, which in 1913 lasted for two weeks. But it was England that acted as arbiter of taste for the new movements in social dance. There the first dance clubs, like the Keen Dancers' Society (later the Boston Club), were founded in 1903. In 1904 the Imperial Society of Teachers of Dancing was established, and in 1910 the periodical *Dancing Times* made its bow. After World War I the English version of the fox-trot was acknowledged as the essence of the internationally acclaimed "English style." Victor Silvester's *Modern Ballroom Dancing* (1928) became the handbook of the dancing world until it was succeeded by Alex Moore's *Ballroom Dancing* (1936). The English style involved strict definitions for the five standard dances—quickstep, waltz, fox-trot, tango, and blues—to which were added after 1945 the Latin-American rumba, samba, calypso, and cha-cha-cha. What was left of the social barriers existing in 1900 between the exclusive and the popular dancing establishments was swept away.

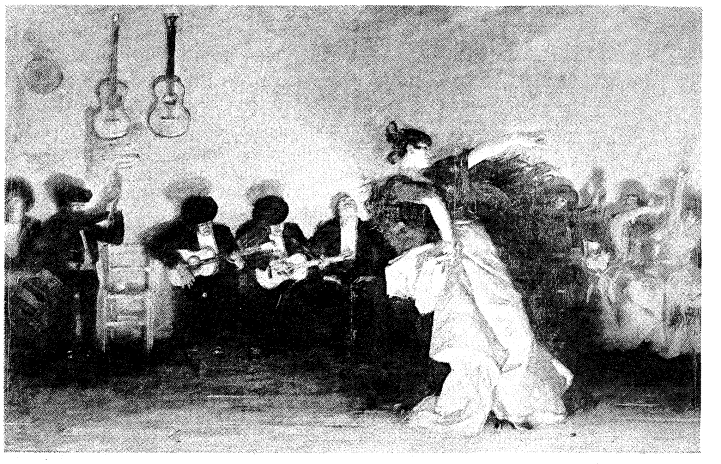
The English style

Many observers were indignant about the changes taking place. Even so liberal a historian as Curt Sachs could not refrain from stating:

> Since the Brazilian *maxixe* of 1890 and the *cakewalk* of 1903 broke up the pattern of turns and glides that dominated the European round dances, our generation has adopted with disquieting rapidity a succession of Central American dances, in an effort to replace what has been lost to modern Europe: multiplicity, power, and expressiveness of movement to the point of grotesque distortion of the entire body. . . . All [of these dances are] compressed into even movement, all emphasizing strongly the erotic element, and all in that glittering rhythm of syncopated four-four measures classified as *ragtime.* (From Curt Sachs, *op. cit.,* pp. 444–445.)

Sachs went on to note the rapid rise and fall in popularity of individual dances and suggested an impermanence to the entire movement.

**Effect on folk dancing.** As social dancing spread with the advent of the radio and the phonograph, the regions where genuine folk dancing was practiced became fewer. It continued least corrupted by the new forms in those countries outside the mainstream of Western urbanization and industrialization. Spain maintained its vigorous tradition of flamenco dancing, and in Hungary the composers Béla Bartók (1881–1945) and Zoltán Kodály (1882–1967) collected the remnants of a wealth of folk song and dance folklore. Minority groups such as the Basques in Spain did likewise to maintain their identity against the overpowering influences of their neighbours.



"El Jaleo," oil painting by John Singer Sargent (1856–1925). In the Isabella Stewart Gardner Museum, Boston.
By courtesy of the Isabella Stewart Gardner Museum, Boston

Folk dancing remained a vital reality in the Soviet Union, especially in those European and Asiatic provinces that had distinctive ethnic populations and were far removed from Moscow, Leningrad, and other centres with Western contacts. In the industrial nations of Europe and the Americas, special nationwide councils and societies were founded to preserve the traditional folk dance that was under threat of extinction.

**Experiments.** Technological progress itself became the subject of dance and dancing. In the Soviet Union, there were experiments during the 1920s with dances created to express urban traffic, the accuracy of machine work, and the grandeur of skyscrapers. In Germany, the painter Oskar Schlemmer (1888–1943) realized his vision of a dance of pure, geometric form in the *Triadisches Ballet* performed in Stuttgart in 1922. In 1926 a sound vision of the technological ages, *Ballet mécanique* (*Mechanical Ballet*), by the American composer George Antheil (1900–59), was scored for mechanical pianos, automobile horns, electric bells, and airplane propellers. It was written not for the live dancer but for an animated film.

### THE DANCE SINCE 1945

Dance of all kinds emerged from World War II, more vital and more expansive than before.

**Social dance.** Postwar social dancing was marked by continuing exuberance and enthusiasm. Dances such as the jitterbug, popular throughout the 1930s and '40s, included lively turns and lifts with rapid footwork. Motion pictures and television helped to spread such rock and roll dances as the twist more rapidly and widely than dances had travelled before. A characteristic of this new generation of jazz-based dances was the lack of bodily contact between the participants, who vibrated their legs, gesticulated with their hands, swung their shoulders, and twitched their heads.

Jazz-based dances

Many observers attempted to draw social implications of all kinds from these dances, which began to spread also among the youth of the Communist countries of Eastern Europe and Asia. Among the more interesting interpretations was that of Frances Rust:

> . . . this type of dancing can be thought of as "progression" rather than "regression." Historically speaking, country-dancing of a communal or group nature gives way, with the break up of communities, to partnered-up ballroom dancing with a concentration on couples rather than groups. This, in turn, is now replaced amongst young people by partner-less dancing, which, although individualistic, seems none-the-less, to be rooted in a striving for community feeling and group solidarity (from *Dance in Society;* Routledge and Kegon Paul, 1969).

In the mid-1970s disco dancing brought a return to dancing with a partner in choreographed steps in dances such as the hustle and the bump. Disco was influenced by modern jazz dancing and became rather athletic, incorporating kicks, turns, and even backflips. Athletic dance moves continued to develop, especially in the 1980s in break dancing, an acrobatic style that featured intricate contortions, mime-like walking moves, and rapid spins on the neck and shoulders. Less complicated dance styles also were found, such as slam dancing, in which the dancers hurled their bodies against each other's, and dances such as the pogo, in which dancers jumped in place to the music's rhythm. Partner dancing never disappeared completely, however, and was especially prominent in the "western-swing" dancing of American country and western music.

**Dance in the theatre.** On the postwar ballet scene there were no revolutionary developments such as those of Diaghilev earlier in the century. The classical ballet style reigned supreme throughout the West and in the Soviet Union. The leading Soviet companies, the Bolshoi Ballet in Moscow and the Kirov Ballet in Leningrad, continued the great 19th-century Russian tradition of full-length dramatic ballets. The popularity of ballet and the establishment of many apparently permanent companies made inevitable wide variations in style and content. International tours were resumed on a large scale. There was also considerable interaction in terms of style and personnel between ballet and modern dance. This was especially true at the New York City Ballet, founded in the late 1940s by George Balanchine and Lincoln Kirstein. The company presented many new works by choreographers such as Jerome Robbins, William Dollar, and Sir Frederick Ashton (the latter principal choreographer and director of Britain's Royal Ballet), but it was Balanchine's style that dominated the company through great ballets such as *The Nutcracker* (1954) and *Don Quixote* (1965) and more abstract works such as *Agon* (1957) and *Jewels* (1967). After Balanchine's death in 1983, Robbins and dancer-choreographer Peter Martins became ballet masters in chief and continued the company's tradition, while at the same time introducing new works.

Another leading company was the American Ballet Theatre, founded in 1939. Its repertoire combined a broad range of works by choreographers such as Antony Tudor and Eliot Feld and balanced classical ballets with established contemporary pieces and newly commissioned works. After the retirement of co-directors Lucia Chase and Oliver Smith in 1980, the great Soviet-born U.S. dancer Mikhail Baryshnikov was named artistic director.

The development of modern dance continued in the work of innovative dancer-choreographers who formed their own companies to explore new styles of dance. Martha Graham's expressive dance centred on mythic and legendary themes, whether ancient, as in *Primitive Mysteries* (1931) and *Clytemnestra* (1958), or modern, as in *Appalachian Spring* (1944). One of Graham's dancers, Merce Cunningham, concentrated on abstract movement that minimized emotional content and experimented with techniques for achieving purity of movement, including arranging sequences of dance steps by flipping a coin. Twyla Tharp was another experimental choreographer whose early work reduced dance to its most fundamental level—movement through open areas, often without music. Her later work melded classical ballet and jazz with modern dance. A different perspective was offered by Arthur Mitchell, who left the New York City Ballet to found the Dance Theatre of Harlem, a company with strong roots in classical ballet.

The American musical theatre benefitted from the techniques of theatrical dance forms. Choreographers of ballet and modern dance also created works for musical comedy. Agnes deMille choreographed *Rodeo* (1942) for the Ballet Russe de Monte Carlo and later created many modern works for the American Ballet Theatre; she also choreographed the stage and film versions of Rodgers and Hammerstein's *Oklahoma!* (1943 and 1955, respectively) and the stage versions of *Carousel* (1945) and *Paint Your Wagon* (1951). Jerome Robbins contributed excellent works for the stage in *The King and I* (1951) and *Fiddler on the Roof* (1967), as well as the stage and film versions of *West Side Story* (1957 and 1961).

Companies presenting dances from India, Ceylon, Bali, and Thailand were no longer considered exotic on Western stages, and their influences contributed to both ballet and modern dance. Numerous ensembles sprang up, their repertoires based on traditional national dances adapted for the stage. Many were modelled on the Soviet Moiseyev folk-dance company, which attracted large audiences during its frequent European and American tours. Similar companies existed in several eastern European countries, in Israel, and in some new African nations, as well as in Brazil, Mexico, and the Philippines.

From the beginning of the 20th century the dance scene became extremely multifaceted and colourful. If some of its manifestations appeared contradictory, that could be regarded as proof of its vitality. No other century granted dance so prominent a role among its social activities. Indications of this prominence included a vast increase in dance research and writing, the opening of colleges and universities in America to special dance faculties, and establishment in the Soviet Union of institutes for the study of choreography. And dance notation promised great advances in recording specific choreographies and as a basic linguistic tool in dance education.

**BIBLIOGRAPHY.** CURT SACHS, *A World History of the Dance* (1937, reprinted 1965; originally published in German, 1933), the most comprehensive, systematic, and factual history of dance in all its epochs and forms, with special emphasis on its earliest beginnings and close attention to dance accompaniment; W.F. RAFFÉ, *Dictionary of the Dance* (1965, reissued 1975), detailed descriptions of the particular dances, their background, and history; ANATOLE CHUJOY and P.W. MANCHESTER (eds.), *The Dance Encyclopedia,* rev. ed. (1967), a collection of articles on all forms of dancing—particularly detailed in its coverage of ballet, including entries on specific productions, artistic biographies, and histories of ballet in various countries; HORST KOEGLER, *The Concise Oxford Dictionary of Ballet,* 2nd ed. (1982), a comprehensive reference work; LINCOLN KIRSTEIN, *Dance: A Short History of Classic Theatrical Dancing* (1935, reprinted 1970), a very thorough book on the pre-balletic forms of dance as well as classic theatrical dance; WALTER SORELL, *The Dance Through the Ages* (1967), a general, readable survey of the worldwide dance scene from prehistoric times through today, with superb pictures of ancient and modern dance; A.H. FRANKS, *Social Dance: A Short History* (1963), the first attempt at relating the origins and developments of the most important social dance forms to their social environment; FRANCES RUST, *Dance in Society* (1969), a study giving documentary evidence of the social dances and their relationships to the changing structures of society, with emphasis on the English scene and the teenage explosion in dance during the 1960s.

(H.Ko./Ed.)

# Dante

Dante Alighieri (1265–1321) is Italy's greatest poet and also one of the towering figures in western European literature. He is best known for his monumental epic poem, *La commedia,* later named *La divina commedia* (*The Divine Comedy*). This great work of medieval literature is a profound Christian vision of man's temporal and eternal destiny. On its most personal level, it draws on the poet's own experience of exile from his native city of Florence; on its most comprehensive level, it may be read as an allegory, taking the form of a journey through hell, purgatory, and paradise. The poem amazes by its array of learning, its penetrating and comprehensive analysis of contemporary problems, and its inventiveness of language and imagery. By choosing to write his poem in Italian rather than in Latin, Dante decisively influenced the course of literary development. Not only did he lend a voice to the emerging lay culture of his own country, but Italian became the literary language in western Europe for several centuries. In addition to poetry Dante wrote important theoretical works ranging from discussions of rhetoric to moral philosophy and political thought. He was fully conversant with the classical tradition, drawing for his own purposes on such writers as Virgil, Cicero, and Boethius. But, most unusual for a layman, he also had an impressive command of the most recent scholastic philosophy and of theology. His learning and his personal involvement in the heated political controversies of his age led him to the composition of *De monarchia,* one of the major tracts of medieval political philosophy.

**Early life and the Vita nuova.** Most of what is known about Dante's life he has told himself. He was born in Florence in 1265 under the sign of Gemini (between May 21 and June 20) and remained devoted to his native city all his life. Dante describes how he fought as a cavalryman against the Ghibellines, a banished Florentine party supporting the imperial cause. He also speaks of his great teacher Brunetto Latini and his gifted friend Guido Cavalcanti, of the poetic culture in which he made his first artistic ventures, his poetic indebtedness to Guido Guinizelli, the origins of his family in his great-great-grandfather, Cacciaguida, whom the reader meets in the central cantos of the *Paradiso* (and from whose wife the family name, Alighieri, derived), and, going back even further, of the pride that he felt in the fact that his distant ancestors were descendants of the Roman soldiers who settled along the banks of the Arno.

Alinari—Mansell/Art Resource



"Dante and His Work" by Domenico di Michelino, 1465. In Florence cathedral.

Yet Dante has little to say about his more immediate family. There is no mention of his father or mother, brother or sister in *The Divine Comedy.* A sister is possibly referred to in the *Vita nuova,* and his father is the subject of insulting sonnets exchanged in jest between Dante and his friend Forese Donati. Because Dante was born in 1265 and the exiled Guelfs, to whose party Dante's family adhered, did not return until 1266, Dante's father apparently was not a figure considerable enough to warrant exile. Dante's mother died when he was young, certainly before he was 14. Her name was Bella, but of which family is unknown. Dante's father then married Lapa di Chiarissimo Cialuffi and they produced a son, Francesco, and a daughter, Gaetana. Dante's father died prior to 1283, since at that time Dante, having come into his majority, was able as an orphan to sell a credit owned by his father. The elder Alighieri left his children a modest yet comfortable patrimony of property in Florence and in the country. About this time Dante married Gemma Donati, to whom he had been betrothed since 1277.

Ghibellines and Guelfs

Dante's life was shaped by the long history of conflict between the imperial and papal partisans called, respectively, Ghibellines and Guelfs. Following the middle of the 13th century the antagonisms were brutal and deadly, with each side alternately gaining the upper hand and inflicting gruesome penalties and exile upon the other. In 1260 the Guelfs, after a period of ascendancy, were defeated in the battle of Montaperti (*Inferno* X, XXXII), but in 1266 a force of Guelfs, supported by papal and French armies, was able to defeat the Ghibellines at Benevento, expelling them forever from Florence. This meant that Dante grew up in a city brimming with postwar pride and expansionism, eager to extend its political control throughout Tuscany. Florentines compared themselves with Rome and the civilization of the ancient city-states.

Influence of Brunetto Latini

Not only did Florence extend its political power, but it was ready to exercise intellectual dominance as well. The leading figure in Florence's intellectual ascendancy was a returning exile, Brunetto Latini. When in the *Inferno* Dante describes his encounter with his great teacher, this is not to be regarded as simply a meeting of one pupil with his master but rather as an encounter of an entire generation with its intellectual mentor. Latini had awakened a new public consciousness in the prominent figures of a younger generation, including Guido Cavalcanti, Forese Donati, and Dante himself, encouraging them to put their knowledge and skill as writers to the service of their city or country. Dante readily accepted the Aristotelian assumption that man is a social (political) being. Even in the *Paradiso* (VIII.117) Dante allows as being beyond any possible dispute the notion that things would be far worse for man were he not a member of a city-state.

A contemporary historian, Giovanni Villani, characterized Latini as the "initiator and master in refining the Florentines and in teaching them how to speak well, and how to guide our republic according to political philosophy [*la politica*]." Despite the fact that Latini's most important book, *Li Livres dou Trèsor* (1262–66; *The Tresor*), was written in French (Latini had passed his years of exile in France), its culture is Dante's culture; it is a repository of classical citation. The first part of Book II contains one of the early translations in a modern European vernacular of Aristotle's *Ethics.* On almost every question or topic of philosophy, ethics, and politics Latini freely quotes from Cicero and Seneca. And, almost as frequently, when treating questions of government, he quotes from the book of Proverbs, as Dante was to do. The Bible, as well as the writings of Aristotle, Cicero, and Seneca, as represented in Latini's work, were the mainstays of Dante's early culture.

Of these Rome presents the most inspiring source of identification. The cult of Cicero began to develop alongside that of Aristotle; Cicero was perceived as not only preaching but as fully exemplifying the intellectual as citizen. A second Roman element in Latini's legacy to become an important part of Dante's culture was the love of glory, the quest for fame through a wholehearted devotion to excelling. For this reason, in the *Inferno* (XV) Latini is praised for instructing Dante in the means by which man makes himself immortal, and in his farewell words Latini commits to Dante's care his *Tresor,* through which he trusts his memory will survive.

Dante was endowed with remarkable intellectual and aesthetic self-confidence. By the time he was 18, as he himself says in the *Vita nuova,* he had already taught himself the art of making verse (chapter III). He sent an early sonnet, which was to become the first poem in the *Vita nuova,* to the most famous poets of his day. He received several responses, but the most important one came from Cavalcanti, and this was the beginning of their great friendship.

Influence of Cavalcanti

As in all meetings of great minds the relationship between Dante and Cavalcanti was a complicated one. In chapter XXX of the *Vita nuova* Dante states that it was through Cavalcanti's exhortations that he wrote his first book in Italian rather than in Latin. Later, in the *Convivio,* written in Italian, and in *De vulgari eloquentia,* written in Latin, Dante was to make one of the first great Renaissance defenses of the vernacular. His later thinking on these matters grew out of his discussions with Cavalcanti, who prevailed upon him to write only in the vernacular. Because of this intellectual indebtedness, Dante dedicated his *Vita nuova* to Cavalcanti—to his best friend (*primo amico*).

Later, however, when Dante became one of the priors of Florence, he was obliged to concur with the decision to exile Cavalcanti, who contracted malaria during the banishment and died in August 1300. In the *Inferno* (X) Dante composed a monument to his great friend, and it is as heartrending a tribute as his memorial to Latini. In both cases Dante records his indebtedness, his fondness, and his appreciation of their great merits, but in each he is equally obliged to record the facts of separation. In order to save himself, he must find (or has found) other, more powerful aesthetic, intellectual, and spiritual sponsorship than that offered by his old friends and teachers.

Beatrice

One of these spiritual guides, for whom Cavalcanti evidently did not have the same appreciation, was Beatrice, a figure in whom Dante created one of the most celebrated fictionalized women in all of literature. In keeping with the changing directions of Dante's thought and the vicissitudes of his career, she, too, underwent enormous changes in his hands—sanctified in the *Vita nuova,* demoted in the canzoni (poems) presented in the *Convivio,* only to be returned with more profound comprehension in *The Divine Comedy* as the woman credited with having led Dante away from the "vulgar herd."

*La vita nuova* (*c.* 1293; *The New Life*) is the first of two collections of verse that Dante made in his lifetime, the other being the *Convivio.* Each is a *prosimetrum,* that is, a work composed of verse and prose. In each case the prose is a device for binding together poems composed over about a 10-year period. The *Vita nuova* brought together Dante's poetic efforts from before 1283 to roughly 1292–93; the *Convivio,* a bulkier and more ambitious work, contains Dante's most important poetic compositions from just prior to 1294 to the time of *The Divine Comedy.*

The *Vita nuova,* which Dante called his *libello,* or small book, is a remarkable work. It contains 42 brief chapters with commentaries on 25 sonnets, one *ballata,* and four canzoni; a fifth canzone is left dramatically interrupted by Beatrice's death. The prose commentary provides the frame story, which does not emerge from the poems themselves (it is, of course, conceivable that some were actually written for other occasions than those alleged). The story is simple enough, telling of Dante's first sight of Beatrice when both are nine years of age, her salutation when they are 18, Dante's expedients to conceal his love for her, the crisis experienced when Beatrice withholds her greeting, Dante's anguish that she is making light of him, his determination to rise above anguish and sing only of his lady's virtues, anticipations of her death (that of a young friend, the death of her father, and Dante's own premonitory dream), and finally the death of Beatrice, Dante's mourning, the temptation of the sympathetic *donna gentile* (a young woman who temporarily replaces Beatrice), Beatrice's final triumph and apotheosis, and, in the last chapter, Dante's determination to write at some later time about her "that which has never been written of any woman."

The impersonal style of the *Vita nuova*

Yet with all of this apparently autobiographical purpose the *Vita nuova* is strangely impersonal. The circumstances it sets down are markedly devoid of any historical facts or descriptive detail (thus making it pointless to engage in too much debate as to the exact historical identity of Beatrice). The language of the commentary also adheres to a high level of generality. Names are rarely used—Cavalcanti is referred to three times as Dante's "best friend"; Dante's sister is referred to as "she who was joined to me by the closest proximity of blood." On the one hand Dante suggests the most significant stages of emotional experience, but on the other he seems to distance his descriptions from strong emotional reactions. The larger structure in which Dante arranged poems written over a 10-year period and the generality of his poetic language are indications of his early and abiding ambition to go beyond the practices of local poets.

**Dante's intellectual development and public career.** A second contemporary poetic figure behind Dante was Guido Guinizelli, the poet most responsible for altering the prevailing local, or "municipal," kind of poetry. Guinizelli's verse provided what Cavalcanti and Dante were looking for—a remarkable sense of joy contained in a refined and lucid aesthetic. What increased the appeal of his poetry was its intellectual, even philosophical, content. His poems were written in praise of the lady and of *gentilezza,* the virtue that she brought out in her admirer. The conception of love that he extolled was part of a refined and noble sense of life. It was Guinizelli's influence that was responsible for the poetic and spiritual turning point of the *Vita nuova.* As reported in chapters XVII to XXI, Dante experienced a change of heart, and rather than write poems of anguish, he determined to write poems in praise of his lady, especially the canzone "Donne ch'avete intelletto d'amore" ("Ladies Who Have Understanding of Love"). This canzone is followed immediately by the sonnet "Amore e 'l cor gentil sono una cosa" ("Love and the Noble Heart Are the Same Thing"), the first line of which is clearly an adaptation of Guinizelli's "Al cor gentil ripara sempre amore" ("In Every Noble Heart Love Finds Its Home"). This was the beginning of Dante's association with a new poetic style, the *dolce stil nuovo* ("the sweet new style"), the significance of which—the simple means by which it transcended the narrow range of the more regional poetry—he dramatically explains in the *Purgatorio* (XXIV).

*Dolce stil nuovo*

This interest in philosophical poetry led Dante into another great change in his life, which he describes in the *Convivio.* Looking for consolation following the death of Beatrice, Dante reports that he turned to philosophy, particularly to the writings of Boethius and Cicero. But what was intended as a temporary reprieve from sorrow became a lifelong avocation and one of the most crucial intellectual events in Dante's career. The *donna gentile* of the *Vita nuova* was transformed into Lady Philosophy, who soon occupied all of Dante's thoughts. He began attending the religious schools of Florence in order to hear disputations on philosophy, and within a period of only 30 months "the love of her [philosophy] banished and destroyed every other thought." In his poem "Voi che 'ntendendo il terzo ciel movete" ("You Who Through Intelligence Move the Third Sphere") he dramatizes this conversion from the sweet old style, associated with Beatrice and the *Vita nuova,* to the rigorous, even severe, new style associated with philosophy. This period of study gave expression to a series of canzoni that were eventually to form the poetic basis for the philosophic commentary of the *Convivio.*

Another great change was Dante's more active political involvement in the affairs of the commune. In 1295 he became a member of the guild of physicians and apothecaries (to which philosophers could belong), which opened

his way to public office. But he entered the public arena at a most perilous time in the city's politics. As it had been during the time of the Guelf and Ghibelline civil strife, in the 1290s Florence once again became a divided city. The ruling Guelf class of Florence became divided into a party of "Blacks," led by Corso Donati, and a party of "Whites," to which Dante belonged. The Whites gained the upper hand and exiled the Blacks.

Political activity

There is ample information concerning Dante's activities following 1295. In May 1300 he was part of an important embassy to San Gimignano, a neighbouring town, whose purpose it was to solidify the Guelf league of Tuscan cities against the mounting ambitions of the new and embattled pope Boniface VIII. When Dante was elected to the priorate in 1300, he presumably was already recognized as a spokesman for those in the commune determined to resist the Pontiff's policies. Dante thus experienced a complete turnabout in his attitudes concerning the extent of papal power. The hegemony of the Guelfs—the party supporting the Pope—had been restored in Florence in 1266 by an alliance forged between the forces of France and the papacy. By 1300, however, Dante had come to oppose the territorial ambitions of the Pope, and this in turn provided the intellectual motivation for another, even greater change: Dante, the Guelf moderate, would in time, through his firsthand experience of the ill effects of papal involvement in political matters, become in the *Convivio,* in the later polemical work the *Monarchia,* and most importantly throughout *The Divine Comedy,* one of the most fervently outspoken defenders of the position that the empire does not derive its political authority from the pope.

Events, moreover, propelled Dante into further opposition to papal policies. A new alliance was formed between the papacy, the French (the brother of King Philip IV, Charles of Valois, was acting in concert with Boniface), and the exiled Black Guelfs. When Charles of Valois wished permission to enter Florence, the city itself was thrown into political indecision. In order to ascertain the nature of the Pope's intentions, an embassy was sent to Rome to discuss these matters with him. Dante was one of the emissaries, but his quandary was expressed in the legendary phrase "If I go, who remains; if I remain, who goes?" Dante was outmaneuvered. The Pope dismissed the other two legates and detained Dante. In early November 1301 the forces of Charles of Valois were permitted entry to Florence. That very night the exiled Blacks surreptitiously reentered Florence and for six days terrorized the city. Dante learned of the deception at first in Rome and then more fully in Siena. In January 1302 he was called to appear before the new Florentine government and, failing to do so, was condemned, along with three other former priors, for crimes he had not committed. Again failing to appear, on March 10, 1302, Dante and 14 other Whites were condemned to be burned to death. Thus Dante suffered the most decisive crisis of his life. In *The Divine Comedy* he frequently and powerfully speaks of this rupture; indeed, he makes it the central dramatic act toward which a long string of prophecies points. But it is also Dante's purpose to show the means by which he triumphed over his personal disaster, thus making his poem into a true "divine comedy."

**Exile, the Convivio, and the De monarchia.** Information about Dante's early years in exile is scanty; nevertheless, enough is known to provide a broad picture. It seems that Dante at first was active among the exiled White Guelfs in their attempts to seek a military return. These efforts proved fruitless. Evidently Dante grew disillusioned with the other Florentine outcasts, the Ghibellines, and was determined to prove his worthiness by means of his writings and thus secure his return. These are the circumstances that led him to compose *Il convivio* (*c.* 1304–07; "The Banquet").

The plan of the *Convivio*

Dante projected a work of 15 books, 14 of which would be commentaries on different canzoni. He completed only four of the books. The finished commentaries in many ways go beyond the scope of the poems, becoming a compendium of instruction with much of the random display of an amateur in philosophy. Dante's intention in the *Convivio,* as in *The Divine Comedy,* was to place the challenging moral and political issues of his day into a suitable ethical and metaphysical framework.

Book I of the *Convivio* is in large part a stirring and systematic defense of the vernacular. (The unfinished *De vulgari eloquentia* [*c.* 1304–07; *Concerning Vernacular Eloquence*], a companion piece, presumably written in coordination with Book I, is primarily a practical treatise in the art of poetry based upon an elevated poetic language.) Dante became the great advocate of its use and in the final sentence of Book I he accurately predicts its glorious future:

> This shall be the new light, the new sun, which shall rise when the worn-out one shall set, and shall give light to them who are in shadow and in darkness because of the old sun, which does not enlighten them.

The revolution Dante described was nothing less than the twilight of the predominantly clerical Latin culture and the emergence of a lay, vernacular urban literacy. Dante saw himself as the philosopher-mediator between the two, helping to educate a newly enfranchised public readership. The Italian literature that Dante heralded was soon to become the leading literature and Italian the leading literary language of Europe, and they would continue to be that for more than three centuries.

In the *Convivio* Dante's mature political and philosophical system is nearly complete. In this work Dante makes his first stirring defense of the imperial tradition and, more specifically, of the Roman Empire. He introduces the crucial concept of *horme,* that is, of an innate desire that prompts the soul to return to God. But it requires proper education through examples and doctrine. Otherwise it can become misdirected toward worldly aims and society torn apart by its destructive power. In the *Convivio* Dante establishes the link between his political thought and his understanding of human appetite: given the pope's craving for worldly power, at the time there existed no proper spiritual models to direct the appetite toward God; and given the weakness of the empire, there existed no law sufficient to exercise a physical restraint on the will. For Dante this explains the chaos into which Italy had been plunged, and it moved him, in hopes of remedying these conditions, to take up the epic task of *The Divine Comedy.*

But a political event occurred that at first raised tremendous hope but then plunged Dante into still greater disillusionment. In November 1308 Henry, the count of Luxembourg, was elected king of Germany, and in July 1309 the French pope, Clement V, who had succeeded Boniface, declared Henry to be king of the Romans and invited him to Rome, where in time he would be crowned Holy Roman emperor in St. Peter's Basilica. The possibility of once again having an emperor electrified Italy; and among the imperial proponents was Dante, who saw approaching the realization of an ideal that he had long held: the coming of an emperor pledged to restore peace while also declaring his spiritual subordination to religious authority. Within a short time after his arrival in Italy in 1310 Henry VII's great appeal began to fade. He lingered too long in the north, allowing his enemies to gather strength. Foremost among the opposition to this divinely ordained moment, as Dante regarded it, was the commune of Florence.

The prospect of a new emperor

During these years Dante wrote important political epistles—evidence of the great esteem in which he was held throughout Italy, of his personal authority, as it were—in which he exalted Henry, urging him to be diligent, and condemned Florence. In subsequent action, however, which was to remind Dante of Boniface's duplicity, Clement himself turned against Henry. This action prompted one of Dante's greatest polemical treatises, his *De monarchia* (*c.* 1313; *On Monarchy*) in which he expands the political arguments of the *Convivio.* In the embittered atmosphere caused by Clement's deceit Dante turned his argumentative powers against papal insistence on its superiority over the political ruler, that is, against the argument that the empire derived its political authority from the pope. In the final passages of the *Monarchia* Dante writes that the ends designed by Providence for

man are twofold: one end is the bliss of this life, which is conveyed in the figure of the earthly paradise; the other is the bliss of eternal life, which is embodied in the image of a heavenly paradise. Yet despite their different ends, these two purposes are not unconnected. Dante concludes his *Monarchia* by assuring his reader that he does not mean to imply "that the Roman government is in no way subject to the Roman pontificate, for in some ways our mortal happiness is ordered for the sake of immortal happiness." Dante's problem was that he had to express in theoretical language a subtle relationship that might be better conveyed by metaphoric language and historical example. Surveying the history of the relationship between papacy and empire, Dante pointed with approval to specific historical examples, such as Constantine's good will toward the church. Dante's disappointment in the failed mission of Henry VII derived from the fact that Henry's original sponsor was apparently Pope Clement and that conditions seemed to be ideal for reestablishing the right relationship between the supreme powers.

Dante's view of imperial and papal authority

**The Divine Comedy.** Dante's years of exile were years of difficult peregrinations from one place to another—as he himself repeatedly says, most effectively in *Paradiso* [XVII], in Cacciaguida's moving lamentation that "bitter is the taste of another man's bread and . . . heavy the way up and down another man's stair." Throughout his exile Dante nevertheless was sustained by work on his great poem, possibly begun prior to 1308 and completed just before his death in 1321. In addition, in his final years Dante was received honourably in many noble houses in the north of Italy, most notably by Guido Novello da Polenta, the nephew of the remarkable Francesca, in Ravenna. There at his death Dante was given an honourable burial attended by the leading men of letters of the time, and the funeral oration was delivered by Guido himself.

The plot of *The Divine Comedy* is simple: a man, generally assumed to be Dante himself, is miraculously enabled to undertake an ultramundane journey, which leads him to visit the souls in Hell, Purgatory, and Paradise. He has two guides: Virgil, who leads him through the *Inferno* and *Purgatorio,* and Beatrice, who introduces him to *Paradiso.* Through these fictional encounters taking place from Good Friday evening in 1300 through Easter Sunday and slightly beyond, Dante learns of the exile that is awaiting him (which had, of course, already occurred at the time of the writing). This device allowed Dante not only to create a story out of his pending exile but also to explain the means by which he came to cope with his personal calamity and to offer suggestions for the resolution of Italy's troubles as well. Thus, the exile of an individual becomes a microcosm of the problems of a country, and it also becomes representative of the fall of man. Dante's story is thus historically specific as well as paradigmatic.

The basic structural component of *The Divine Comedy* is the canto. The poem consists of 100 cantos, which are grouped together into three sections, or canticles, *Inferno, Purgatorio,* and *Paradiso.* Technically there are 33 cantos in each canticle and one additional canto, contained in the *Inferno,* which serves as an introduction to the entire poem. For the most part the cantos range from about 136 to about 151 lines. The poem's rhyme scheme is the terza rima (*aba, bcb, cdc,* etc.) Thus, the divine number of three is present in every part of the work.

Comparison with the *Odyssey* and the *Aeneid*

Dante's *Inferno* differs from its great classical predecessors in both position and purpose. In Homer's *Odyssey* (Book XII) and Virgil's *Aeneid* (Book VI) the visit to the land of the dead occurs in the middle of the poem because in these centrally placed books the essential values of life are revealed. Dante, while adopting the convention, transforms the practice by beginning his journey with the visit to the land of the dead. He does this because his poem's spiritual pattern is not classical but Christian: Dante's journey to Hell represents the spiritual act of dying to the world, and hence it coincides with the season of Christ's own death. (In this way, Dante's method is similar to that of Milton in *Paradise Lost,* where the flamboyant but defective Lucifer and his fallen angels are presented first.) The *Inferno* represents a false start during which Dante, the character, must be disabused of harmful values that somehow prevent him from rising above his fallen world. Despite the regressive nature of the *Inferno,* Dante's meetings with the roster of the damned are among the most memorable moments of the poem: the Neutrals, the virtuous pagans, Francesca da Rimini, Filipo Argenti, Farinata degli Uberti, Piero delle Vigne, Brunetto Latini, the simoniacal popes, Ulysses, and Ugolino impose themselves upon the reader's imagination with tremendous force.

The visit to Hell is, as Virgil and later Beatrice explain, an extreme measure, a painful but necessary act before real recovery can begin. This explains why the *Inferno* is both aesthetically and theologically incomplete. For instance, readers frequently express disappointment at the lack of dramatic or emotional power in the final encounter with Satan in canto XXXIV. But because the journey through the *Inferno* primarily signifies a process of separation and thus is only the initial step in a fuller development, it must end with a distinct anticlimax. In a way this is inevitable because the final revelation of Satan can have nothing new to offer: the sad effects of his presence in human history have already become apparent throughout the *Inferno.*

Purgatory as the realm of spiritual renewal

In the *Purgatorio* the protagonist's painful process of spiritual rehabilitation commences; in fact, this part of the journey may be considered the poem's true moral starting point. Here the pilgrim Dante subdues his own personality in order that he may ascend. In fact, in contrast to the *Inferno,* where Dante is confronted with a system of models that needs to be discarded, in the *Purgatorio* few characters present themselves as models; all of the penitents are pilgrims along the road of life. Dante, rather than being an awed if alienated observer, is an active participant. If the *Inferno* is a canticle of enforced and involuntary alienation, in which Dante learns how harmful were his former allegiances, in the *Purgatorio* he comes to accept as most fitting the essential Christian image of life as a pilgrimage. As Beatrice in her magisterial return in the earthly paradise reminds Dante, he must learn to reject the deceptive promises of the temporal world.

Despite its harsh regime, the *Purgatorio* is the realm of spiritual dawn, where larger visions are entertained. Whereas in only one canto of the *Inferno* (VII), in which Fortuna is discussed, is there any suggestion of philosophy, in the *Purgatorio,* historical, political, and moral vistas are opened up. It is, moreover, the great canticle of poetry and the arts. Dante meant it literally when he proclaimed, after the dreary dimensions of Hell: "But here let poetry rise again from the dead." There is only one poet in Hell proper and not more than two in the *Paradiso,* but in the *Purgatorio* the reader encounters the musicians Casella and Belacqua and the poet Sordello and hears of the fortunes of the two Guidos, Guinizelli and Cavalcanti, the painters Cimabue and Giotto, and the miniaturists. In the upper reaches of Purgatory, the reader observes Dante reconstructing his classical tradition and then comes even closer to Dante's own great native tradition (placed higher than the classical tradition) when he meets Forese Donati, hears explained—in an encounter with Bonagiunta da Lucca—the true resources of the *dolce stil nuovo,* and meets with Guido Guinizelli and hears how he surpassed in skill and poetic mastery the reigning regional poet, Guittone d'Arezzo. These cantos resume the line of thought presented in the *Inferno* (IV), where among the virtuous pagans Dante announces his own program for an epic and takes his place, "sixth among that number," alongside the classical writers. In the *Purgatorio* he extends that tradition to include Statius (whose *Thebaid* did in fact provide the matter for the more grisly features of the lower inferno), but he also shows his more modern tradition originating in Guinizelli. Shortly after his encounter with Guinizelli comes the long-awaited reunion with Beatrice in the earthly paradise. Thus, from the classics Dante seems to have derived his moral and political understanding as well as his conception of the epic poem, that is, a framing story large enough to encompass the most important issues of his day, but it was from his native tradition that he acquired the philosophy of love that forms the Christian matter of his poem.

The rejection of Virgil

This means of course that Virgil, Dante's guide, must give way to other leaders, and in a canticle generally de-

void of drama the rejection of Virgil becomes the single dramatic event. Dante's use of Virgil is one of the richest cultural appropriations in literature. To begin, in Dante's poem he is an exponent of classical reason. He is also a historical figure and is presented as such in the *Inferno* (I): ". . . once I was a man, and my parents were Lombards, both Mantuan by birth. I was born *sub Julio,* though late in his time, and I lived in Rome under the good Augustus, in the time of the false and lying gods." Virgil, moreover, is associated with Dante's homeland (his references are to contemporary Italian places), and his background is entirely imperial. (Born under Julius Caesar, he extolled Augustus Caesar.) He is presented as a poet, the theme of whose great epic sounds remarkably similar to that of Dante's poem: "I was a poet and sang of that just son of Anchises who came from Troy after proud Ilium was burned." So, too, Dante sings of the just son of a city, Florence, who was unjustly expelled, and forced to search, as Aeneas had done, for a better city, in his case the heavenly city.

Virgil is a poet whom Dante had studied carefully and from whom he had acquired his poetic style, the beauty of which has brought him much honour. But Dante had lost touch with Virgil in the intervening years, and when the spirit of Virgil returns it is one that seems weak from long silence. But the Virgil that returns is more than a stylist; he is the poet of the Roman Empire, a subject of great importance to Dante, and he is a poet who has become a *saggio,* a sage, or moral teacher.

Though an exponent of reason, Virgil has become an emissary of divine grace, and his return is part of the revival of those simpler faiths associated with Dante's earlier trust in Beatrice. And yet, of course, Virgil by himself is insufficient. It cannot be said that Dante rejects Virgil; rather he sadly found that nowhere in Virgil's work, that is, in his consciousness, was there any sense of personal liberation from the enthrallment of history and its processes. Virgil had provided Dante with moral instruction in survival as an exile, which is the theme of his own poem as well as Dante's, but he clung to his faith in the processes of history, which, given their culmination in the Roman Empire, were deeply consoling. Dante, on the other hand, was determined to go beyond history because it had become for him a nightmare.

In the *Paradiso* true heroic fulfillment is achieved. Dante's poem gives expression to those figures from the past who seem to defy death. Their historical impact continues and the totality of their commitment inspires in their followers a feeling of exaltation and a desire for identification. In his encounters with such characters as his great-great-grandfather Cacciaguida and SS. Francis, Dominic, and Bernard, Dante is carried beyond himself. The *Paradiso* is consequently a poem of fulfillment and of completion. It is the fulfillment of what is prefigured in the earlier canticles. Aesthetically it completes the poem's elaborate system of anticipation and retrospection.

**Assessment and influence.** The recognition and the honour that were the due of Dante's *Divine Comedy* did not have to await the long passage of time: by the year 1400 no fewer than 12 commentaries devoted to detailed expositions of its meaning had appeared. Giovanni Boccaccio wrote a life of the poet and then in 1373–74 delivered the first public lectures on *The Divine Comedy* (which means that Dante was the first of the moderns whose work found its place with the ancient classics in a university course). Dante became known as the *divino poeta,* and in a splendid edition of his great poem published in Venice in 1555 the adjective was applied to the poem's title; thus, the simple *Commedia* became *La divina commedia,* or *The Divine Comedy.*

The *Commedia* becomes *La divina commedia*

Even when the epic lost its appeal and was replaced by other art forms (the novel, primarily, and the drama) Dante's own fame continued. In fact, his great poem enjoys the kind of power peculiar to a classic: successive epochs have been able to find reflected in it their own intellectual concerns. In the post-Napoleonic 19th century, readers identified with the powerful, sympathetic, and doomed personalities of the *Inferno.* In the early 20th century they found the poem to possess an aesthetic power of verbal realization independent of and at times in contradiction to its structure and argument. Later readers have been eager to show the poem to be a polyphonic masterpiece, as integrated as a mighty work of architecture, whose different sections reflect and, in a way, respond to one another. Dante created a remarkable repertoire of types in a work of vivid mimetic presentations, as well as a poem of great stylistic artistry in its prefigurations and correspondences. Moreover, he incorporated in all of this important political, philosophical, and theological themes and did so in a way that shows moral wisdom and lofty ethical vision.

Dante's *Divine Comedy* is a poem that has flourished for more than 650 years: in the simple power of its striking imaginative conceptions it has continued to astonish generations of readers; for more than a hundred years it has been a staple in all higher educational programs in the Western world; and it has continued to provide guidance and nourishment to the major poets of our own times. William Butler Yeats called Dante "the chief imagination of Christendom"; and T.S. Eliot elevated Dante to a preeminence shared by only one other poet in the modern world, William Shakespeare: "[They] divide the modern world between them. There is no third." In fact, they rival one another in their creation of types that have entered into the world of reference and association of modern thought. Like Shakespeare, Dante created universal types from historical figures, and in so doing he considerably enhanced the treasury of modern myth.

### MAJOR WORKS

Because Dante lived and worked long before book printing began, it is not possible to discuss first editions of his works. The following begins with a list of early printed editions of his works in the original languages.

INDIVIDUAL WORKS: *La commedia* (1472); *Vita nuova* (1576).

LYRIC POETRY: *Canzoni e madrigali di Dante, di Mess. Gino da Pistoja e di Giraldo Novello* (1518); *Rime di diversi antichi autori toscani in dieci libri* (1532).

TREATISES: *Convivio di Dante Alighieri fiorentino* (1490); *De vulgari eloquentia libri duo* (1577); *Dantis Aligherii Florentini Monarchia* (1740).

LATIN ECLOGUES: *I versi latini di Giovanni del Virgilio e di Dante Allighieri* (1845).

RECOMMENDED MODERN EDITIONS: *La commedia secondo l'antica vulgata,* ed. by Giorgio Petrocchi (1966–67); *La divina commedia,* ed. by Natalino Sapegno, 3rd ed. (1985), with excellent commentary; *The Divine Comedy of Dante Alighieri,* trans. by John D. Sinclair (1958), with superb small essays for each canto; *The Divine Comedy,* trans. with a commentary by Charles S. Singleton (1970–75), the most useful work in English, published in the Bollingen series.

*La vita nuova,* ed. by Michele Barbi (1932), a critical edition; *Dante's Vita Nuova,* trans. by Mark Musa, new ed. (1973).

*Rime della "Vita nuova" e della giovinezza,* ed. by Michele Barbi and F. Maggini (1956); *Rime della maturità e dell'esilio,* ed. by Michele Barbi and V. Pernicone (1969); *Rime,* ed. by Gianfranco Contini (1965); *Dante's Lyric Poetry,* ed. and trans. by K. Foster and P. Boyde (1967).

*Il convivio,* 2nd ed., ed. by G. Busnelli, G. Vandelli, and Antonio E. Quaglio (1968); *Dante's Convivio,* trans. by William Walrond Jackson (1909).

*De vulgari eloquentia,* ed. by Aristide Marigo (1957); *Dante's Treatise "De Vulgari Eloquentiâ,"* trans. by A.G. Ferrers Howell (1890); *Literary Criticism of Dante Alighieri,* trans. and ed. by Robert S. Haller (1973).

*Monarchia,* ed. by Pier Giorgio Ricci (1965); *On World Government, or, De Monarchia,* trans. by Herbert W. Schneider, 2nd rev. ed. (1957).

*Dante and Giovanni del Virgilio, Including a Critical Edition of the Text of Dante's "Eclogae Latinae" and of the Poetic Remains of Giovanni del Virgilio,* ed. by Philip H. Wicksteed and Edmund G. Gardner (1902).

COLLECTED WORKS: *Le opere di Dante: testo critico della società dantesca italiana,* ed. by Michele Barbi *et al.,* 2nd ed. (1960); *Le opere di Dante Alighieri,* ed. by E. Moore and Paget Toynbee, 5th ed. (1963).

### BIBLIOGRAPHY

*Biographies:* RICARDO J. QUINONES, *Dante Alighieri* (1979, reprinted 1985), an overview; CECIL GRAYSON (ed.), *The World of Dante: Essays on Dante and His Times* (1980); and WILLIAM ANDERSON, *Dante the Maker* (1980), a critical biographical study, with the emphasis on Dante's creative processes. See also PATRICK BOYDE, *Dante, Philomythes and Philosopher: Man in the Cosmos* (1981), an examination of Dante's intellectual concerns.

*Commentaries:* For extracts from early commentaries, see GUIDO BIAGI *et al.* (eds.), *La Divina Commedia nella figurazione artistica e nel secolare commento* (1921–40), issued in separate parts, and its useful bibliography. Modern commentaries on *The Divine Comedy* include those by GIUSEPPE VANDELLI, CARLO GRABHER, MANFREDI PORENA, ATTILIO MOMIGLIANO, and NATALINO SAPEGNO in their respective editions of *La divina commedia.* See also FRANCESCO MAZZONI, *Saggio di un nuovo commento alla Divina Commedia* (1967). For English-speaking readers, the commentary of CHARLES H. GRANDGENT (ed.), *La Divina Commedia di Dante Alighieri* (1933; rev. ed. by CHARLES S. SINGLETON, 1972), is excellent; as is that of CHARLES S. SINGLETON (trans. and ed.), *The Divine Comedy* (1970–75). See also GEORGE HOLMES, *Dante* (1980), a brief study; MARK MUSA, *Advent at the Gates* (1974), a study of seven cantos; and DAVID NOLAN (ed.), *Dante Commentaries: Eight Studies of the Divine Comedy* (1977), and *Dante Soundings: Eight Literary and Historical Essays* (1981).

*Introductory works:* Of the general works available to English-speaking readers, see especially ERNEST HATCH WILKINS and THOMAS GODDARD BERGIN, *A Concordance to the Divine Comedy of Dante Alighieri* (1965). EDWARD S. SHELDON and ALAIN C. WHITE, *Concordanza delle opere italiane in prosa e del Canzoniere di Dante Alighieri* (1905, reprinted 1969 with *Supplementary Concordance to the Minor Italian Works of Dante,* comp. by LEWIS H. GORDON); and EDWARD KENNARD RAND and ERNEST HATCH WILKINS, *Dantis Alagherii Operum Latinorum Concordantiae,* to the Latin works (1912, reprinted 1970), are also useful. PAGET TOYNBEE, *A Dictionary of Proper Names and Notable Matters in the Works of Dante,* new ed., rev. by CHARLES S. SINGLETON (1968), is invaluable. Excellent introductions to Dante include UMBERTO COSMO, *A Handbook to Dante Studies* (1947, reprinted 1978; originally published in Italian, 1947); MICHELE BARBI, *Life of Dante* (1954, reprinted 1966; originally published in Italian, 1933); and THOMAS GODDARD BERGIN, *Dante* (1965, reprinted 1976). Also useful are NICOLA ZINGARELLI, *La vita, i tempi e le opere di Dante,* 3rd ed., 2 vol. (1931); and ALDO VALLONE, *Dante,* 2nd ed. (1981). Essential information is found in *Codice diplomatico dantesco,* ed. by RENATO PIATTOLI, 2nd ed. (1950); and in *Enciclopedia dantesca,* 2nd ed., 5 vol. (1984).

*General studies:* EDWARD MOORE, *Studies in Dante,* 4 vol. (1896–1917, reprinted with new introductory matter ed. by COLIN HARDIE, 1969); PAGET TOYNBEE, *Dante Studies* (1921); BENEDETTO CROCE, *The Poetry of Dante* (1922, reissued 1971; originally published in Italian, 1920); T.S. ELIOT, *Dante* (1929; reprinted 1974); JOHN FRECCERO (ed.), *Dante: A Collection of Critical Essays* (1965); UBERTO LIMENTANI (ed.), *The Mind of Dante* (1965); OXFORD DANTE SOCIETY, *Centenary Essays on Dante* (1965); WILLIAM J. DE SUA and GINO RIZZO (eds.), *A Dante Symposium* (1965); FRANCESCO MAZZONI, *Contributi di filologia dantesca* (1966).

*Specialized studies:* On the *Vita nuova,* see CHARLES S. SINGLETON, *An Essay on the Vita Nuova* (1949, reprinted 1977); on the canzoni, PATRICK BOYDE, *Dante's Style in His Lyric Poetry* (1971); on Dante's philosophical thought, ÉTIENNE GILSON, *Dante the Philosopher* (1948, reissued 1963; originally published in French, 1939); on Dante's political thought, ALESSANDRO PASSERIN D'ENTRÈVES, *Dante as a Political Thinker* (1952, reprinted 1965); EWART K. LEWIS, *Medieval Political Ideas,* 2 vol. (1954, reprinted 1974); CHARLES T. DAVIS, *Dante and the Idea of Rome* (1957); and *Dante's Italy* (1984); on *The Divine Comedy,* WILLIAM H.V. READE, *The Moral System of Dante's Inferno* (1909, reprinted 1969); KARL VOSSLER, *Medieval Culture: An Introduction to Dante and His Times,* 2 vol. (1929, reissued 1970; originally published in German, 1907–10); ERNST ROBERT CURTIUS, *European Literature and the Latin Middle Ages* (1953, reissued 1973; originally published in German, 1948); FRANCIS FERGUSSON, *Dante's Drama of the Mind* (1953, reissued 1981); CHARLES S. SINGLETON, *Dante Studies:* vol. 1, *Commedia: Elements of Structure* (1954, reprinted 1977), and vol. 2, *Journey to Beatrice* (1958); JOHAN CHYDENIUS, *The Typological Problem in Dante* (1958); JOSEPH A. MAZZEO, *Structure and Thought in the Paradiso* (1958, reissued 1968), and *Medieval Cultural Tradition in Dante's Comedy* (1960, reprinted 1968); IRMA BRANDEIS, *The Ladder of Vision: A Study of Dante's Comedy* (1960); HELEN F. DUNBAR, *Symbolism in Medieval Thought and Its Consummation in the Divine Comedy* (1929, reissued 1961); THOMAS GODDARD BERGIN, *Perspectives on the Divine Comedy* (1967), and *A Diversity of Dante* (1969).

*Illuminated manuscripts:* PETER H. BRIEGER, MILLARD MEISS, and CHARLES S. SINGLETON, *Illuminated Manuscripts of the Divine Comedy,* 2 vol. (1969).

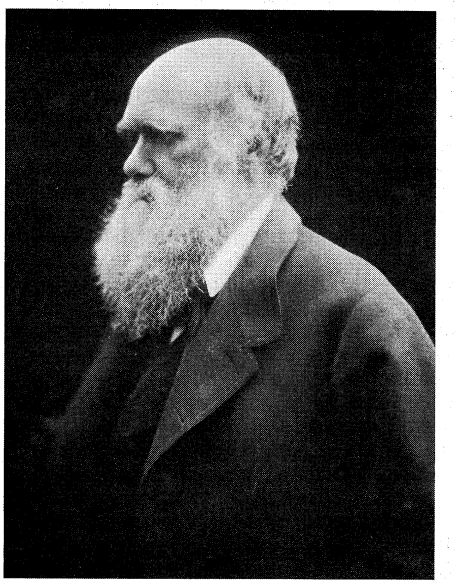
*Bibliographies:* PAUL COLOMB DE BATINES, *Bibliografia dantesca,* trans. from the French, 2 vol. in 3 (1845–46), supplemented by GUIDO BIAGI, *Giunte e correzioni inediti alla Bibliografia dantesca* (1888); continued also in CARLO F. CARPELLINI, *Della letteratura dantesca degli ultimi venti anni dal 1845 a tutto il 1865* (1866); CORNELL UNIVERSITY LIBRARY, *Catalogue of the Dante Collection,* comp. by THEODORE WESLEY KOCH, 2 vol. (1898–1900), and *Catalogue of the Dante Collection Additions 1898–1920,* comp. by MARY FOWLER (1921); and GIULIANO MAMBELLI, *Gli annali delle edizioni dantesche* (1931, reprinted 1965). For post-World War II studies, see ALDO VALLONE, *Gli studi danteschi dal 1940 al 1949* (1950); and ENZO ESPOSITO, *Gli studi danteschi dal 1950 al 1964* (1965). Annual bibliographies of Dante studies published in the United States are printed in *Dante Studies* (annual), published by the Dante Society of America (founded 1881). Bodies specializing in Dante studies have been established in many countries. Apart from the Società Dantesca Italiana (founded 1888), of special interest to English-speaking readers is the Oxford Dante Society (founded 1876).

(R.J.Q.)

# Darwin

Charles Robert Darwin, the naturalist whose discovery of the theory of evolution by natural selection revolutionized biology, was born in Shrewsbury, Eng., on Feb. 12, 1809. He was the son of Robert Waring Darwin, who had one of the largest medical practices outside of London, and the grandson of the physician Erasmus Darwin, the author of *Zoonomia, or the Laws of Organic Life,* and the artisan-entrepreneur Josiah Wedgwood. Darwin enjoyed a secure position in the professional upper middle class that provided him with considerable social and professional advantages.

**Youth and education.** Darwin's mother died when he was eight years old. Otherwise he enjoyed a golden childhood, cosseted and encouraged by adoring sisters, an older brother, and the large Darwin and Wedgwood clans. He was keenly interested in specimen collecting and chemical investigations, but at the Shrewsbury school, where he was an uninspired student, the headmaster, Dr. Samuel Butler, stressed the classics and publicly rebuked Darwin for wasting his time with chemical experiments. At age 16 he was sent to study medicine at the University of Edinburgh, where he was repelled by surgery performed without anesthetics. During his two years in Scotland Darwin benefited from friendships with the zoologist Robert Grant, who introduced him to the study of marine animals, and the geologist Robert Jameson, who fed his growing interest in the history of the Earth.



Darwin, carbon print photograph by Julia Margaret Cameron, 1868.

By courtesy of the International Museum of Photography at George Eastman House, Rochester, N.Y.

Cambridge years

Disappointed by Darwin's lack of enthusiasm for medicine, his father sent him to the University of Cambridge in 1827 to study divinity. At the time Darwin adhered to the conventional beliefs of the Church of England. His academic record at Christ's College was as undistinguished as it had been at Edinburgh. He socialized considerably with hunting, shooting, riding, and sporting friends. Cambridge did not yet offer a degree in the natural sciences, but, guided by his older cousin William Darwin Fox (an entomologist who inspired in him a lifelong passion for collecting beetles), Darwin met the circle of Cambridge scientists led by the cleric-botanist John Stevens Henslow. Soon a regular at Henslow's "open houses," Darwin accompanied him on daily walks and became known as "the man who walks with Henslow." Henslow encouraged Darwin's excitement about science and confidence in his own abilities.

On leaving Cambridge in the spring of 1831 Darwin, in preparation for a scientific trip to the Canary Islands, read Alexander von Humboldt's *Personal Narrative of Travels to the Equinoctial Regions of the New Continent,* a scientific travelogue of a journey to Central and the northern parts of South America. At Henslow's recommendation he accompanied Adam Sedgwick, Woodwardian professor of geology at Cambridge, on a three-week tour of North Wales to learn geologic fieldwork.

In August 1831, at Henslow's recommendation to the Admiralty, Darwin was invited to sail as the unpaid naturalist on HMS *Beagle.* The ship was to survey the east and west coasts of South America and continue to the Pacific islands to establish a chain of chronometric stations. Henslow suggested Darwin as both an acute observer and a companion for the aristocratic young captain, Robert FitzRoy. (The *Beagle* already had a naturalist-surgeon, but one whom FitzRoy found socially unsuitable.) Robert Darwin first refused permission on grounds that it was dangerous and would not advance Charles in his career. But upon the intercession of his brother-in-law, Josiah Wedgwood II, he changed his mind.

On Dec. 27, 1831, Charles Darwin sailed from Plymouth, Eng., on the *Beagle,* a 10-gun brig that had been refitted as a three-masted bark. The voyage, planned for two years, lasted five, during which Darwin kept meticulous notes and sent back geologic and biologic specimens.

**The voyage of the "Beagle."** In a letter to FitzRoy accepting the post Darwin explained that he expected the voyage to be a "second birth." There is no doubt that the years he spent exploring the South American continent and the offshore islands of the Galápagos honed his skills as a collector, observer, and theorist. Often seasick, Darwin rested horizontally in a hammock during the worst motion and spent long periods of time ashore whenever the opportunity arose. He delighted in the exotica of the tropics. Adventurous, he braved his way through armed political rebellions, rode with the gauchos in Argentina, and on collecting and shooting expeditions justified his earlier devotion to sport. He joined the crew in towing the ship's boats upstream and once rescued the expedition by running to save a boat from a tidal wave. He seemed to relish danger and was sustained in the considerable discomfort by a lively curiosity. He wrote to one of his sisters, "We have in truth the world before us. Think of the Andes; the luxuriant forest of the Guayquil [sic]; the islands of the South Sea & New South Wales. How many magnificent & characteristic views, how many & curious tribes of men we shall see.—what fine opportunities for geology & for studying the infinite host of living beings: Is this not a prospect to keep up the most flagging spirit?"

The influence of Charles Lyell's *Principles of Geology*

Darwin had brought his own books to augment the ship's extensive library. The most important scientific work was the first volume of Charles Lyell's *Principles of Geology,* which Henslow had urged him to read though not to believe. Lyell argued that the face of the Earth had changed gradually over long periods of time through the continuing, cumulative effects of local disturbances, such as eruptions, earthquakes, erosion, and deposition. Such disturbances had existed in the distant past and could be observed in the present. This view differed dramatically from that held by most contemporary geologists, who hewed to the belief that changes in the face of the Earth resulted from short-lived events of great violence that could raise mountains or flood the entire planet. During the first months of his journey Darwin was converted to Lyell's views by his own observations.

About 1,800 miles southwest of the Canary Islands the *Beagle* visited São Tiago, a volcanic island in the Cape

Verde Islands. From the harbour Darwin saw a band of white rock extending horizontally at a height of about 45 feet above the base of the sea cliffs. The formation was calcareous and contained numerous shells, almost all of which could be found on the coast. Darwin reasoned that a stream of lava from the ancient volcanoes had flowed over what had been ancient seabed, baking it to form the hard white rock. The whole island had subsequently been heaved up to make the sea cliff from the white band downward. Darwin also realized that the island's surface had been formed by a succession of volcanic events, not a single dramatic one. He discerned an initial subsidence, the settling of the surface around the original craters, its building up from new lava spills from different craters, and further subsidence and building up over a long period of time.

Later, in Chile, Darwin witnessed his first earthquake. He saw the land rise before his eyes. Then, after crossing the Andes in 1835, he wrote to his sister that he could understand "to a certain extent the description & manner of the force, which has elevated this great line of mountains." He had found fossil shells at an elevation of 12,000 feet, and he theorized that a chain of suboceanic volcanoes had poured forth enormous quantities of lava that formed the Andes through a further process of upheaval and fracturing. Darwin marveled at the whole South American continent, which he read as a vast testing ground for Lyell's ideas. To his cousin William Fox he wrote, "Everything in America is on such a grand scale. The same formations extend for 5 or 600 miles without the slightest change—for such geology one requires 6 league boots."

The data Darwin collected on the *Beagle* provided him with material for three books on South American geology. Although his theories of continental change have been superseded by the theory of plate tectonics, his descriptions in letters to Henslow, which Henslow excerpted and read before the Cambridge Philosophical Society and the Geological Society of London, brought him celebrity in scientific circles even before his return.

Darwin's theory concerning the formation of coral reefs

While still on the voyage he challenged Lyell's view of the formation of coral reefs by volcanic action. Darwin contended that the reefs were part of a process of gradual changes in the Earth's crust resulting from the subsidence of some landmasses and corresponding elevations elsewhere. He explained that coral, which only grows in shallow waters, forms a reef by building up on the seafloor as the floor subsides. He predicted that if a whole island sank below the ocean's surface, and the coral continued to grow, a reef would turn into an atoll around a lagoon. Lyell was convinced and supported Darwin's reinterpretation, which deep-sea borings in the 20th century have confirmed.

Darwin's geologic ponderings were important for geology and to his scientific development. Many of the rocks he examined contained fossils, and his constant exposure to the evidence of extinct species and the similarity of many of them to living species kept one problem at the fore: By what mechanism did new species replace extinct ones?

During the voyage Darwin developed confidence in his own observations as well as the ability to grasp a problem and work at it steadily. The isolation of the voyage, combined with the exposure to new phenomena, taught him to think for himself within the familiar scientific culture of his time. He developed a rare combination of strengths: a dedication to careful fact gathering and a propensity to theorize about the facts. His geologic pondering on a continental scale encouraged him to search for universal laws. The voyage of the *Beagle* transformed Darwin into an independent and adventurous scientist who had the courage to embrace the heretical idea of the transmutation of species.

Return to England

**"On the Origin of Species."** When Darwin returned to England in 1836 he was welcomed by the scientific fraternity as a colleague and was promptly made a fellow of the Geological Society. The next year he was elected to its governing council. In 1838 Darwin was elected to the Athenaeum, the exclusive club for men distinguished in literature, art, or science, and in 1839 he was elected to the Royal Society. Through his older brother, Erasmus, he met the historian Thomas Carlyle and the feminist Harriet Martineau. He was also a friend of Charles Babbage, whose computing machine was one of a host of scientific interests.

At this time, however, Darwin began to lead something of a double life. To the world he was busy preparing his *Journal of Researches into the Geology and Natural History of the Various Countries Visited by H.M.S. Beagle,* which was published in 1839. This book, modeled in part on von Humboldt's, established the lucid style enlivened by the sharp descriptions that makes all of Darwin's works both accessible and convincing. Darwin was also preparing his geology books and superintending the analysis and publication by specialists of *The Zoology of the Voyage of H.M.S. Beagle* (published between 1839 and 1843 with the help of a £1,000 government grant). Privately Darwin had begun a remarkable series of notebooks in which he initiated a set of questions and answers about "the species problem." He proceeded to collect facts about species through letters and discussions with breeders, gardeners, naturalists, and zookeepers, as well as through extensive reading.

Notebooks on "the species problem"

Darwin kept this interest secret while he gathered evidence to substantiate his theory of organic evolution. He was mindful of the fate of other unorthodox scientists. He jotted in his notebook, "Mention persecution of early astronomers—then add chief good of individual scientific men, is to push their science a few years in advance only of their age." Darwin's ideas were not only scientifically radical but also could have been interpreted as actionable under the laws governing blasphemy and sedition. England at the time was intensely evangelical, and the natural world was understood as one in which the spirit of God could be seen in the creation of new species of plants and animals that appeared to come into existence to replace those that became extinct. Darwin gradually became intellectually uncomfortable with this view of life as he confronted puzzling evidence.

Upon his return from the voyage Darwin had turned over his specimens to cataloging experts in Cambridge and London. In South America he had found fossils of extinct armadillos that were similar but not identical to the living animals. In Argentina he had seen species vary geographically; for example, the giant ostriches (rheas) on the pampas were replaced to the south in Patagonia by much smaller species, both of which were akin to but different from the African ostrich. He had been disturbed by the fact that the birds and tortoises of the Galápagos Islands off the western coast of Ecuador tended to resemble species found on the nearby continent, while inhabitants of similar neighbouring islands in the Galápagos had quite different animal populations. In London Darwin learned that the finches he had brought from the Galápagos belonged to different species, not merely different varieties, as he had originally believed. He also learned that the mockingbirds were of three distinct species and that the Galápagos tortoises represented at least two species and that, like many of the specimens from the archipelago, they were native to the islands but to neither of the American continents.

After Darwin received these reports, his doubts about the fixity of species crystallized into a belief in transmutation. In March 1837 he confided in his notebook that species changed from one place to another or from one era to the next. He continued analyzing his data, searching for a mechanism for this process. Then in October 1838 Darwin read Thomas Malthus' *An Essay on the Principle of Population.* Malthus argued that population growth is geometric, while the food supply increases only arithmetically, and thus that population increase is always checked by a limited food supply. Darwin recalled in his *Autobiography* his realization that given the struggle for existence everywhere, "favourable variations would tend to be preserved, and unfavourable ones to be destroyed. . . . The result of this would be the formation of new species. Here, then, I had at last got a theory"—the principle of natural selection—"by which to work."

The principle of natural selection

Darwin's originality extended beyond perceiving the savagery of the natural world. Other scientists and philosophers had noted the brutality of species against species,

of the lion devouring the lamb. Darwin saw competition between individuals of a single species. He recognized that within a local population the individual with, for example, the sharper beak, the longer horn, or the brighter feather might have a better chance to survive and reproduce than other individuals. If such advantageous traits were passed on to new generations, they would eventually be predominant in future populations. Darwin thus shifted the focus of evolutionary analysis from between to within species. He saw natural selection as the mechanism by which advantageous variations passed on to succeeding generations and by which the traits of individuals that were less competitive gradually disappeared from populations. (Later generations of biologists came to understand variations within a species as variations in the genes of its individual members, and they explained evolution as the action of natural selection upon genes responsible for advantageous traits.) After he had hit upon natural selection, Darwin was eager to verify it, and he stepped up his inquiries to plant and animal breeders. He hoped to learn from their experience with artificial selection how natural selection worked.

Darwin still faced the problem of divergence, that is, the evolutionary development of dissimilar characteristics in closely related species that have descended from a single organic ancestor. As he had observed during his voyage, divergent species appeared on different landmasses. Darwin solved this puzzle of geographic distribution by assigning the dissemination of populations of ocean islands to the power of wind and water. The theory of the evolution of species thus solved many puzzles in comparative anatomy, embryology, and paleontology. (For further discussion of the details of Darwin's theory, see EVOLUTION, THE THEORY OF; for details of the evolution of humans, see EVOLUTION, HUMAN.)

Lamarck

The idea of organic evolution was not new. It had been suggested a generation earlier by Erasmus Darwin and in France by Buffon, Montesquieu, Maupertuis, Diderot, and most recently Lamarck. Lamarck had drawn the first evolutionary diagram—a ladder leading from unicellular organisms to man. But none of these earlier evolutionists had presented either a mechanism or persuasive evidence for the process. Lamarck offered the hypothesis that spontaneous generation occurs constantly, that organisms possess an "inner feeling" toward perfection, and that the traits an animal acquires to adapt to a changing environment are passed on to its descendants.

Darwin was aware of these other explanations but not influenced by them. His theory was rooted in direct observation and an attempt to discover universal laws. His evolutionary sketch was a branching tree, not a single ladder. Above all, Darwin rejected the prevailing view that organisms are perfectly adapted to their environment. He viewed the natural world, instead, as caught in an incessant struggle between competing individuals that have different degrees of fitness. Others had seen struggles but always between species, never within them. By moving the battle from interspecies struggles to intraspecies competition, Darwin introduced the concept of populations, that is, localized groups consisting of members of a given species in which each individual differs from its sibling. He recognized that it is the competition within a species leading to the survival of individuals with adaptationally advantageous traits that eventually brings about the evolution of a new species.

By 1842 Darwin was confident enough in his theory to draft a short sketch, and in 1844 he composed a longer version, which he showed to his friend, the botanist Joseph Dalton Hooker. Wary of presenting his theory to the public, Darwin spent the next decade concentrating on a treatise on barnacles, in which he hinted but did not actually say that species were the product of natural selection. In the meantime the intellectual atmosphere in England altered and discussions about evolution became commonplace. Darwin still withheld publishing his thesis. When he would have determined that the time was ripe is impossible to know, but the decision was removed from him when on June 18, 1858, he received from Alfred Russel Wallace, a naturalist working in the Malay Archipelago, a paper that perfectly summarized the theory that Darwin had been elaborating for 20 years. Disheartened by this apparent preemption of his life's work, Darwin was saved by his friends and confidants, Lyell, Hooker, and T.H. Huxley, who arranged for a joint paper by Darwin and Wallace to be read to the Linnean Society of London on July 1, 1858.

Alfred Russel Wallace

Darwin then began work on what he called an "abstract" of the larger manuscript that he had begun two years earlier. This abstract, *On the Origin of Species by Means of Natural Selection, or The Preservation of Favoured Races in the Struggle for Life,* was published on Nov. 24, 1859. The first edition sold out immediately, and by 1872 the work had run through six editions. The theory was accepted quickly in most scientific circles. With the exception of holdouts like his old colleague Adam Sedgwick and individuals such as the biologist Richard Owen, who attacked Darwin personally, most opposition was from the clergy. They realized that the theory of evolution was inconsistent with a literal interpretation of the book of Genesis. Orthodox Christians felt threatened by the suggestion that the natural (or living) world worked according to laws as did the physical world. There was no place in Darwin's world for divine intervention, nor was mankind placed in a position of superiority vis-à-vis the rest of the animal world. Darwin saw man as part of a continuum with the rest of nature, not separated by divine injunction.

Publication of *On the Origin of Species*

After the publication of the *Origin* Darwin continued to write, while friends, especially Huxley, defended the theory before the public. In June 1860, at the Oxford meeting of the British Association for the Advancement of Science, Huxley confronted the bishop of Oxford, Samuel Wilberforce, who had been coached by Richard Owen. Wilberforce patronized Huxley, asking whether it was through his father or his mother that he was descended from an ape. Huxley replied that he was not ashamed of having descended from an ape but would be ashamed of an ancestor who used gifts of eloquence in the service of falsehoods. Huxley and Hooker annihilated Wilberforce's position at the Oxford debate and continued spreading what was tantamount to a gospel of evolution.

Darwin completed the elucidation of his theory in his next three books, which were all continuations of the *Origin.* In *The Variation of Animals and Plants Under Domestication* (1868) Darwin proposed his hypothesis of pangenesis (an ill-founded attempt to account for the acquisition of hereditary characteristics, a process eventually explained in the development of cell biology and genetics).

Darwin's views on human evolution

Darwin met the issue of human evolution head-on in *The Descent of Man, and Selection in Relation to Sex* (1871), in which he elaborated on the controversial subject only alluded to in the *Origin.* He expanded the scope of evolution to include moral and spiritual as well as physical traits and underscored man's psychological as well as physiological similarities to the great apes, predicting, "the time will before long come when it will be thought wonderful that naturalists, who were well acquainted with the comparative structure and development of man and other animals, should have believed that each was the work of a separate act of creation."

The second half of the book elaborated upon the theory of sexual selection. Darwin observed that in some species males battle other males for access to certain females. But in other species, such as peacocks, there is a social system in which the females select males according to such qualities as strength or beauty. Twentieth-century biologists have expanded this theory to the selection by females of males who can contribute toward the survival of their offspring; *i.e.,* female selection secures traits that make the next generation more competitive. Although Darwin's description of female choice was roundly rejected by most scientists at the time, he adamantly defended this insight until the end of his life. While not universally accepted today, the theory of female choice has many adherents among evolutionary biologists.

The last of Darwin's sequels to the *Origin, The Expression of the Emotions in Man and Animals* (1872), was an attempt to erase the last barrier presumed to exist between human and nonhuman animals—the idea that the expres-

*The Expression of the Emotions in Man and Animals*

sion of such feelings as suffering, anxiety, grief, despair, joy, love, devotion, hatred, and anger is unique to human beings. Darwin connected studies of facial muscles and the emission of sounds with the corresponding emotional states in man and then argued that the same facial movements and sounds in nonhuman animals express similar emotional states. This book laid the groundwork for the study of ethology, neurobiology, and communication theory in psychology.

**Later works.** Throughout his career Darwin wrote two kinds of books—those with a broad canvas, such as the evolution quartet, and those with a narrow focus, such as the treatise on barnacles. His interests shifted over the years from geology to zoology to botany. In these later works, however, he included theoretical interpretation, whereas his earlier works had contained mostly data. In *On the Various Contrivances by Which British and Foreign Orchids Are Fertilised by Insects* (1862) he demonstrated that plants exhibit complicated characteristics that are adaptive and that increase the survival of a species. One such characteristic, for example, is cross-pollination (the mechanism by which pollen is transferred from one flower to another). In explaining the interdependence of bees and orchids, Darwin noted that flowers that are pollinated by the wind have little colour, while those that need to attract insects have brightly coloured petals and sweet-smelling nectaries. In *The Different Forms of Flowers on Plants of the Same Species* (1877) he observed that flowers in some species differ in the lengths of their anthers and styles, which is another adaptation for cross-pollination. Darwin experimented in his garden at Down House in Kent where he raised two large beds of *Linaria vulgaris,* one from cross-pollinated and the other from self-pollinated seeds, both of which he obtained from the same parent plant. He observed, "To my surprise, the crossed plants when fully grown were plainly taller and more vigorous than the self-fertilized ones." He continued horticultural experiments for another 12 years on 57 species and described his results in *The Effects of Cross and Self Fertilisation in the Vegetable Kingdom* (1876). Here he developed the theme that there are hereditary advantages in having two sexes in both the plant and animal kingdoms—to ensure cross-fertilization, which, as he knew from experiments, produced healthier, more vigorous offspring.

*The Effects of Cross and Self Fertilisation in the Vegetable Kingdom*

In "On the Movements and Habits of Climbing Plants" (1875) Darwin advanced an adaptive explanation for the tendency of the stems of certain plants to spiral toward heat and light, bending either clockwise or counterclockwise. Through experimentation he had discovered that a twining plant would not twine around an object larger than six inches in diameter. This characteristic Darwin interpreted as preventing a vine from climbing up a large tree where the shade from the upper branches would deprive it of sunlight. Still interested in the mechanism that enables some plants to climb and bend, Darwin continued experimenting and pinpointed "some matter in the upper part which is acted upon by light, and which transmits its effects to the lower part." He reported these researches in *The Power of Movement in Plants* (1880).

A chance observation of flies caught on the leaf of the common sundew initiated Darwin's investigation of carnivorous plants. He was especially impressed by the fact that the living cells of plants possess a similar capacity for irritability and response as the cells of animals. Darwin published these findings in *Insectivorous Plants* (1875). In his last botanical work, *The Formation of Vegetable Mould, Through the Action of Worms,* which appeared only six months before his death in 1881, he demonstrated the service that worms perform in digesting leaves and recirculating organic matter. It was a pioneering study in the field of quantitative ecology.

**A man of his time.** Darwin worked alone at home, leading the life of an independent scientist (a privileged existence open to a fortunate few in Victorian England). Money from Robert Darwin made it unnecessary for Charles to seek employment. After his return from the voyage Darwin knew he would never become a clergyman like his mentor, Henslow. Nor would he remain a bachelor like his brother, Erasmus, who was a man-about-town. After drawing up lists of the benefits and drawbacks of marriage, he proposed to his first cousin Emma Wedgwood, whom he married on Jan. 29, 1839. She brought fortune, devotion, and considerable housewifely skills that enabled him to work in peace for the next 40 years. Newly married, the Darwins moved into a house on Gower Street in London, but within a few years Darwin's increasingly poor health prompted them to move to the country. In 1842 the Darwins moved into Down House in the village of Downe, Kent, only 16 miles from London but remote from easy access to the city.

Charles and Emma Darwin had 10 children: two died in infancy and a third, Anne, died at age 10. The surviving five sons went away to school. George, Francis, and Horace became distinguished scientists, and Leonard, a major in the royal army, was an engineer and eugenicist. William Erasmus was undistinguished, as were his sisters, who prepared at home to follow their mother into marriage. Henrietta married; Elizabeth remained single at Down. Darwin was devoted to his wife and daughters but treated them as children, obliging Emma to ask him for the only key to the drawers containing all the keys to cupboards and other locked depositories.

Darwin noted in *The Descent* that the young of both sexes resemble the adult female in most species and reasoned that males are more evolutionarily advanced than females. His attitude toward women coloured his scientific insights. "The female is less eager than the male," he wrote, "She is coy," and when she takes part in choosing a mate, she chooses "not the male which is most attractive to her, but the one which is least distasteful."

Attitude toward women

His medical school experience had left him sympathetic to the popular antivivisectionist movement, but he admonished women for their involvement in it in a letter to *The Times* of London on June 23, 1876: "Women, who from the tenderness of their hearts and from their profound ignorance are the most vehement opponents of such experiments, will I hope pause when they learn that a few such experiments performed under the influence of an anaesthetic, have saved and will save through all future time thousands of women from a dreadful and lingering death." All expressions of cruelty offended him, and he was an ardent opponent of slavery.

Comfortable in English society, Darwin treasured his place and feared alienating those who he knew would be offended by his theory. He benefited at the beginning of his career from the scientific fraternity in London, who helped him understand the specimens from the *Beagle,* and he appreciated his intellectual give and take with Henslow, Hooker, Lyell, and Huxley. He was a beneficiary of this conservative English society, and his fear of ostracism was one of the forces that prevented him from publishing his theory sooner. He also dreaded the hurt he knew that his ideas would inflict on his close friend Henslow and especially on Emma, both devout Christians, for whom his theory was heresy.

The conflict between his science and his realization of what publication would imply for the society he was so much a part of manifested itself in physical pain. The once adventurous young naturalist was a semi-invalid before his 40th year. Darwin's illness has been the subject of extensive speculation. Some of the symptoms—painful flatulence, vomiting, insomnia, palpitations—appeared in force as soon as he began his first transmutation notebook, in 1837. Although he was exposed to insects in South America and could possibly have caught Chagas' or some other tropical disease, a careful analysis of the attacks in the context of his activities points to psychogenic origins. Throughout the next decades Darwin's maladies waxed and waned. But during the last decade of his life, when he concentrated on botanical research and no longer speculated about evolution, he experienced the best health since his years at Cambridge.

Darwin's illness

Darwin made his home at Down into his laboratory, where he experimented in his garden and observed the local fauna. Dogs and cats were part of the Darwin household, which also was not without a child under school age between 1839 and 1856. By no means a recluse, Darwin often attended scientific meetings in London; he

was away from home for about 2,000 days between 1842 and 1881. He was a member of 57 leading foreign learned societies and was no less a prominent figure in the village of Downe, treasurer of the Friendly Club and a Justice of the Peace. Darwin sent his children to village dances, and, even though he was a skeptical agnostic, he participated in church functions that were part of village life.

Darwin died at Down House on April 19, 1882. Within hours the news reached London, and a Parliamentary petition won him burial in Westminster Abbey. By this time the theory of evolution through natural selection was generally accepted. His ideas were modified by later developments in genetics and molecular biology, but his work remains central to modern evolutionary theory.

**MAJOR WORKS**

*Journal of Researches into the Geology and Natural History of the Various Countries Visited by H.M.S. Beagle* . . . (1839); *The Structure and Distribution of Coral Reefs* (1842); *On the Origin of Species by Means of Natural Selection, or, The Preservation of Favoured Races in the Struggle for Life* (1859); *On the Various Contrivances by Which British and Foreign Orchids Are Fertilised by Insects, and On the Good Effects of Intercrossing* (1862); "On the Movements and Habits of Climbing Plants" (1865); *The Variation of Animals and Plants Under Domestication* (1868); *The Descent of Man, and Selection in Relation to Sex,* 2 vol. (1871); *The Expression of the Emotions in Man and Animals* (1872); *Insectivorous Plants* (1875); *The Effects of Cross and Self Fertilisation in the Vegetable Kingdom* (1876); *The Different Forms of Flowers on Plants of the Same Species* (1877); *The Power of Movement in Plants* (1880); *The Formation of Vegetable Mould, Through the Action of Worms, with Observations on Their Habits* (1881).

**BIBLIOGRAPHY.** FRANCIS DARWIN (ed.), *The Life and Letters of Charles Darwin, Including an Autobiographical Chapter,* 3 vol. (1887, reissued in 2 vol., 1972), is a standard life, by Darwin's son, but was written at a time when the full significance of evolution by natural selection was not understood. It is supplemented and continued by FRANCIS DARWIN and A.C. SEWARD (eds.), *More Letters of Charles Darwin,* 2 vol. (1903, reprinted 1972). FREDERICK BURKHARDT and SYDNEY SMITH (eds.), *The Correspondence of Charles Darwin* (1985– ), provides a vivid portrait of the man and his thoughts on evolution in many previously unpublished letters; 2 vol., covering the years 1821–43, had appeared by 1987.

PAUL H. BARRETT (ed.), *The Collected Papers of Charles Darwin,* 2 vol. (1977, reprinted, 2 vol. in 1, 1980), is a collection of his minor works. Darwin's earliest exposition of his discoveries is contained in CHARLES DARWIN, *The Foundations of "The Origin of Species": Two Essays Written in 1842 and 1844* (1909), edited by FRANCIS DARWIN, and reedited by GAVIN DE BEER in *Evolution by Natural Selection* (1958, reprinted 1971). Several of Darwin's notebooks have been edited and published; four are presented in GAVIN DE BEER (ed.), "Darwin's Notebooks on Transmutation of Species," *Bulletin of the British Museum (Natural History), Historical Series,* vol. 2, no. 2–5 (1959–63). See also SANDRA HERBERT (ed.), *The Red Notebook of Charles Darwin* (1980). Darwin's original manuscript on natural selection, of which *On the Origin of Species* was but an abstract, is presented in R.C. STAUFFER (ed.), *Charles Darwin's Natural Selection* (1975).

Biographical works include NORA BARLOW (ed.), *The Autobiography of Charles Darwin, 1809–1882* (1958); GERTRUDE HIMMELFARB, *Darwin and the Darwinian Revolution* (1959, reissued 1968); GAVIN DE BEER (ed.), *Charles Darwin: Evolution by Natural Selection* (1963, reprinted 1976), also published with title *Charles Darwin: A Scientific Biography* (1963, reprinted 1965), a work that takes into account Darwin's notebooks on the transmutation of species and Mendel's attitude toward Darwin; and PETER BRENT, *Charles Darwin, "A Man of Enlarged Curiosity"* (1981, reissued 1983), an extensive treatment. An examination of Darwin's botanical studies is MEA ALLAN, *Darwin and His Flowers: The Key to Natural Selection* (1977). WILMA GEORGE, *Darwin* (1982), discusses Darwin's biology. See also RALPH COLP, JR., *To Be an Invalid: The Illness of Charles Darwin* (1977).

Studies of Darwin and his contemporaries include DAVID KOHN (ed.), *The Darwinian Heritage* (1985), an important collection of essays, including revised interpretations of Darwin's position among his contemporary scientists; EDWARD MANIER, *The Young Darwin and His Cultural Circle: A Study of Influences Which Helped Shape the Language and Logic of the First Drafts of the Theory of Natural Selection* (1978), an intellectual history not only of Darwin but also of early Victorian society; and MICHAEL RUSE, *The Darwinian Revolution* (1979, reissued 1981), a study of the reactions of the British scientific community between 1830 and 1875 to the idea of evolution. NEAL C. GILLESPIE, *Charles Darwin and the Problem of Creation* (1979), is a study of how Darwin and his contemporaries dealt with the idea of special creation.

Important reinterpretations of Darwin's work can be found in DOV OSPOVAT, *The Development of Darwin's Theory: Natural History, Natural Theology, and Natural Selection, 1838–1859* (1981), a revisionist suggestion that Darwin's theory of natural selection underwent a radical change between the late 1830s and the publication of *On the Origin of Species*; and four articles in *Journal of the History of Biology*: SANDRA HERBERT, "The Place of Man in the Development of Darwin's Theory of Transmutation," parts 1 and 2, 7(2):249–257 (Fall 1974) and 10(2):155–227 (Fall 1977), an important reinterpretation of Darwin's development; and a revised account of Darwin's interpretation of the Galápagos animals, FRANK SULLAWAY, "Darwin and His Finches: The Evolution of a Legend," 15(1):1–53 (Spring 1982), and "Darwin's Conversion: The Beagle Voyage and Its Aftermath," 15(3):325–396 (Fall 1982), convincing arguments that Darwin came to his conclusions after his return from the voyage.

(B.Ke.)

# Death

During the latter half of the 20th century, death has become a strangely popular subject. Before that time, perhaps rather surprisingly, it was a theme largely eschewed in serious scientific, and to a lesser extent, philosophical speculations. It was neglected in biological research and, being beyond the physician's ministrations, was deemed largely irrelevant by medical practice. In modern times, however, the study of death has become a central concern in all these disciplines and in many others.

"So many more people seem to die nowadays," an elderly lady is alleged to have said, scanning the obituary columns of a famous daily. This was not just a comment on the documented passing of a cohort. Various journals now not only list the dead but also describe what they died of, at times in some detail. They openly discuss subjects considered too delicate or personal less than a generation ago. Television interviewers question relatives of the dying—or even the dying themselves—and films depict murders or executions in gruesome and often quite accurate detail. Death is no longer enshrined in taboos. Popular readiness to approach these matters and a general desire to be better informed about them reflect a change in cultural attitudes perhaps as great as that which accompanied the more open discussion of sex after World War I.

Thanatology—the study of death—delves into matters as diverse as the cultural anthropology of the notion of soul, the burial rites and practices of early civilizations, the location of cemeteries in the Middle Ages, and the conceptual difficulties involved in defining death in an individual whose brain is irreversibly dead but whose respiration and heartbeat are kept going by artificial means. It encompasses the biological study of programmed cell death, the understanding care of the dying, and the creation of an informed public opinion as to how the law should cope with the stream of problems generated by intensive-care technology. Legal and medical quandaries regarding the definition of death and the rights of the terminally ill (or their families) to refuse life-prolonging treatments force physicians to think like lawyers, lawyers like physicians, and both like philosophers. In his *Historia Naturalis* (*Natural History*), the Roman author Pliny the Elder wrote that "so uncertain is men's judgment that they cannot determine even death itself." The challenge remains, but if humans now fail to provide some answers it will not be for lack of trying.

This article is divided into the following sections:


The meaning of death 982
The biological problems 983
  Death: process or event 983
  The "point of no return" 983
Cell death 984
Clinical death 985
  Functions of the brain stem 985
  Mechanisms of brain-stem death 985
  Evolution of the concept of brain-stem death 985
  Diagnosis of brain-stem death 986
The cultural background 986
  Ancient Egypt 987
  Mesopotamia 987
  Judaism 988
  Hinduism 989
  Islām 991
  The modern Western context 992
Bibliography 994


## THE MEANING OF DEATH

This subject can be approached from a variety of perspectives. It can, for example, be viewed historically, in terms of how popular perceptions of death have been reflected in poetry, literature, legend, or pictorial art. Illustrations of those killed in battle and of their severed parts find particular prominence in ancient Egyptian art. The campaign of the 13th-century-BC Egyptian king Ramses II against the Hittites, in particular the Battle of Kadesh, is recorded in gruesome detail on the battle reliefs of 19th- and 20th-dynasty temples in Upper Egypt. Assyrian art, too, made great play of illustrating cadavers. Those slaughtered by the king Ashurbanipal (flourished 7th century BC) in his campaign against the Arabian king Uate are shown having their eyes plucked out by vultures. These very concrete depictions of the meaning of death seem to have had mainly propagandistic value, boosting the self-confidence of the victors and inspiring fear among the defeated. Deities of the dead were features of many early cultures, but apart from ancient Egypt neither such deities nor those over whom they held dominion were the subject of any significant artistic representation. In Egypt, sepulchral iconography was to reach truly impressive heights, particularly after the democratization of the Osirian cult with its promise of an afterlife for everyone. Well-known sculptors produced some striking individual tombstones in both ancient Greece and Rome, but it was medieval Christianity that gave real impetus to this practice, which can be thought of as an attempt to perpetuate among the living a vivid memory of the dead. The representation of death itself, usually personified in the form of a skeleton, seems to have developed on a large scale only in medieval Christian art.

Artistic representations of the dead

An alternative approach is to look at the meaning of death in terms of various eschatologies (beliefs regarding death and the end of the world). Human beings have been the only species to bury their dead in a systematic way, often with implements to be used in a further existence. The study of death rites and customs illustrates impressively the relation between religious belief and popular practice in the presence of the dead. Such an approach starts from the meaning of death in those cultures (such as Phoenician, early Judaic, Homeric, Epicurean, and Stoic) in which only a shadowy afterlife or no afterlife at all was envisaged; it analyzes other traditions (such as Sumero-Akkadian) in which ambiguities and contradictions abounded; and it finally searches for death's meaning in those cultures (such as ancient Egyptian, Zoroastrian, Hindu, Orphic, Platonic, Christian, Pharisaic Judaic, and Islāmic) in which a very "physical" afterlife, or the presence of an eternal soul, played central roles.

Both the historical and the eschatological approaches share a common advantage: they need not be preceded by a definition of death. They accept death as an easily determined empirical fact, not requiring discussion or further elaboration. But a conceptual crisis has arisen in modern medicine and biology, a crisis that stems precisely from the realization that the definition of death—taken for granted for millennia—requires reexamination. To approach the subject of death from the biological angle, which is perhaps the most difficult and arguably the most challenging perspective, certainly reflects some of the most pressing needs of modern times.

The difficulty of defining death

Many dictionaries define death as "the extinction or cessation of life" or as "ceasing to be." As life itself is notoriously difficult to define—and as everyone tends to think of things in terms of what is known—the problems in defining death are immediately apparent. The most useful definitions of life are those that stress function, whether at the level of physiology, of molecular biology

and biochemistry, or of genetic potential. Death should be thought of as the irreversible loss of such functions.

The remainder of this article first explores the recurrent problems involved in seeking a biological definition of death. It then examines the implications of these problems in relation to human death. In this context, the article raises two major points: (1) death of the brain is the necessary and sufficient condition for death of the individual; and (2) the physiological core of brain death is the death of the brain stem. Finally, the article surveys notions about the meaning of human death that have prevailed throughout history in a wide variety of cultural contexts. By so doing, it attempts to show that brain-stem death, far from being a radically new idea, turns out to have always provided both an ultimate mechanism of death and a satisfactory anatomical basis for a wide range of philosophical concepts relating to death.

## THE BIOLOGICAL PROBLEMS

Whether one considers the death of individual cells, the death of small multicellular organisms, or the death of a human being, certain problems are repeatedly met. The physicist may encounter difficulties in trying to define death in terms of entropy change and the second law of thermodynamics. So may the histologist looking at the ultrastructure of dying tissue through an electron microscope. Pope Pius XII, speaking to an International Congress of Anesthesiologists in 1957, raised the question of when, in the intensive care unit, the soul actually left the body. More secularly inclined philosophers have meanwhile pondered what it was that was so essential to the nature of man that its loss should be called death. The questions of what may or may not be legitimately demanded of a "beating-heart cadaver" (in terms of supplying donor organs for transplants or of serving as a subject for physiological experimentation) has given new poignancy to the quip made by the English author Sir Thomas Browne in 1643: "With what strife and pains we come into the world we know not, but 'tis commonly no easy matter to get out of it." Common conceptual difficulties underlie many of these questions.

**Death: process or event.** The American physician and writer Oliver Wendell Holmes said "to live is to function" and "that is all there is in living." But who or what is the subject who lives because it functions? Is death the irreversible loss of function of the whole organism (or cell); that is, of every one of its component parts? Or is it the irreversible loss of function of the organism (or cell) as a whole; that is, as a meaningful and independent biological unit? To perceive the difference between the two questions is to understand many modern controversies about death. The described dichotomy is clearly part of a much wider one: civilizations fall apart yet their component societies live on; societies disintegrate but their citizens survive; individuals die while their cells, perversely, still metabolize; finally, cells can be disrupted yet the enzymes they release may, for a while, remain active.

Such problems would not arise if nature were tidier. In nearly all circumstances human death is a process rather than an event. Unless caught up in nuclear explosions people do not die suddenly, like the bursting of a bubble.

Death as a process

A quiet, "classical" death provides perhaps the best illustration of death as a process. Several minutes after the heart has stopped beating, a mini-electrocardiogram may be recorded, if one probes for signals from within the cardiac cavity. Three hours later, the pupils still respond to pilocarpine drops by contracting, and muscles repeatedly tapped may still mechanically shorten. A viable skin graft may be obtained from the deceased 24 hours after the heart has stopped, a viable bone graft 48 hours later, and a viable arterial graft as late as 72 hours after the onset of irreversible asystole (cardiac stoppage). Cells clearly differ widely in their ability to withstand the deprivation of oxygen supply that follows arrest of the circulation.

Similar problems arise, but on a vastly larger scale, when the brain is dead but the heart (and other organs) are kept going artificially. Under such circumstances, it can be argued, the organism as a whole may be deemed dead, although the majority of its cells are still alive.

**The "point of no return."** To claim that death is a process does not imply that this process unfurls at an even rate, or that within it there are not "points of no return." The challenge is to identify such points with greater precision for various biological systems. At the clinical level, the irreversible cessation of circulation has for centuries been considered a point of no return. It has provided (and still provides) a practical and valid criterion of irreversible loss of function of the organism as a whole. What is new is the dawning awareness that circulatory arrest is a mechanism of death and not in itself a philosophical concept of death; that cessation of the heartbeat is only lethal if it lasts long enough to cause critical centres in the brain stem to die; and that this is so because the brain stem is irreplaceable in a way the cardiac pump is not. These are not so much new facts as new ways of looking at old ones.

Failure to establish beyond all doubt that the point of no return had been reached has, throughout the ages, had interesting effects on medical practice. The Thracians, according to the ancient Greek historian Herodotus, kept their dead for three days before burial. The Romans kept the corpse considerably longer; the Roman author Servius, in his commentary on Virgil, records that "on the eighth day they burned the body and on the ninth put its ashes in the grave." The practice of cutting off a finger, to see whether the stump bled, was often resorted to. Even the most eminent proved liable to diagnostic error. The 16th-century Flemish physician Andreas Vesalius, probably the greatest anatomist of all time, professor of surgery in Padua for three years and later physician to the Holy Roman emperor Charles V, had to leave Spain in a hurry in 1564. He was performing a postmortem when the subject, a nobleman he had been attending, showed signs of life. This was at the height of the Spanish Inquisition and Vesalius was pardoned only on the condition that he undertake a pilgrimage to the Holy Sepulchre in Jerusalem.

Anxieties over premature burial

Fears of being buried alive have long haunted humankind. During the 19th century, for example, accounts of "live sepulture" appeared in medical writing and led to repeated demands that putrefaction—the only sure sign of death of the whole organism—be considered an essential prerequisite to a diagnosis of death. Anxieties had become so widespread following the publication of some of U.S. author Edgar Allan Poe's macabre short stories that Count Karnice-Karnicke, a Russian nobleman, patented a coffin of particular type. If the "corpse" regained consciousness after burial, it could summon help from the surface by activating a system of flags and bells. Advertisements described the price of the apparatus as "exceedingly reasonable, only about twelve shillings."

At the turn of the century, a sensation-mongering press alleged that there were "many ugly secrets locked up underground." There may have been some basis for these claims: instances of collapse and apparent death were not uncommon during epidemics of plague, cholera, and smallpox; hospitals and mortuaries were overcrowded, and there was great fear of the spread of infection. This agitation resulted in stricter rules concerning death certification. In the United Kingdom, statutory obligations to register deaths date only from 1874, and at that time it was not even necessary for a doctor to have viewed the corpse.

The impact of medical technology

The second half of the 20th century has seen tremendous developments in the field of intensive care and the emergence of new controversies concerning the point of no return. Modern technology now makes it possible to maintain ventilation (by respirators), cardiac function (by various pumping devices), feeding (by the intravenous route), and the elimination of the waste products of metabolism (by dialysis) in a body whose brain is irreversibly dead. In these macabre by-products of modern technology, a dissociation has taken place between the various components of death so that the most important—the death of the brain—occurs before, rather than after, the cessation of other functions, such as circulation. Such cases have presented both practical and conceptual problems, but the latter need not have arisen had what happens during decapitation been better appreciated.

"Beating-heart cadavers" were of course familiar to the observant long before the days of intensive care units. A

photograph of a public decapitation in a Bangkok square in the mid-1930s illustrates such a case. The victim is tied to a stake and the head has been severed, but jets of blood from the carotid and vertebral arteries in the neck show that the heart is still beating. It is doubtful that anyone would describe the executed man—as distinct from some of his organs—as still alive. This gruesome example stresses three points: it reiterates the fact, admittedly from an unusual angle, that death is a process rather than an event; it emphasizes the fact that in this process there is a point of no return; and it graphically illustrates the difference between the death of the organism as a whole and the death of the whole organism. In thinking the implications through, one takes the first steps toward understanding brain death. The executed man has undergone anatomical decapitation. Brain death is physiological decapitation: it arises when intracranial pressure exceeds arterial pressure, thereby depriving the brain of its blood supply as efficiently as if the head had been cut off. The example serves as an introduction to the proposition that the death of the brain is the necessary and sufficient condition for the death of the individual.

These issues were authoritatively discussed in 1968, at the 22nd World Medical Assembly in Sydney, Australia. The assembly stated that "clinical interest lies not in the state of preservation of isolated cells but in the fate of a person. The point of death of the different cells and organs is not as important as the certainty that the process has become irreversible." The statement had a profound effect on modern medical thinking. "Irreversible loss of function of the organism as a whole" became an accepted clinical criterion of death.

Semantic confusion may underlie some of the controversies outlined in this section. In many languages, including English, the word *death* may be used in various ways. The *Concise Oxford Dictionary* for instance defines death both as "dying" (a process) and as "being dead" (a state). Expressions such as "a painful death" and "a lingering death" show how often the word is used in the former sense. Many people are afraid of dying yet can face the prospect of being dead with equanimity. Another source of confusion that bedevils discussions about death is what the great English mathematician and philosopher Alfred North Whitehead called the "fallacy of misplaced concreteness." This occurs when one treats an abstraction (however useful it may be to denote the behaviour or properties of objects under specific circumstances) as if it were itself a material thing. "O death, where is thy sting?" may be a searching metaphorical question, but such queries can only confuse the biologist. When the poet John Milton wrote of "the pain of death denounced, whatever *thing* death be," the conceptual problem was of his own making.

The next two sections of this article illustrate these general principles concerning death from each end of the spectrum of living things: from the level of the cell and from that of the fully developed human being.

## CELL DEATH

A vast amount of work has been devoted since the late 19th century to discovering how cells multiply. The study of how and why they die is a relatively recent concern: a rubric entitled "cell death" only appeared in the *Index Medicus,* an index to medical literature, in 1979.

Cell necrosis

What most textbooks of pathology describe as cell death is coagulative necrosis. This is an abnormal morphological appearance, detected in tissue examined under the microscope. The changes, which affect aggregates of adjacent cells or functionally related cohorts of cells, are seen in a variety of contexts produced by accident, injury, or disease. Among the environmental perturbations that may cause cell necrosis are oxygen deprivation (anoxia), hyperthermia, immunological attack, and exposure to various toxins that inhibit crucial intracellular metabolic processes. Coagulative necrosis is the classical form of cell change seen when tissues autolyze (digest themselves) in vitro.

But cells may die by design as well as by accident. Research in developmental pathology has stressed the biological importance of this other kind of cell death, which has been referred to as programmed cell death. In vertebrates it has been called apoptosis and in invertebrates, cell deletion. Programmed cell death plays an important role in vertebrate ontogeny (embryological development) and teratogenesis (the production of malformations), as well as in the spectacular metamorphoses that affect tadpoles or caterpillars. Such programmed events are essential if the organism as a whole is to develop its normal final form. Waves of genetically driven cell death are critical to the proper modeling of organs and systems. The inflections (curvatures) of the developing mammalian brain and spinal cord, for instance, or the achievement of a proper numerical balance between functionally related cell groups, cannot be understood without an appreciation of how the death of some (or many) cells is necessary for others to reach maturity. Localized cell death, occurring at precise moments during normal ontogeny, explains phenomena as varied as the fashioning of the digits or the involution of phylogenetic vestiges. Several congenital abnormalities can be attributed to disorders of programmed cell death. Cell death occurs spontaneously in normally involuting tissues such as the thymus. It can be initiated or inhibited by a variety of environmental stimuli, both physiological and pathological. Cell death even occurs in some of the cells of untreated malignant tumours, and it is seen during tumour regression induced by X rays or radiomimetic cytotoxic agents. Programmed cell death may also play a part in the process of aging, cells being designed to die after a certain number of mitotic divisions. Groups of cells responsible for the colour of human hair, for instance, may cease to function years before the hair itself loses the capacity to grow: the result is the "uncoloured" white hair of old age.

Programmed cell death

The two types of cell death—imposed from without or programmed from within—have different morphological features. Furthermore, different intracellular mechanisms have been incriminated in their production.

Necrosis is characterized by early swelling of the cytoplasm and of the mitochondria (energy-releasing organelles) within it. Later changes include the appearance of localized densities, possibly related to calcium deposition, in the matrix (ground substance) of the mitochondria. This is followed by the dissolution of other cytoplasmic organelles and the separation of affected cells from their neighbours through shearing of intercellular junctions. Nuclear alterations occur late and are relatively unremarkable. The nucleus swells, becomes darker (pyknosis), and ruptures (karyolysis) at about the same time as does the plasma membrane, the outer envelope of the cell. The basic mechanism of necrosis is thought to be a loss of control over cell volume, related to changes in the permeability of the cell membrane. These changes form the basis of several of the tests used to diagnose a necrotic cell in the laboratory. The affected membrane rapidly loses its ion-pumping capacity, and there are dramatic increases in the intracellular concentrations of sodium and calcium ions. This is followed by osmotic shock and the development of intracellular acidosis. The early injury to the mitochondria has profound repercussions on intracellular oxidative metabolism. The point of no return is reached with irreversible damage to mitochondrial structure and function. Later still, the lysosomes (membranous bags of hydrolytic enzymes found in most cells) rupture, releasing their acid enzymes into the cytoplasm of the cell. All this produces an ionic milieu unsuitable to the survival of the nucleus. Loss of the cell's capacity to synthesize protein is the ultimate proof that it is functionally dead.

Mechanisms of necrosis

Programmed cell death usually affects scattered single cells. Early ultrastructural features are the disintegration of cell junctions and condensations of the cytoplasm. The cells shrivel up instead of swelling. Lumps of chromatin aggregate at the surface of the nucleus. The nuclear membrane develops folds, and the nucleus splits into a number of membrane-bound, ultrastructurally well-preserved fragments, which are shed and promptly taken up by specialized scavenger cells or even by ordinary cells in the neighbourhood. Energy-producing mitochondria are preserved until quite late. The nuclear changes seem to be energy-dependent; they may reflect the fact that genes in the

nucleus are beginning to express themselves in new ways, in response to unknown stimuli. One of these responses seems to be the activation of endogenous endonucleases, enzymes in the cell nucleus that "suicidally" disrupt its cardinal functions.

Time alone will tell whether the distinctions between the two types of cell death are valid or spurious, and whether the concept of apoptosis will gain wide acceptance. Reality will probably turn out to be a great deal more complex. Meanwhile, one should retain, without overemphasis, the twin visions of cell death—one in which death approaches the cell from the outside and the other in which death starts from within the living core of the cell itself.

## CLINICAL DEATH

At the opposite end of the spectrum from cell death lies the death of a human being. It is obvious that the problems of defining human death cannot be resolved in purely biological terms, divorced from all ethical or cultural considerations. This is because there will be repercussions (burial, mourning, inheritance, etc.) from any decisions made, and because the decisions themselves will have to be socially acceptable in a way that does not apply to the fate of cells in tissue culture.

Unless death is defined at least in outline, the decision that a person is "dead" cannot be verified by any amount of scientific investigation. Technical data can never answer purely conceptual questions. Earlier in this article it was suggested that the death of the brain was the necessary and sufficient condition for the death of the individual, but the word *death* was not given much content beyond the very general definition of "irreversible loss of function." If one seeks to marry conceptions of death prevalent in the oldest cultures with the most up-to-date observations from intensive care units, one might think of human death as the irreversible loss of the capacity for consciousness combined with the irreversible loss of the capacity to breathe. The anatomical basis for such a concept of human death resides in the loss of brain-stem function.

Human death defined

**Functions of the brain stem.** The brain stem is the area at the base of the brain that includes the mesencephalon (midbrain), the pons, and the medulla. It contains the respiratory and vasomotor centres, which are responsible, respectively, for breathing and the maintenance of blood pressure. Most importantly, it also contains the ascending reticular activating system, which plays a crucial role in maintaining alertness (*i.e.,* in generating the capacity for consciousness); small, strategically situated lesions in the medial tegmental portions of the midbrain and rostral pons cause permanent coma. All of the motor outputs from the cerebral hemispheres—for example, those that mediate movement or speech—are routed through the brain stem, as are the sympathetic and parasympathetic efferent nerve fibres responsible for the integrated functioning of the organism as a whole. Most sensory inputs also travel through the brain stem. This part of the brain is, in fact, so tightly packed with important structures that small lesions there often have devastating effects. By testing various brain-stem reflexes, moreover, the functions of the brain stem can be assessed clinically with an ease, thoroughness, and degree of detail not possible for any other part of the central nervous system.

Capacity for consciousness

It must be stressed that the capacity for consciousness (an upper brain-stem function) is not the same as the content of consciousness (a function of the cerebral hemispheres); it is, rather, an essential precondition of the latter. If there is no functioning brain stem, there can be no meaningful or integrated activity of the cerebral hemispheres, no cognitive or affective life, no thoughts or feelings, no social interaction with the environment, nothing that might legitimize adding the adjective *sapiens* ("wise") to the noun *Homo* ("man"). The "capacity for consciousness" is perhaps the nearest one can get to giving a biological flavour to the notion of "soul."

The capacity to breathe is also a brain-stem function, and apnea (respiratory paralysis) is a crucial manifestation of a nonfunctioning lower brain stem. Alone, of course, it does not imply death; patients with bulbar poliomyelitis, who may have apnea of brain-stem origin, are clearly not dead. Although irreversible apnea has no strictly philosophical dimension, it is useful to include it in any concept of death. This is because of its obvious relation to cardiac function—if spontaneous breathing is lost the heart cannot long continue to function—and perhaps because of its cultural associations with the "breath of life." These aspects are addressed in the later discussion of how death has been envisaged in various cultures.

**Mechanisms of brain-stem death.** From as far back as medical records have been kept, it has been known that patients with severe head injuries or massive intracranial hemorrhage often die as a result of apnea: breathing stops before the heart does. In such cases, the pressure in the main (supratentorial) compartment of the skull becomes so great that brain tissue herniates through the tentorial opening, a bony and fibrous ring in the membrane that separates the spaces containing the cerebral hemispheres and the cerebellum. The brain stem runs through this opening, and a pressure cone formed by the herniated brain tissue may dislocate the brain stem downward and cause irreversible damage by squeezing it from each side. An early manifestation of such an event is a disturbance of consciousness; a late feature is permanent apnea. This was previously nature's way out.

Pressure cone formation

With the widespread development of intensive care facilities in the 1950s and '60s, more and more such moribund patients were rushed to specialized units and put on ventilators just before spontaneous breathing ceased. In some cases the effect was dramatic. When a blood clot could be evacuated, the primary brain damage and the pressure cone it had caused might prove reversible. Spontaneous breathing would return. In many cases, however, the massive, structural intracranial pathology was irremediable. The ventilator, which had taken over the functions of the paralyzed respiratory centre, enabled oxygenated blood to be delivered to the heart, which went on beating. Physicians were caught up in a therapeutic dilemma partly of their own making: the heart was pumping blood to a dead brain. Sometimes the intracranial pressure was so high that the blood could not even enter the head. Modern technology was exacting a very high price: the beating-heart cadaver.

Brain-stem death may also arise as an intracranial consequence of extracranial events. The main cause in such cases is circulatory arrest. The usual context is delayed or inadequate cardiopulmonary resuscitation following a heart attack. The intracranial repercussions depend on the duration and severity of impaired blood flow to the head. In the 1930s the British physiologist John Scott Haldane had emphasized that oxygen deprivation "not only stopped the machine, but wrecked the machinery." Circulatory arrest lasting two or three minutes can cause widespread and irreversible damage to the cerebral hemispheres while sparing the brain stem, which is more resistant to anoxia. Such patients remain in a "persistent vegetative state." They breathe and swallow spontaneously, grimace in response to pain, and are clinically and electrophysiologically awake, but they show no behavioral evidence of awareness. Their eyes are episodically open (so that the term *coma* is inappropriate to describe them), but their retained capacity for consciousness is not endowed with any content. Some patients have remained like this for many years. Such patients are not dead, and their prognosis depends in large part on the quality of the care they receive. The discussion of their management occasionally abuts onto controversies about euthanasia and the "right to die." These issues are quite different from that of the "determination of death," and failure to distinguish these matters has been the source of great confusion.

Circulatory arrest

If circulatory arrest lasts for more than a few minutes, the brain stem—including its respiratory centre—will be as severely damaged as the cerebral hemispheres. Both the capacity for consciousness and the capacity to breathe will be irreversibly lost. The individual will then show all the clinical features of a dead brain, even if the heart can be restarted.

**Evolution of the concept of brain-stem death.** It was against this sort of background that French neurologists, in 1958, described a condition they called *coma dépassé*

(literally, "a state beyond coma"). Their patients all had primary, irremediable, structural brain lesions; were deeply comatose; and were incapable of spontaneous breathing. They had not only lost their ability to react to the external world, but they also could no longer control their own internal environment. They became poikilothermic (*i.e.,* they could not control their body temperature, which varied with that of the environment). They could not control their blood pressure or vary their heart rate in response to appropriate stimuli. They could not even retain body water and would pass great volumes of urine. The organism as a whole had clearly ceased to function. *Coma dépassé* was considered a "frontier state" between life and death. Ventilation was continued in the vast majority of such cases until the heartbeat ceased, usually a few days later.

Criteria for recognizing brain death

In 1968 the Ad Hoc Committee of the Harvard Medical School published a report entitled "A Definition of Irreversible Coma" in *The Journal of the American Medical Association.* This watershed article listed criteria for the recognition of the "brain-death syndrome." It stated that the persistence of a state of apneic coma with no evidence of brain-stem and spinal reflexes and a flat electroencephalogram over a period of 24 hours implied brain death, provided the cause of the coma was known and provided reversible causes of brain dysfunction (such as hypothermia or drug intoxication) had been excluded. The report explicitly identified brain death with death (without seeking to define death) and endorsed the withdrawal of respiratory support in such cases. No evidence was published to legitimize the contention that the coma was irreversible; *i.e.,* that if artificial ventilation was continued no such patient ever recovered consciousness, and that all invariably developed asystole. There was wide medical experience among the members of the committee, however, and its contentions have since been massively validated. Not a single exception has come to light.

The next few years witnessed increasing sophistication in the techniques used to diagnose brain death, none of which, however, surpassed basic clinical assessment. In 1973 two neurosurgeons in Minneapolis, Minn., identified the death of the brain stem as the point of no return in the diagnosis of brain death. In 1976 and 1979, the Conference of Royal Colleges and Faculties of the United Kingdom published important memoranda on the subject. The first described the clinical features of a dead brain stem, the second identified brain-stem death with death. In 1981 in the United States, the President's Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research published a report ("Defining Death") and a list of guidelines very similar to the British ones. The commission also proposed a model statute, called the Uniform Determination of Death Act, which was subsequently endorsed by the American Medical Association, the American Bar Association, and the National Conference of Commissioners on Uniform State Laws and became law in many states. International opinion and practice has moved along similar lines in accepting the concept of brain-stem death.

**Diagnosis of brain-stem death.** The diagnosis is not technically difficult. In more and more countries, it is made on purely clinical grounds. The aim of the clinical tests is not to probe every neuron within the intracranial cavity to see if it is dead—an impossible task—but to establish irreversible loss of brain-stem function. This is the necessary and sufficient condition for irreversible unconsciousness and irreversible apnea, which together spell a dead patient. Experience has shown that instrumental procedures (such as electroencephalography and studies of cerebral blood flow) that seek to establish widespread loss of cortical function contribute nothing of relevance concerning the cardiac prognosis. Such tests yield answers of dubious reliability to what are widely felt to be the wrong questions. As the concept of brain-stem death is relatively new, most countries rightly insist that the relevant examinations be carried out by physicians of appropriate seniority. These doctors (usually neurologists, anesthetists, or specialists in intensive care) must be entirely separate from any who might be involved in using the patient's organs for subsequent transplants.

Stages in diagnosing brain-stem death

The diagnosis of brain-stem death involves three stages. First, the cause of the coma must be ascertained, and it must be established that the patient (who will always have been in apneic coma and on a ventilator for several hours) is suffering from irremediable, structural brain damage. Damage is judged "irremediable" based on its context, the passage of time, and the failure of all attempts to remedy it. Second, all possible causes of reversible brain-stem dysfunction, such as hypothermia, drug intoxication, or severe metabolic upset, must be excluded. Finally, the absence of all brain-stem reflexes must be demonstrated, and the fact that the patient cannot breathe, however strong the stimulus, must be confirmed.

It may take up to 48 hours to establish that the preconditions and exclusions have been met; the testing of brain-stem function takes less than half an hour. When testing the brain-stem reflexes, doctors check for the following normal responses: (1) constriction of the pupils in response to light, (2) blinking in response to stimulation of the cornea, (3) grimacing in response to firm pressure applied just above the eye socket, (4) movements of the eyes in response to the ears being flushed with ice water, and (5) coughing or gagging in response to a suction catheter being passed down the airway. All responses have to be absent on at least two occasions. Apnea, which also must be confirmed twice, is assessed by disconnecting the patient from the ventilator. (Prior to this test, the patient is fully oxygenated by being made to breathe 100 percent oxygen for several minutes, and diffusion oxygenation into the trachea is maintained throughout the procedure. These precautions ensure that the patient will not suffer serious oxygen deprivation while disconnected from the ventilator.) The purpose of this test is to establish the total absence of any inspiratory effort as the carbon dioxide concentration in the blood (the normal stimulus to breathing) reaches levels more than sufficient to drive any respiratory centre cells that may still be alive.

Tests of brain-stem reflexes

The patient thus passes through a tight double filter of preconditions and exclusions before he is even tested for the presence of a dead brain stem. This emphasis on strict preconditions and exclusions has been a major contribution to the subject of brain-stem death, and it has obviated the need for ancillary investigations. Thousands of patients who have met criteria of this kind have had ventilation maintained: all have developed asystole within a few hours or a few days, and none has ever regained consciousness. There have been no exceptions. The relevant tests for brain-stem death are carried out systematically and without haste. There is no pressure from the transplant team.

The developments in the idea and diagnosis of brain-stem death came as a response to a conceptual challenge. Intensive-care technology had saved many lives, but it had also created many brain-dead patients. To grasp the implications of this situation, society in general—and the medical profession in particular—was forced to rethink accepted notions about death itself. The emphasis had to shift from the most common mechanism of death (*i.e.,* irreversible cessation of the circulation) to the results that ensued when that mechanism came into operation: irreversible loss of the capacity for consciousness, combined with irreversible apnea. These results, which can also be produced by primary intracranial catastrophes, provide philosophically sound, ethically acceptable, and clinically applicable secular equivalents to the concepts of "departure of the soul" and "loss of the 'breath of life,' " which were so important to some earlier cultures.

## THE CULTURAL BACKGROUND

Throughout history, specific cultural contexts have always played a crucial role in how people perceived death. Different societies have held widely diverging views on the "breath of life" and on "how the soul left the body" at the time of death. Such ideas are worth reviewing (1) because of the light they throw on important residual elements of popular belief; (2) because they illustrate the distance traveled (or not traveled) between early beliefs and current ones; and (3) because of the relevance of certain old ideas to contemporary debates about brain-stem death and about the philosophical legitimacy of organ transplanta-

tion. The following discussion therefore focuses on how certain cultural ideas about death compare or contrast with the modern concept. For an overview of various eschatologies from a cross-cultural perspective, see RITES AND CEREMONIES, SACRED: *Death rites and customs.*

**Ancient Egypt.** Two ideas that prevailed in ancient Egypt came to exert great influence on the concept of death in other cultures. The first was the notion, epitomized in the Osirian myth, of a dying and rising saviour god who could confer on devotees the gift of immortality; this afterlife was first sought by the pharaohs and then by millions of ordinary people. The second was the concept of a postmortem judgment, in which the quality of the deceased's life would influence his ultimate fate. Egyptian society, it has been said, consisted of the dead, the gods, and the living. During all periods of their history, the ancient Egyptians seem to have spent much of their time thinking of death and making provisions for their afterlife. The vast size, awe-inspiring character, and the ubiquity of their funerary monuments bear testimony to this obsession.

The physical preservation of the body was central to all concerns about an afterlife; the Egyptians were a practical people, and the notion of a disembodied existence would have been totally unacceptable to them. The components of the person were viewed as many, subtle, and complex; moreover, they were thought to suffer different fates at the time of death. The physical body was a person's *khat,* a term that implied inherent decay. The *ka* was the individual's doppelgänger, or double; it was endowed with all the person's qualities and faults. It is uncertain where the *ka* resided during life, but "to go to one's *ka*" was a euphemism for death. The *ka* denoted power and prosperity. After death it could eat, drink, and "enjoy the odour of incense." It had to be fed, and this task was to devolve on a specific group of priests. The *ka* gave comfort and protection to the deceased: its hieroglyphic sign showed two arms outstretched upward, in an attitude of embrace.

The *ba*

The *ba* (often translated as "the soul") conveyed notions of "the noble" and "the sublime." It could enter the body or become incorporeal at will. It was represented as a human-headed falcon, presumably to emphasize its mobility. The *ba* remained sentimentally attached to the dead body, for whose well-being it was somehow responsible. It is often depicted flying about the portal of the tomb or perched on a nearby tree. Although its anatomical substratum was ill-defined, it could not survive without the preserved body.

Other important attributes were an individual's *khu* ("spiritual intelligence"), *sekhem* ("power"), *khaibit* ("shadow"), and *ren* ("name"). In the pyramid of King Pepi I, who ruled during the 6th dynasty (*c.* 2345–*c.* 2182 BC), it is recorded how the dead king had "walked through the iron which is the ceiling of heaven. With his panther skin upon him, Pepi passeth with his flesh, he is happy with his name, and he liveth with his double." The depictions of the dead were blueprints for immortality. Conversely, to blot out a person's name was to destroy that individual for all eternity, to eliminate him from the historical record. The Stalinist and Maoist regimes in the Soviet Union and China were later to resort to the same means, with the same end in mind. They also, however, invented the concept of "posthumous rehabilitation."

The heart played a central part in how the Egyptians thought about the functioning of the body. Political and religious considerations probably lay behind the major role attributed to the heart. Many of the so-called facts reported in the Ebers papyrus (a kind of medical encyclopaedia dating from the early part of the 18th dynasty; *i.e.,* from about 1550 BC) are really just speculations. This is surprising in view of how often bodies were opened during embalmment. A tubular system was rightly said to go from the heart "to all members" and the heart was said "to speak out of the vessels of every limb." But the vessels were thought to convey a mixture of air, blood, tears, urine, saliva, nasal mucus, semen, and at times even feces. During the process of embalming, the heart was always left in situ or replaced in the thorax. According to the renowned Orientalist Sir Wallis Budge, the Egyptians saw the heart as the "source of life and being," and any damage to it would have resulted in a "second death" in which everything (*ka, ba, khu,* and *ren*) would be destroyed. In some sarcophagi one can still read the pathetic plea "spare us a second death."

The anatomical heart was the *haty,* the word *ib* referring to the heart as a metaphysical entity embodying not only thought, intelligence, memory, and wisdom, but also bravery, sadness, and love. It was the heart in its sense of *ib* that was weighed in the famous judgment scene depicted in the Ani papyrus and elsewhere. After the deceased had enumerated the many sins he had not committed (the so-called negative confession), the heart was weighed against the feather of Ma'at (*i.e.,* against what was deemed right and true). It had to prove itself capable of achieving balance with the symbol of the law. The deceased who was judged pure was introduced to Osiris (in fact, became an Osiris). The deceased who failed was devoured by the monster Am-mit, the "eater of the dead." It was never the physical body on earth that was resurrected, but a new entity (the *Sahu*) that "germinated" from it and into which the soul would slip.

Judgment of the dead

The Egyptians were concerned that the dead should be able to breathe again. The Pyramid Texts describe the ceremony of the "opening of the mouth," by which this was achieved. Immediately before the mummy was consigned to the sepulchral chamber, specially qualified priests placed it upright, touched the face with an adz, and proclaimed "thy mouth is opened by Horus with his little finger, with which he also opened the mouth of his father Osiris." It has proved difficult to relate this ritual, in any meaningful way, to specific beliefs about the *ka* or *ba.*

The brain is not mentioned much in any of the extant medical papyruses from ancient Egypt. It is occasionally described as an organ producing mucus, which drained out through the nose; or it is referred to by a generic term applicable to the viscera as a whole. Life and death were matters of the heart, although the suggested relationships were at times bizarre—for example, it was said that the "mind passed away" when the vessels of the heart were contaminated with feces. The only reference that might relate death to the brain stem is the strange statement in the Ebers papyrus (gloss 854f) to the effect that "life entered the body through the left ear, and departed through the right one."

It is clear why the Egyptians never cremated their dead: to do so would have destroyed for the deceased all prospects of an afterlife. Fortunately, there was no question of organ transplantation; in the prevailing cultural context, it would never have been tolerated. Whether the pharaohs would have been powerful enough—or rash enough—to transgress accepted norms had transplantation been feasible is quite another matter.

**Mesopotamia.** The Mesopotamian (Sumerian, Babylonian, and Assyrian) attitudes to death differed widely from those of the Egyptians. They were grim and stark: sickness and death were the wages of sin. This view was to percolate, with pitiless logic and simplicity, through Judaism into Christianity. Although the dead were buried in Mesopotamia, no attempts were made to preserve their bodies.

According to Mesopotamian mythology, the gods had made humans of clay, but to the clay had been added the flesh and blood of a god specially slaughtered for the occasion. God was, therefore, present in all people. The sole purpose of humanity's creation was to serve the gods, to carry the yoke and labour for them. Offended gods withdrew their support, thereby opening the door to demons, whose activities the malevolent could invoke.

Notions of the afterlife

The main strands of Sumero-Akkadian thought held no prospect of an afterlife, at any rate of a kind that anyone might look forward to. In the Gilgamesh epic, the aging folk hero, haunted by the prospect of his own death, sets off to visit Utnapishtim, who, with his wife, was the only mortal to have achieved immortality. He meets Siduri, the wine maiden, who exhorts him to make the most of the present for "the life which thou seekest thou wilt not find." There was no judgment after death, a common fate awaiting the good and the bad alike. Death was conceived

of in terms of appalling grimness, unrelieved by any hope of salvation through human effort or divine compassion. The dead were, in fact, among the most dreaded beings in early Mesopotamian demonology. In a myth called "The Descent of Ishtar to the Underworld," the fertility goddess decides to visit *kur-nu-gi-a* ("the land of no return"), where the dead "live in darkness, eat clay, and are clothed like birds with wings." She threatens the doorkeeper: "If thou openest not that I may enter I will smash the doorpost and unhinge the gate. I will lead up the dead, that they may eat the living." Given this background, it is not surprising that offerings to the dead were made in a spirit of fear; if not propitiated they would return and cause all kinds of damage.

The Babylonians did not dissect bodies, and their approach to disease and death was spiritual rather than anatomical or physiological. They did not speculate about the functions of organs but considered them the seat of emotions and mental faculties in general. The heart was believed to be the seat of the intellect, the liver of affectivity, the stomach of cunning, the uterus of compassion, and the ears and the eyes of attention. Breathing and life were thought of in the same terms. The Akkadian word *napistu* was used indifferently to mean "the throat," "to breathe," and "life" itself.

**Judaism.** The canonical writings of biblical Judaism record the relations between certain outstanding individuals and their god. The events described are perceived as landmarks in the unfurling of a national destiny, designed and guided by that god. Jewish eschatology is in this sense unique: its main concern is the fate of a nation, not what happens to an individual at death or thereafter.

Absence of an afterlife in classical Judaism

In classical Judaism death closes the book. As the anonymous author of Ecclesiastes bluntly put it: "For the living know that they will die, but the dead know nothing, and they have no more reward" (Eccles. 9:5). The death of human beings was like that of animals: "As one dies, so dies the other. They all have the same breath, and man has no advantage over the beasts . . . all are from the dust, and all turn to dust again" (Eccles. 3:19–20). Life alone mattered: "A living dog is better than a dead lion" (Eccles. 9:4). Even Job, whose questioning at times verges on subverting Yahwist doctrine, ends up endorsing the official creed: "Man dies, and is laid low . . . . As waters fail from a lake, and a river wastes away and dries up, So man lies down and rises not again; till the heavens are no more he will not awake, or be roused out of his sleep" (Job 14:10–12).

Yet such views were far from universal. The archaeological record suggests that the various racial elements assimilated to form the Jewish nation each had brought to the new community its own tribal customs, often based on beliefs in an afterlife. Both Moses (Deut. 14:1) and Jeremiah (Jer. 16:6) denounced mortuary practices taken to imply such beliefs. Necromancy, although officially forbidden, was widely practiced, even in high places. Saul's request to the witch of Endor to "bring up" the dead prophet Samuel for him (I Sam. 28:3–20) implied that the dead, or at least some of them, still existed somewhere or other, probably in Sheol, "the land of gloom and deep darkness" (Job 10:21). In Sheol, the good and the wicked shared a common fate, much as they had in the Babylonian underworld. The place did not conjure up images of an afterlife, for nothing happened there. It was literally inconceivable, and this is what made it frightening: death was utterly definitive, even if rather ill-defined.

Many were unsatisfied by the idea that individual lives only had meaning inasmuch as they influenced the nation's destiny for good or ill. There was only one life, they were told, yet their everyday experience challenged the view that it was on earth that Yahweh rewarded the pious and punished the wicked. The Book of Job offered little solace: it was irrelevant that the good suffered and that the wicked prospered. One did not pray to improve one's prospects. The worship of God was an end in itself; it was what gave meaning to life. Against this backdrop of beliefs, the longing for personal significance was widespread.

It is difficult to determine when the notion of soul first emerged in Jewish writings. The problem is partly philological. The word *nefesh* originally meant "neck" or "throat," and later came to imply the "vital spirit," or *anima* in the Latin sense. The word *ruach* had at all times meant "wind" but later came to refer to the whole range of a person's emotional, intellectual, and volitional life. It even designated ghosts. Both terms were widely used and conveyed a wide variety of meanings at different times, and both were often translated as "soul."

Emerging ideas of an afterlife

The notion of a resurrection of the dead has a more concrete evolution. It seems to have originated during Judaism's Hellenistic period (4th century BC–2nd century AD). Isaiah announced that the "dead shall live, their bodies shall rise," and the "dwellers in the dust" would be enjoined to "awake and sing" (Isa. 26:19). Both the good and the wicked would be resurrected. According to their deserts, some would be granted "everlasting life," others consigned to an existence of "shame and everlasting contempt" (Dan. 12:2). The idea that a person's future would be determined by conduct on earth was to have profound repercussions. The first beneficiaries seem to have been those killed in battle on behalf of Israel. Judas Maccabeus, the 2nd-century-BC Jewish patriot who led a struggle against Seleucid domination and Greek cultural penetration, found that his own supporters had infringed the law. He collected money and sent it to Jerusalem to expiate their sins, acting thereby "very well and honorably, taking account of the resurrection. For if he were not expecting that those who had fallen would rise again, it would have been superfluous and foolish to pray for the dead" (II Macc. 12:43–45).

Sheol itself became departmentalized. According to the *First Book of Enoch,* a noncanonical work believed to have been written between the 2nd century BC and the 2nd century AD, Sheol was composed of three divisions, to which the dead would be assigned according to their moral deserts. The real Ge Hinnom ("Valley of Hinnom"), where the early Israelites were said to have sacrificed their children to Moloch (and in which later biblical generations incinerated Jerusalem's municipal rubbish), was transmuted into the notion of Gehenna, a vast camp designed for torturing the wicked by fire. This was a clear precursor of things to come—the Christian and Islāmic versions of hell.

Orphic and Platonic ideas also came to exert a profound influence on the Judaic concept of death. These were perhaps expressed most clearly in the apocryphal text known as the Wisdom of Solomon, written during the 1st century BC and reflecting the views of a cultured Jew of the Diaspora. The author stressed that a "perishable body weighs down the soul" (Wisd. Sol. 9:15) and stated that "being good" he had "entered an undefiled body" (Wisd. Sol. 8:20), a viewpoint that was quintessentially Platonic in its vision of a soul that predated the body. Flavius Josephus, the Jewish historian of the 1st century AD, recorded in *Bellum Judaicum* (*History of the Jewish War*) how doctrinal disputes about death, the existence of an afterlife, and the "fate of the soul" were embodied in the views of various factions. The Sadducees (who spoke for a conservative, sacerdotal aristocracy) were still talking in terms of the old Yahwist doctrines, while the Pharisees (who reflected the views of a more liberal middle class) spoke of immortal souls, some doomed to eternal torment, others promised passage into another body). The Essenes held views close to those of the early Christians.

Following the destruction of the Temple (AD 70) and, more particularly, after the collapse of the last resistance to the Romans (*c.* 135), rabbinic teaching and exegesis slowly got under way. These flowered under Judah ha-Nasi ("Judah the Prince"), who, during his reign (*c.* 175–*c.* 220) as patriarch of the Jewish community in Palestine, compiled the collection of rabbinic law known as the Mishna. During the next 400 years or so, rabbinic teaching flourished, resulting in the production and repeated reelaboration first of the Palestinian (Jerusalem) and then of the Babylonian Talmuds. These codes of civil and religious practice sought to determine every aspect of life, including attitudes toward the dead. The concepts of immortality and resurrection had become so well established that in the Eighteen Benedictions (recited daily in syna-

Rabbinic teachings on immortality

gogues and homes) God was repeatedly addressed as "the One who resurrects the dead." Talmudic sources warned that "anyone who said there was no resurrection" would have no share in the world to come (tractate *Sanhedrin* 10:1). Over the centuries, a radical doctrinal shift had occurred. One would have to await the great political volte-faces of the 20th century to witness again such dramatic gyrations of decreed perspective.

One of the strangest notions to be advanced by rabbinic Judaism—and of relevance to the evolution of the concept of death—was that of the "bone called Luz" (or *Judenknöchlein,* as it was to be called by early German anatomists). In his *Glossa magna in Pentateuchum* (AD 210), Rabbi Oshaia had affirmed that there was a bone in the human body, just below the 18th vertebra, that never died. It could not be destroyed by fire, water, or any other element, nor could it be broken or bruised by any force. In his exceeding wisdom, God would use this bone in the act of resurrection, other bones coalescing with it to form the new body that, duly breathed upon by the divine spirit, would be raised from the dead. The name of the bone was derived from *lus,* an old Aramaic word meaning "almond." The emperor Hadrian had apparently once asked Rabbi Joshua, son of Chanin, how God would resurrect people in the world to come. The rabbi had answered "from the bone Luz in the spinal column." He had then produced a specimen of such a bone, which could not be softened in water or destroyed by fire. When struck with a hammer, the bone had remained intact while the anvil upon which it lay had been shattered. The bone had apparently been called Aldabaran by the Arabs. In some of the most interesting writings of polemical anatomy, Vesalius showed, in 1543, that the bone did not exist.

Orthodox Jewish responses to current medical controversies concerning death are based on biblical and Talmudic ethical imperatives. First, nothing must be done that might conceivably hasten death. Life being of infinite worth, a few seconds of it are likewise infinitely valuable. Causing accidental death is seen as only one step removed from murder. When a patient is in the pangs of death the bed should not be shaken, as even this might prove to be the last straw. Such invasive diagnostic procedures as four-vessel angiography (to assess cerebral blood flow) would almost certainly be frowned upon. Even a venipuncture (say, for tissue typing) could be conceived of as *shpikhut damim,* a spilling of blood with nefarious intent. In secular medical practice, however, problems of this sort are unlikely to arise. Much more important is the conceptual challenge presented by the beating-heart cadaver. Here it must be stressed that absence of a heartbeat was never considered a cardinal factor in the determination of death (Bab. Talmud, tractate *Yoma* 85A). Talmudic texts, moreover, clearly recognized that death was a process and not an event: "the death throes of a decapitated man are not signs of life any more than are the twitchings of a lizard's amputated tail" (Bab. Talmud, tractate *Chullin* 21A; Mishna, *Oholoth* 1:6). The decapitated state itself defined death (Maimonides: *Tumath Meth* 1:15). Brain-stem death, which is physiological decapitation, can readily be equated with death in this particular perspective.

Orthodox responses to the beating-heart cadaver

What mattered, in early Jewish sources, was the capacity to breathe spontaneously, which was seen as an indicator of the living state. The Babylonian Talmud (tractate *Yoma* 85A) explained that when a building collapsed, all life-saving activities could legitimately cease on determination that the victim was no longer breathing. The instructions were quite explicit: "As soon as the nose is uncovered no further examination need be made, for the Tanach (Bible) refers to 'all living things who have the breath of life in their nostrils.' "

Apnea alone, of course, does not constitute death; it is a necessary but not a sufficient condition for such a diagnosis. But if apnea is conjoined to all that is implied in the notion of the decapitated state (in terms of the irreversible loss of the capacity for consciousness, for instance), one finds that the concepts of death in the Talmud and in the most modern intensive care unit are virtually identical.

The issue of transplantation is more complex. The Talmud forbids the mutilation of a corpse or the deriving of any benefit from a dead body, but these considerations can be overridden by the prescriptions of *pikuakh nefesh* ("the preservation of life"). The Chief Rabbi of Israel has even argued that, as a successful graft ultimately becomes part of the recipient, prohibitions related to deriving benefit from the dead do not, in the long run, apply.

**Hinduism.** Among the collected hymns of the Rigveda (which may date from 1500 BC and probably constitute the earliest known book in the world), there is a "Song of Creation." "Death was not there," it states, "nor was there aught immortal." The world was a total void, except for "one thing, breathless, yet breathed by its own nature." This is the first recorded insight into the importance of respiration to potential life.

Later, by about 600 BC, the *Upaniṣad*s (a collection of searching, intellectually stimulating Indo-Aryan texts) record the quest for a coordinating principle that might underlie such diverse functions of the individual as speech, hearing, and intellect. An essential attribute of the living was their ability to breathe (*an*). Their *praṇa* ("breath") was so vital that on its cessation the body and its faculties became lifeless and still. The word for "soul," *ātman,* is derived from *an,* thus placing the concept of breath at the very core of the individual self or soul.

The Hindu concept of the soul is central to an understanding of most Hindu practices related to death. In *The Discovery of India,* Jawaharlal Nehru described Hinduism as a faith that was "vague, amorphous, many-sided and all things to all men." The practices that the religion inspires do indeed entail acts that appear contradictory. What is unique to Hinduism, however, is that these are not perceived as contradictions. A common thread unites the most abstract philosophical speculations and childish beliefs in ghosts; a deep respect for nonviolence and the bloodiness of certain sacrificial rites; extreme asceticism and the sexual aspects of Tantric worship. At very different levels of sophistication, these all represent attempts to expand human perception of the truth and to achieve a cosmic consciousness. To the intellectually inclined Hindu, the eternal, infinite, and all-pervasive principle of Brahman alone is real, and the acquisition of cosmic consciousness allows humans to become one with it. The individual soul (*ātman*) is merely a particle of this cosmic principle, the relationship being likened to that between air, temporarily trapped in an earthen jar, and the endless space without; or to that between a particular wave and the ocean as a whole.

Death practices are probably more important in Hinduism than in any other religion. At one level they derive from explicit religious premises. Each being is predestined to innumerable rebirths (*saṃsāra*), and one's aggregate moral balance sheet (*karman*) determines both the length of each life and the specific form of each rebirth. Moral attributes are minutely quantifiable causal agents: every grain sown in this existence is reaped in the next. The prospect of innumerable lives is therefore envisaged with dismay. To escape the dreaded rebirths is to achieve final emancipation (*mokṣa*). "Life everlasting" (at least of the type already sampled) is the last thing a Hindu would aspire to. *Mokṣa* can be achieved only by the saintly, or perhaps by those who have died in Vārānasi and had their ashes strewed on the Ganges River. For others, the wages of worldliness is inevitable reincarnation.

The cycle of rebirths

Hindu death practices, however, also reflect popular beliefs and fears, as well as local customs. They thus may vary considerably from region to region or from sect to sect, bearing a rather variable relation to religious doctrine. Many practices are derived from the *Dharma-śastra* of Manu, the most authoritative of the books of Hindu sacred law. The alleged author of the book is the mythical sage Manu, who combined flood-surviving attributes (like Noah of Jews and Christians, and Utnapishtim of the Mesopotamians) with law-giving propensities (like Moses and Hammurabi). The book, which grew by repeated additions over many centuries, reflects the evolving interests of a male Brahman priesthood: its prescriptions are overwhelmingly recorded in terms of what is appropriate for men. Women are seldom referred to, and then often in derogatory terms.

*Death practices.* Hindus hold that a span of 120 years has been allotted to human life, a strange notion in a country where the average life expectancy was under 30 into the 20th century. They have no difficulty with the concept of death as a process. Mythological beliefs involving early Vedic gods held that the god reigning over the ears departed early, as did the gods of the eyes, hands, and mind.

When devout Hindus sense death approaching, they begin repeating the monosyllable *Om.* (This word refers to Brahman and is widely used in religious observance to help concentrate the mind on what matters.) If it is the last word on a person's lips, it guarantees a direct passage to *mokṣa.* When the dying are judged to have only an hour or so left, they are moved from their bed to a mattress on the floor and their heads are shaved. The space between ground and the ceiling is thought to symbolize the troubled area between earth and sky, and those dying there may return after death as evil spirits. A space on the ground is sanctified with Ganges water and various other ingredients, including cow dung, barley, and sesame seeds. A Hindu should never die in bed, but lying on the ground. As they take their last breaths, the dying are moved from the mattress to ground. Experienced members of the family are usually present to help decide the opportune moment. Water taken from the confluence of the rivers Ganges and Yamuna (at Allahābād) is poured into the mouth, into which is also placed a leaf of the tulsi plant (*Ocimum sanctum*). The forehead is smeared with white clay (*gopi candana*). A woman whose death precedes her husband's is considered so fortunate that her face, and especially her forehead, may be smeared with red. Sometimes, if there is doubt as to whether death has occurred, a lump of ghee (clarified butter) is placed on the forehead; if it does not melt, it is taken as a sign that life is extinct—an interesting but potentially misleading practice in the light of modern awareness of how hypothermia can mimic death. The dead body is wrapped in clean cloth of varying colours that indicate age. In the home the relatives walk clockwise around the body; they will walk around the funeral pyre in the opposite direction.

Rituals for the dying

The body is looked upon as an offering to Agni, god of fire. According to the Vedas, the Indo-Aryans used to bury their dead. Why the Hindus and Buddhists burn theirs has been the subject of much controversy. It has been variously interpreted as a gesture of purification, as the most efficient means of releasing the soul from the corrupted body, as a public health measure with important ecological benefits in a crowded country, or as a symbol of the transitory nature of any particular life and the desire that it should end in permanent anonymity. Fire taken from the deceased's home is transported to the cremation ground in a black earthen pot; this is carried immediately in front of the deceased, and nothing must come between them. For many years women were not allowed to follow the cortege, and only the wives of Brahmans could walk around the pyre. At the cremation site, a lighted torch is handed to the eldest son or grandson, who ignites the pyre, near the feet of the dead woman, at the head of the dead man. While the body is burning the soul is thought to seek refuge within the head. The intense heat usually explodes the skull, liberating the soul; when this does not happen spontaneously, the skull is deliberately shattered by blows from a cudgel. Other traditions hold that the soul passes out through the nose, eyes, and mouth. Some believe it is better still if it leaves through the anterior fontanel, an opening in the skull that normally closes during early childhood. Such theorists hold that if the deceased has practiced yoga or intense meditation, this opening will reopen, allowing free passage to the soul. In some parts of India it is believed that the souls of the really wicked depart through the rectum, and in so doing acquire such defilement that endless purification is necessary.

Cremation

Children under the age of two are not cremated but buried. When dying, they are not placed on the ground; instead they are allowed to expire in their mothers' arms. There are no special death rites; it is felt the child must have been a monster of iniquity in its previous life to have incurred such a terrible *karman.* Infant mortality is clearly attributed to the child's own wickedness and carries a load of 84 *lakh*s of rebirths (*i.e.,* the child has to be reborn 8,400,000 times). The ceremonial defilement of relatives is short, lasting only three days. Among the very high-caste Nagaras, when a pregnant woman dies the fetus is removed and buried, while the mother is cremated.

Ascetics, too, are buried rather than burnt, usually in an upright posture with the body surrounded with salt. Lepers and smallpox victims used to be buried in a recumbent position. Smallpox has been eradicated, and leprosy victims are usually cremated. If a Hindu "breaks caste" by becoming either a Muslim or a Christian, a death ceremony is conducted, the relatives bathe to purge their defilement, and the person's name is never mentioned again. The concept of death clearly influences what is deemed appropriate death behaviour, as was argued earlier in this article.

*The fate of the soul.* What happens between death and reincarnation is seldom discussed in articles about Hinduism. This is regrettable, for the perception of these events helps explain some of the rites of the religion and provides unique insights into the human preference, when thinking about death, to conceptualize metaphysical developments in very concrete terms.

Immediately after death, the soul is not clothed in a physical body but in a vaporous thumb-sized structure (*linga śarīra*). This is immediately seized by two servants of Yama, the god of death, who carry it to their master for a preliminary identity check. Afterward, the soul is promptly returned to the abode of the deceased, where it hovers around the doorstep. It is important that the cremation be completed by the time of the soul's return, to prevent it from reentering the body. By the 10th day, the near relatives have purged some of the defilement (*mṛitaka sutaka*) they incurred from the death, and the chief mourner and a priest are ready to carry out the first *śrāddha* (ritual of respect). This is a step toward the reconstitution of a more substantial physical body (*yatana śarīra*) around the disembodied soul (*preta*) of the deceased. A tiny trench is dug in a ritually purified piece of land by a river, and the presence of Vishnu is invoked. Ten balls of barley flour mixed with sugar, honey, milk, curds, ghee, and sesame seeds are then placed, one by one, in the soil. As the first ball is offered, the priest says (and the son repeats after him), "May this create a head"; with the second ball, "May this create neck and shoulders"; with the third, "May this create heart and chest"; and so on. The 10th request is for the ball to create the capacity to digest, thereby satisfying the hunger and thirst of the newly created body. Bungled ceremonies can have catastrophic effects. Prayers are offered to Vishnu to help deliver the new entity (now perceived as some 18 inches [46 centimetres] long) into the power of Yama. The balls of barley are picked up from the trench and thrown into the river. Further *śrāddha*s are performed at prescribed times, varying according to caste; one of these rituals makes the soul an ancestral spirit, or *pitṛi.* With the completion of these rituals, the soul of the deceased leaves this world for its yearlong and perilous journey to Yama's kingdom. The family is now formally cleansed. The men shave their heads, and the women wash their hair. The family's tutelary god (removed by a friend at the time of the death) can be returned to its home. A feast is offered to Brahmans, neighbours, and beggars—even the local cows are given fresh grass. There is a sense of general relief: if the *śrāddha*s had not been performed, the *preta* could have become a *bhūta* (malignant spirit), repeatedly turning up to frighten the living. For the deceased, things would have been worse: the *preta* would have been left errant. (A similar fate befalls the soul of a person who commits suicide.) The horror of dying unshriven that haunted people in medieval Europe resembles the despair of the devout Hindu at the prospect of having no son to perform the *śrāddha*s.

The first *śrāddha*

The soul, in its substantial envelope, is meanwhile proceeding on its journey, holding onto a cow's tail to cross the Vaitarani, a horrible river of blood and filth that marks the boundary of Yama's kingdom. Throughout, it is sustained by further *śrāddha*s, during which friends on

Later rituals

earth seek to provide it with shoes, umbrellas, clothing, and money. These they give to a Brahman, in the hope that the deceased will benefit. During such rituals relatives have to avoid all sewing, which might occlude the *pitṛi*'s throat, rendering it incapable of ever breathing or drinking again. After a year, the *pitṛi* in its *yatana śarīra* reaches Yama's seat of judgment, where it is sentenced to a strictly limited term in heaven (*svarga*) or hell (*naraka*) according to its deserts. This completed, it moves into another body (the *karaṇa śarīra*), whose form depends on the individual's *karman*. It could be a plant, a cockroach, a canine intestinal parasite, a mouse, or a human being. Unlike Jains, Hindus believe that whatever body the soul eventually moves into, it inhabits as sole tenant, not as a tenement lodger.

**Islām.** Probably no religion deals in such graphic detail as does Islām with the creation, death, "life in the tomb," and ultimate fate of humankind. Yet the Qur'ān, the holy book of Islām, itself provides no uniform or systematic approach to these problems. It is only in its later parts (which date from the period when the small Muslim community in Medina had come into contact with other religious influences) that problems such as the relation of sleep to death, the significance of breathing, and the question of when and how the soul leaves the body are addressed in any detail. Popular Muslim beliefs are based on still later traditions. These are recorded in the *Kitāb al-rūḥ* ("Book of the Soul") written in the 14th century by the Ḥanbalī theologian Muḥammad ibn Abī-Bakr ibn Qayyīm al-Jawzīyah.

Predestination in Islāmic teachings

The basic premise of all Qur'ānic teaching concerning death is Allāh's omnipotence: he creates human beings, determines their life span, and causes them to die. The Qur'ān states: "Some will die early, while others are made to live to a miserable old age, when all that they once knew they shall know no more (22:5; *i.e., sūrah* [chapter] 22, verse 5). Damnation and salvation are equally predetermined: "Allāh leaves to stray whom he willeth, and guideth whom he willeth" (35:8). As for those whom Allāh leaves astray, the Qur'ān states that "for them there will be no helpers" (30:29). Allāh has decided many will fail: "If We had so willed We could certainly cause proper guidance to come to every soul, but true is My saying 'assuredly I shall fill Jihannam' " (32:13).

In this perspective the individual's fate (including the mode and time of death) appears inescapably predetermined. The very term *Islām,* Arabic for "surrender," implies an absolute submission to the will of God. But what freedom does this allow those predestined to continue in the path of error, or to reject God's will? And if there is no such freedom, what sense was there in the mission of the Prophet Muḥammad (Islām's founder) and his appeal to people to alter their ways? It is hardly surprising that arguments about free will and predestination broke out soon after the Prophet's death. The ensuing tensions dominated theological (and other) controversies within Islām during many centuries.

Nature of the soul

Questions concerning the meaning of life and the nature of the soul are dealt with patchily in both the Qur'ān and the Ḥadīth (the record of the sayings attributed to the Prophet). The Qur'ān records that, when asked about these matters by local leaders of the Jewish faith, the Prophet answered that "the spirit cometh by command of God" and that "only a little knowledge was communicated to man" (17:85). Humanity was created from "potter's clay, from mud molded into shape" into which Allāh has "breathed his spirit" (15:28–29). A vital spirit or soul (*nafs*) is within each human being. It is associated, if not actually identified, with individuality and also with the seat of rational consciousness. It is interesting to speculate on the possible relation of the term *nafs* to such Arabic words as *nafas* ("breath") and *nafīs* ("precious"), particularly in a language where there are no written vowels.

Death is repeatedly compared with sleep, which is at times described as "the little death." God takes away people's souls "during their sleep" and "upon their death." He "retains those against whom he has decreed death, but returns the others to their bodies for an appointed term" (39:42–43). During death, the soul "rises into the throat" (56:83) before leaving the body. These are interesting passages in the light of modern medical knowledge. The study of sleep has identified the episodic occurrence of short periods during which the limbs are totally flaccid and without reflexes, as would be the limbs of the recently dead. Modern neurophysiology, moreover, stresses the role of structures in the upper part of the brain stem in the maintenance of the waking state. Lesions just a little higher (in the hypothalamus) cause excessively long episodes of sleep. Irreversible damage at these sites is part of the modern concept of death. Finally, various types of breathing disturbance are characteristic of brain-stem lesions and could have been attributed, in former times, to occurrences in the throat. Nothing in these passages outrages the insights of modern neurology. The absence of any cardiological dimension is striking.

Events at the moment of death

It is orthodox Muslim belief that when someone dies the Angel of Death (*malāk al-mawt*) arrives, sits at the head of the deceased, and addresses each soul according to its known status. According to the *Kitāb al-rūh,* wicked souls are instructed "to depart to the wrath of God." Fearing what awaits them, they seek refuge throughout the body and have to be extracted "like the dragging of an iron skewer through moist wool, tearing the veins and sinews." Angels place the soul in a hair cloth and "the odour from it is like the stench of a decomposing carcass." A full record is made, and the soul is then returned to the body in the grave. "Good and contented souls" are instructed "to depart to the mercy of God." They leave the body, "flowing as easily as a drop from a waterskin"; are wrapped by angels in a perfumed shroud, and are taken to the "seventh heaven," where the record is kept. These souls, too, are then returned to their bodies.

Two angels coloured blue and black, known as Munkar and Nakīr, then question the deceased about basic doctrinal tenets. In a sense this trial at the grave (*fitnat al-Qabr*) is a show trial, the verdict having already been decided. Believers hear it proclaimed by a herald, and in anticipation of the comforts of *al-jannah* (the Garden, or "paradise") their graves expand "as far as the eye can reach." Unbelievers fail the test. The herald proclaims that they are to be tormented in the grave; a door opens in their tomb to let in heat and smoke from *jihannam* ("hell"), and the tomb itself contracts "so that their ribs are piled up upon one another." The period between burial and the final judgment is known as *al-barzakh.*

Final judgment

At the final judgment (*yaum al-Hisāb*), unbelievers and the god-fearing are alike resurrected. Both are endowed with physical bodies, with which to suffer or enjoy whatever lies in store for them. The justified enter Gardens of Delight, which are described in the Qur'ān in terms of prevalent, but essentially masculine, tastes (37:42–48). At the reception feast on the Day of Judgment unbelievers fill their bellies with bitter fruit, and "drink down upon it hot water, drinking as drinks the camel crazed with thirst" (56:52–55). They then proceed to hell, where they don "garments of fire" (22:19) and have boiling water poured over their heads. Allāh has made provision against the annihilation of the body of the damned, promising that "whenever their skins are cooked to a turn, We shall substitute new skins for them, that they may feel the punishment" (4:56). Pleas for annihilation are disregarded. Although this is sometimes referred to as the "second death," the Qur'ān is explicit that in this state the damned "neither live nor die" (87:13).

A special fate is reserved for the martyrs of Islām; *i.e.,* for those who fall in a *jihād* ("holy war"). Their evil deeds are instantly expiated and the formalities of judgment are waived; they enter the Garden immediately. Similar dispensations are promised to "those who had left their homes, or been driven therefrom, or who had suffered harm" in the divine cause (3:195). For the Shīʿites, followers of the smaller of Islām's two major branches, the prospects for martyrdom are even wider. A major event of the origin of Shīʿism, moreover, was the slaughter of the Prophet's grandson, Ḥusayn, in 680; this heritage has imbued Shīʿism with a zeal for martyrdom. Some of the behaviour of Islāmic fundamentalists is explicable from this perspective.

A gentler strand in Islāmic eschatology produced, over

the centuries, a series of reinterpretations or adaptations of the original doctrine, some of whose tenets were even claimed to have been only metaphorical. These tendencies, which stressed individual responsibility, were often influenced by the Ṣūfīs (Islāmic mystics).

Muslims accord a great respect to dead bodies, which have to be disposed of very promptly. The mere suggestion of cremation, however, is viewed with abhorrence. The philosophical basis, if any, of this attitude is not clear. It is not stated, for instance, that an intact body will be required at the time of resurrection. It is unlikely, moreover, that the abhorrence—which Orthodox Jews share—arose out of a desire to differentiate Islāmic practices from those of other "people of the Book" (*i.e.,* Jews and Christians). The attitude toward dead bodies has had practical consequences; for instance, in relation to medical education. It is almost impossible to carry out postmortem examinations in many Islāmic countries. Medical students in Saudi Arabia, for example, study anatomy on corpses imported from non-Islāmic countries. They learn pathology only from textbooks; many complete their medical training never having seen a real brain destroyed by a real cerebral hemorrhage.

Attitudes toward transplantation

In 1982 organ donation after death was declared *ḥallāl* ("permissible") by the Senior 'Ulamā' Commission, the highest religious authority on such matters in Saudi Arabia (and hence throughout the Islāmic world). Tales inculcated in childhood continue, however, to influence public attitudes in Islāmic nations. The widely told story of how the Prophet's uncle Ḥamzah was murdered by the heathen Hind, who then opened the murdered man's belly and chewed up his liver, has slowed public acceptance of liver transplantation. Kidney transplantation is more acceptable, perhaps because the Ḥadīth explicitly states that those entering the Garden will never more urinate.

**The modern Western context.** *The Christian legacy.* The spread of rationalistic and scientific ideas since the 18th century has undermined many aspects of religion, including many Christian beliefs. The church, moreover, although still seeking to exert its influence, has ceased to dominate civil life in the way it once did. Religion is no longer the pivot of all social relations as it once was in ancient Egypt and still is in some Islāmic countries. The decline of the church is epitomized by the fact that, while it is still prepared to speak of the symbolic significance of the death of Jesus Christ (and of human death in general), it has ceased to emphasize many aspects of its initial eschatology and to concern itself, as in the past, with the particular details of individual death. In the age of Hiroshima and Nagasaki, the elaborate descriptions of heaven, purgatory, and hell in Dante's *Divine Comedy,* while remaining beautiful literature, at best raise a smile if thought of as outlines for humanity's future.

The centrality of death to Christianity

Death is at the very core of the Christian religion. Not only is the cross to be found in cemeteries and places of worship alike, but the premise of the religion is that, by their own action, humans have forfeited immortality. Through abuse of the freedom granted in the Garden of Eden, Adam and Eve not only sinned and fell from grace, but they also transmitted sin to their descendants: the sins of the fathers are visited on the children. And as "the wages of sin is death" (Rom. 6:23), death became the universal fate: "Therefore as sin came into the world through one man and death through sin, and so death spread to all men" (Rom. 5:12). Christian theologians spent the best part of two millennia sorting out these implications and devising ways out of the dire prognosis implicit in the concept of original sin. The main salvation was to be baptism into the death of Jesus Christ (Rom. 6:3–4).

Among early Christians delay in the promised Second Coming of Christ led to an increasing preoccupation with what happened to the dead as they awaited the resurrection and the Last Judgment. One view was that there would be an immediate individual judgment and that instant justice would follow: the deceased would be dispatched forthwith to hell or paradise. This notion demeaned the impact of the great prophecy of a collective mass resurrection, followed by a public mass trial on a gigantic scale. Moreover, it deprived the dead of any chance of a postmortem (*i.e.,* very belated) expiation of their misdeeds. The Roman Catholic notion of purgatory sought to resolve the latter problem; regulated torture would expiate some of the sins of those not totally beyond redemption.

The second view was that the dead just slept, pending the mass resurrection. But as the sleep might last for millennia, it was felt that the heavenly gratification of the just was being arbitrarily, and somewhat unfairly, deferred. As for the wicked, they were obtaining an unwarranted respite. The Carthaginian theologian Tertullian, one of the Church Fathers, outlined the possibility of still further adjustments. In his *Adversus Marcionem,* written about 207, he described "a spatial concept that may be called Abraham's bosom for receiving the soul of all people." Although not celestial, it was "above the lower regions and would provide refreshment (*refrigerium*) to the souls of the just until the consummation of all things in the great resurrection." The Byzantine Church formally endorsed the concept, which inspired some most interesting art in both eastern and western Europe.

During its early years, the Christian Church debated death in largely religious terms. The *acerbitas mortis* ("bitterness of death") was very real, and pious deathbeds had to be fortified by the acceptance of pain as an offering to God. Life expectancy fell far short of the promised threescore years and 10. Eastern medicine remained for a long time in advance of that practiced in the West, and the church's interventions were largely spiritual. It was only during the Renaissance and the later age of Enlightenment that an intellectual shift became perceptible.

*Descartes, the pineal soul, and brain-stem death.* The first attempts to localize the soul go back to classical antiquity. The soul had originally been thought to reside in the liver, an organ to which no other function could, at that time, be attributed. Empedocles, Democritus, Aristotle, the Stoics, and the Epicureans had later held its abode to be the heart. Other Greeks (Pythagorus, Plato, and Galen) had opted for the brain. Herophilus (flourished *c.* 300 BC), a famous physician of the Greek medical school of Alexandria, had sought to circumscribe its habitat to the fourth ventricle of the brain; that is, to a small area immediately above the brain stem. Controversy persisted to the very end of the 16th century.

Concepts of the soul

The departure of the soul from the body had always been central to the Christian concept of death. But the soul had come to mean different things to various classical and medieval thinkers. There was a "vegetative soul," responsible for what we would now call autonomic function; a "sensitive soul," responsible for what modern physiologists would describe as reflex responses to environmental stimuli; and, most importantly, a "reasoning soul," responsible for making a rational entity (*res cogitans*) of human beings. The reasoning soul was an essentially human attribute and was the basis of thought, judgment, and responsibility for one's actions. Its departure implied death. The *Anatome Corporis Humani* (1672) of Isbrand van Diemerbroeck, professor at Utrecht, appears to have been the last textbook of anatomy that discussed the soul within a routine description of human parts. Thereafter, the soul disappeared from the scope of anatomy.

The modern and entirely secular concept of brain-stem death can, perhaps rather surprisingly, find both a conceptual and a topographical foundation in the writings of René Descartes (1596–1650), the great French philosopher and mathematician who sought to bring analytical geometry, physics, physiology, cosmology, and religion into an integrated conceptual framework. Descartes considered the body and the soul to be ontologically separate but interacting entities, each with its own particular attributes. He then sought to specify both their mode and site of interaction; the latter he deduced to be the pineal gland. The pineal was to become, in the words of Geoffrey Jefferson, "the nodal point of Cartesian dualism."

Before Descartes, the prevailing wisdom, largely derived from Greece, had regarded the soul both as the motive force of all human physiological functions and as the conscious agent of volition, cognition, and reason. Descartes succeeded in eliminating the soul's general physiological role altogether and in circumscribing its cognitive role to

the human species. Descartes's writings about death show that his concept of the soul clearly implied both mind and the immaterial principle of immortality. It had to mean both things, for no one had ever conceived of survival after death without a mind to verify the fact of continued existence, to enjoy its pleasures, and to suffer its pains.

The relation between body and soul had been discussed in patristic literature, and, because of his Jesuit education, Descartes would have been familiar with these discussions. The church's interest in these matters was strictly nonmedical, seeking only to reconcile earlier Greek theories with its own current doctrines. Descartes was the first to tackle these problems in a physiological way. With one foot still firmly on consecrated ground (and with Galileo's difficulties with the Inquisition very much in mind), he sought to give a materialistic, even mechanistic, dimension to the discussion. In this sense, his *De Homine* (*On Man*; published posthumously in 1662) can be thought of as an updating of Plato's *Timaeus*. His contemporaries viewed Descartes as having delivered the coup de grace to an earlier Greek tradition (dating back to several centuries before Christ) that had claimed that animals, as well as humans, had souls. This had been the subject of much discussion in the early Christian Church. During the 4th century, St. John Chrysostom (onetime archbishop of Constantinople) had denounced the idea, attributing it to the devil, who had allegedly managed by various maneuvers to deceive people as varied as Pythagoras, Plato, Pliny, and even Zoroaster.

Descartes probably was impressed by the central location of the unpaired pineal gland, situated where neural pathways from the retinas converge with those conveying feelings from the limbs. This "general reflector of all sorts of sensation" is, moreover, sited in the immediate proximity of the brain ventricles, from which (according to the wisdom of the day) "animal spirits" flowed into the hollow nerves, carrying instructions to the muscles. In his *Excerpta Anatomica,* Descartes had even likened the pineal to a penis obturating the passage between the third and fourth ventricles.

Descartes's ideas regarding the pineal gland

Descartes proved wrong in his beliefs that all sensory inputs focused on the pineal gland and that the pineal itself was a selective motor organ, suspended in a whirl of "animal spirits," dancing and jigging "like a balloon captive above a fire," yet capable in humans of scrutinizing inputs and producing actions "consistent with wisdom." He was also wrong when he spoke of the "ideas formed on the surface" of the pineal gland, and in his attribution to the pineal of such functions as "volition, cognition, memory, imagination, and reason." But he was uncannily correct in his insight that a very small part of this deep and central area of the brain was relevant to some of the functions he stressed. We now know that immediately below the pineal gland there lies the mesencephalic tegmentum (the uppermost part of the brain stem), which is crucial to generating alertness (the capacity for consciousness), without which, of course, there can be no volition, cognition, or reason.

It is a matter of vocabulary whether one considers the mesencephalic tegmentum either as being involved in generating a "capacity for consciousness" or as preparing the brain for the exercise of what Descartes would have considered the "functions of the soul" (volition, cognition, and reason). In either case, the total and irreversible loss of these functions dramatically alters the ontological status of the subject. Descartes specifically considered the example of death. In "La Description du corps humain" (1664) he wrote that "although movements cease in the body when it is dead and the soul departs, one cannot deduce from these facts that the soul produced the movements." In a formulation of really modern tenor, he then added "one can only infer that the same single cause (a) renders the body incapable of movement and (b) causes the soul to absent itself." He did not, of course, say that this "same single cause" was the death of the brain stem. Some 300 years later, in 1968, the Harvard Committee spoke of death in terms of "irreversible coma" (where Descartes had spoken of the "now absent soul") and stressed, as had Descartes, the immobility of the comatose body. The religious and secular terms seem to describe the same reality.

There have been other neurological controversies concerning the locus of the soul. Early in the 18th century Stephen Hales, an English clergyman with a great interest in science, repeated an experiment originally reported by Leonardo da Vinci. Hales tied a ligature around the neck of a frog and cut off its head. The heart continued to beat for a while, as it usually does in the brain dead. Thirty hours later, the limbs of the animal still withdrew when stimulated. In fact, the elicited movements only ceased when the spinal cord itself had been destroyed. This observation gave rise to a great controversy. Reflex action at spinal cord level was not then fully understood, and it was argued that the irritability implied sentience, and that sentience suggested that the soul was still present. The "spinal cord soul" became the subject of much debate. It is now known that such purely spinal reflex movements may occur below a dead brain. It was shown during the 19th century that individuals executed on the guillotine might retain the knee jerk reflex for up to 20 minutes after decapitation.

Modern theological concerns

The church is still concerned with the diagnosis of death, but the theological argument has, during the last half of the 20th century, moved to an entirely different plane. As mentioned earlier, in 1957 Pope Pius XII raised the question whether, in intensive care units, doctors might be "continuing the resuscitation process, despite the fact that the soul may already have left the body." He even asked one of the central questions confronting modern medicine, namely whether "death had already occurred after grave trauma to the brain, which has provoked deep unconsciousness and central breathing paralysis, the fatal consequences of which have been retarded by artificial respiration." The answer, he said, "did not fall within the competence of the Church."

*Public attitudes.* Until about 100 years ago, people had by and large come to terms with death. They usually died in their homes, among their relatives. In villages, in the 18th or early 19th centuries, passers-by might join the priest bearing the last sacrament on his visit to the dying man or woman. Doctors even stressed the public health hazards this might cause. Numerous pictures attest to the fact that children were not excluded from deathbeds, as they were to be during the 20th century.

Changing attitudes toward death

The general acceptance of death was to be subverted by the advances of modern medicine and by the rapid spread of rationalist thought. This led, during a period of only a few decades, to a striking change of attitudes. In the advanced industrial countries, a large number of people now die in hospitals. The improvement in life expectancy and the advances of modern surgery and medicine have been achieved at a certain price. A mechanistic approach has developed, in which the protraction of dying has become a major by-product of modern technology. The philosophy of modern medicine has been diverted from attention to the sick and has begun to reify the sickness. Instead of perceiving death as something natural, modern physicians have come to see it as bad or alien, a defeat of all their therapeutic endeavours, at times almost as a personal defeat. Sickness is treated with all possible weapons, often without sufficient thought for the sick person—at times even without thought as to whether there is still a "person" at all. The capacity to "care" for biological preparations, with no other human attribute than physical form, is part of the context in which the reevaluation of death described earlier has taken place.

Parallel developments have taken place at the level of the psyche of the dying person and of the person's relatives. Elisabeth Kübler-Ross, an American psychiatrist, has outlined the stages (denial, anger, bargaining, preparatory grief, and acceptance) through which people, informed of their own approaching death, are said to pass. Her writings are based on a wide but essentially American experience, and their universality has not been tested, particularly in other cultural contexts. They may well prove somewhat ethnocentric.

The development of the death industry (satirized in Evelyn Waugh's *Loved One* and explored in Jessica Mitford's *American Way of Death*) is also a by-product of the technological revolution and of modern attitudes to

death. Undertakers have become "morticians" and coffins "caskets." Embalming has enjoyed a new vogue. Drive-in cemeteries have appeared, for those seeking to reconcile devotion to the dead with other pressing engagements. Cryogenic storage of the corpse has been offered as a means to preserve the deceased in a form amendable to any future therapies that science may devise. Commercial concerns have entered the scene: nonpayment of maintenance charges may result in threats of thawing and putrefaction. In a contentious environment, the law has even invaded the intensive care unit, influencing the decisions of physicians concerning the withdrawal of treatment or the determination of death. A wit has remarked that in the modern era, the only sure sign that a man is dead is that he is no longer capable of litigation.

**BIBLIOGRAPHY**

*Biological aspects:* The concept of apoptosis (programmed cell death) is outlined in A. GLÜCKSMAN, "Cell Deaths in Normal Vertebrate Ontogeny," *Biological Reviews of the Cambridge Philosophical Society,* 26:59–86 (1951); A.H. WYLLIE, J.F.R. KERR, and A.R. CURRIE, "Cell Death: The Significance of Apoptosis," *International Review of Cytology,* 68:251–306 (1980); I.D. BOWEN and R.A. LOCKSHIN (eds.), *Cell Death in Biology and Pathology* (1981); I. DAVIES and D.C. SIGEE (eds.), *Cell Ageing and Cell Death* (1985). The development of the idea of brain death (and of its evolution into the concept of brain-stem death) can be followed in P. MOLLARET and M. GOULON, "Le Coma dépassé," *Revue Neurologique,* 101(1):3–15 (July 1959); AD HOC COMMITTEE OF THE HARVARD MEDICAL SCHOOL TO EXAMINE THE DEFINITION OF BRAIN DEATH, "A Definition of Irreversible Coma," *J.A.M.A.,* 205(6):337–340 (Aug. 5, 1968); JULIUS KOREIN (ed.), *Brain Death: Interrelated Medical and Social Issues* (1978); A. EARL WALKER, *Cerebral Death,* 3rd ed. (1985); PRESIDENT'S COMMISSION FOR THE STUDY OF ETHICAL PROBLEMS IN MEDICINE AND BIOMEDICAL AND BEHAVIORAL RESEARCH, *Defining Death: A Report on the Medical, Legal, and Ethical Issues in the Determination of Death* (1981, reprinted 1983); BRYAN JENNETT, JOHN GLEAVE, and PETER WILSON, "Brain Death in Three Neurosurgical Units," *Br.Med.J.,* 282:533–539 (Feb. 14, 1981); CHRISTOPHER PALLIS, *ABC of Brain Stem Death* (1983), and his "Brain-stem Death: The Evolution of a Concept," in PETER J. MORRIS (ed.), *Kidney Transplantation: Principles and Practice,* 2nd ed., pp. 101–127 (1984); JAMES L. BERNAT, "The Definition, Criterion, and Statute of Death," *Seminars in Neurology,* 4(1):45–51 (March 1984).

*Philosophical and cultural aspects:* E.A. WALLIS BUDGE, *Egyptian Ideas of the Future Life: Egyptian Religion* (1899, reprinted 1979 as *Egyptian Religion: Egyptian Ideas of the Future Life*); ANGE P. LECA, *La Médecine égyptienne au temps des pharaons* (1971); ALEXANDRE PIANKOFF (ed.), *Le "Cœur" dans les textes égyptiens depuis l'ancien jusqu' à la fin du nouvel empire* (1930); and HENRY E. SIGERIST, *A History of Medicine,* 2 vol. (1951–61), are useful reviews of the notion of death in ancient Egypt. Mesopotamian concepts are described in J. HACKIN *et. al., Asiatic Mythology* (1932, reissued 1963); and SAMUEL GEORGE FREDERICK BRANDON, *Man and His Destiny in the Great Religions* (1962, reprinted 1963). The latter and F.H. GARRISON, "The Bone Called 'Luz,' " *New York Medical Journal,* 92(4):149–151 (July 23, 1910), also contain much useful information on Judaic attitudes. Hindu perceptions and practices are detailed in PAUL THOMAS, *Hindu Religion, Customs and Manners,* 6th ed. (1975); and NIRAD C. CHAUDHURI, *Hinduism: A Religion to Live By* (1979, reprinted 1980). MUḤAMMAD IBN ABĪ-BAKR IBN QAYYĪM AL-JAWZĪYAH, *Kitāb al-rūḥ,* 2nd ed. (1324); and FRANK E. REYNOLDS and EARLE H. WAUGH (eds.), *Religious Encounters with Death: Insights from the History and Anthropology of Religions* (1977), present Islāmic attitudes. More recent developments are discussed in T.S.R. BOASE, *Death in the Middle Ages: Mortality, Judgment and Remembrance* (1972); and PHILIPPE ARIÈS, *Western Attitudes Toward Death: From the Middle Ages to the Present,* translated from the French (1974, reprinted 1975), and *The Hour of Our Death* (1981, reissued 1982; originally published in French, 1977). Information about the "pineal soul" is found in RENÉ DESCARTES, *Treatise of Man,* translated from the 1664 French edition, by THOMAS STEELE HALL (1972), originally published in a Latin translation, 1662; and about the "spinal cord soul" in EDWARD GEORGE TANDY LIDDELL, *The Discovery of Reflexes* (1960). See also GEOFFREY JEFFERSON, "René Descartes on the Localisation of the Soul," *Irish Journal of Medical Science,* 285:691–706 (Sept. 1949); and G. CORNER, "Anatomists in Search of the Soul," *Annals of Medical History,* 2(1):1–7 (Spring 1919). Modern attitudes on death form the basis of JESSICA MITFORD, *The American Way of Death* (1963, reprinted 1978); ELISABETH KÜBLER-ROSS, *On Death and Dying* (1969, reprinted 1979); and ROBERT M. VEATCH, *Death, Dying and the Biological Revolution: Our Last Quest for Responsibility* (1976). HERMAN FEIFEL (ed.), *The Meaning of Death* (1959, reissued 1965); and JAMES P. CARSE, *Death and Existence: A Conceptual History of Human Mortality* (1980), both present excellent overviews.

(C.A.P.)